DEPARTAMENT DE GENÈTICA

GENOME EVOLUTION AND SYSTEMS BIOLOGY IN BACTERIAL ENDOSYMBIONTS OF INSECTS.

EUGENI BELDA CUESTA

UNIVERSITAT DE VALÈNCIA
Servei de Publicacions
2011

Aquesta Tesi Doctoral va ser presentada a València el dia 29 d'octubre de 2010 davant un tribunal format per:

- Dra. Amparo Latorre Castillo
- Dr. Mario Alí Farés
- Dr. Abdelaziz Heddi
- Dr. Hernán Javier Dopazo
- Dr. Juli Peretó i Magraner

Va ser dirigida per:
Dr. Francisco Silva Moreno
Dr. Andrés Moya Simarro

Dipòsit legal: V-4120-2011
I.S.B.N.: 978-84-370-8071-0
Edita: Universitat de València
Servei de Publicacions
C/ Arts Gràfiques, 13 baix
46010 València
Spain
Telèfon:(0034)963864115

Facultat de Ciències Biològiques

Departament de Genètica

Instituto Cavanilles de Biodiversidad y Biología Evolutiva

GENOME EVOLUTION AND SYSTEMS BIOLOGY IN BACTERIAL ENDOSYMBIONTS OF INSECTS

Memoria presentada por Eugeni Belda Cuesta para optar al grado de doctor en ciencias Biológicas por la Universitat de València

Directores:

Dr. Francisco Silva Moreno. Catedrático de la U.V.E.G.

Dr. Andrés Moya Simarro. Catedrático de la U.V.E.G.

Valencia 2010

D. Francisco Silva Moreno, Doctor en Ciencias Biológicas y Catedrático del Departamento de Genética de la Universitat de València

D. Andrés Moya Simarro, Doctor en Ciencias Biológicas y Catedrático del Departamento de Genética de la Universitat de València

CERTIFICAN: Que Eugeni Belda Cuesta, Licenciado en Ciencias Biológicas por la Universitat de València, ha realizado bajo su dirección el trabajo que lleva por título "Genome evolution and systems biology in bacterial endosymbionts of insects", para optar al Grado de doctor en Ciencias Biológicas por la Universitat de València

Fdo.: Dr. Francisco Silva Moreno

Fdo.: Dr. Andrés Moya Simarro

AGRADECIMIENTOS

A la hora de recordar a todas las personas que han compartido conmigo todos estos años de doctorado se mezclan un monton de sentimientos. Alegría por el monton de gente que he tenido la enorme fortuna de conocer durante todo este tiempo, y cuya amistad espero conservar para siempre, nostalgia por todos los momentos, los buenos pero tambien los no tan buenos, compartidos en cafés, comidas y cenas, y congresos y viajes. Y pena al recordar a todas aquellas personas que ya no estan aunque seguro que en algun lugar se alegran de que haya llegado hasta aqui. Necesitaría un capítulo adicional para mencionarlos a todos, y si de algo estoy plenamente convencido es que sin su ayuda no lo hubiera conseguido, asi que muchas gracias a todos por vuestra ayuda y apoyo durante todo este tiempo.

A mis directores de tesis, por darme la oportunidad de formarme como investigador en el mejor lugar posible tanto a nivel académico como especialmente humano, como ha sido el grupo de Genética Evolutiva del Instituto Cavanilles. A Paco, porque fuiste el que me reclutó y me ofreció la posibilidad de hacer la Tesis bajo tu supervisión y por haberme ayudado mil y una veces con todas las dudas y problemas que han surgido todo este tiempo, y a Andrés, por ayudarme a ver el camino cuando más oscuro lo veia y por tu interés y preocupación en los momentos que más lo he necesitado, muchisimas gracias a ambos.

A Laura y Vicente Pérez, por toda su ayuda durante los primeros años cuando yo no tenia ni idea de que iba todo esto. A Iñaki, el mejor instructor posible en temas de bioinformática siempre dispuesto a ayudar en cualquier cosa, y a Vicente Sentandreu, compañero de Catedral y de muchas otras cosas, porque con ambos he pasado los mejores momentos de esta etapa. A Maria Jose, Vicky y Mireia, que empezaron conmigo el doctorado, por todos los momentos de cursos, trabajos, y neuras que hemos compartido todos estos años. A Araceli, Teresa y Alicia, que vinieron luego pero con las que he compartido las mismas alegrias y desencantos. A Yolima,. A Pepa, Silvia, Nuria, y Carmen. A los últimos en llegar, Rafa, Peris, Ana, Sergio, Diego y Alejandro, a los cuales espero ver como doctores dentro de no mucho tiempo. Sin la ayuda y el apoyo de todos ellos no creo que hubiera podido llegar hasta aqui, muchisimas gracias a todos.

A Daniel, Juanan y Ronald, por haber estado siempre ahi cuando más lo he necesitado, por todos los momentos que hemos vivido juntos desde hace tanto tiempo que casi ni recuerdo, y por ser los mejores amigos que jamas hubiera podido imaginar.

A mi madre, porque todo lo que soy se lo debo a ella. Son tantas cosas que necesitaria un capitulo adicinal para ella. A mi tia Alicia y mis primos Pau y Maria, porque lo único bueno del 2007 fue el unirme mas a ellos y conocerlos y quererlos

mas. A mi Tio Emilio, a mis primos Oscar, Emilio y Isabel, a Vicente y a toda la gente de la Sala Alameda, por su confianza y cariño durante todos estos años.

A mi novia Maria Jose, por haberme aguantado todo este tiempo, por todo lo que hemos vivido y lo que espero que vivamos en esta nueva etapa. Sin tu apoyo y cariño no lo hubiera logrado.

A mis abuelos y a mi tio Manolo, que son el espejo en el que me miro cada dia y las mejores personas que he conocido jamás. Solo aspiro a parecerme a ellos algun dia.

Index

1. General introduction

The main objective of this thesis is the study of the dynamics of genome evolution in bacteria, specially focused in the process of reductive genome evolution that affects to bacterial endosymbionts of insects at first steps of adaptation to endosymbiotic lifestyle. This process is covered by two main methodological approaches, a comparative genomics approach to study the evolution of genome structure in γ-subdivision of proteobacteria, that harbors most of the genomes of bacterial endosymbionts sequenced to date, and an approach based on systems biology to study the particular situation of the genome of *Sodalis glossinidius*, the secondary endosymbiont of tsetse flies that is in an initial stage of adaptation to endosymbiotic lifestyle, to model the process of genome reduction at different stages of the endosymbiotic association. For this purpose we made an exhaustive re-annotation of its genome to evaluate the impact of gene inactivation and mobile genetic elements proliferation in the reductive evolution initiated by this endosymbiotic bacteria since its symbiotic association with the tsetse fly and the consequences of this process over the metabolic capabilities of the bacteria considering the whole endosymbiotic system comprised by *S. glossinidius* and the primary endosymbiont *W. glossinidia*. In this introduction I first describe briefly the main evolutionary process that generates the huge diversity in genomic structures presents in the prokaryotic world that has been revealed by the large number of complete genome sequences available in public databases, that have converted comparative genomics in one of the most promising fields in the study of prokaryotic evolution and diversity, and then I describe more deeply the special case of reductive evolution that shapes the genome of intracellular bacteria, both parasitic and mutualistic, and how systems biology have the potential to fill the gap between the information comprised by genome sequences and their projection in the functionality of the complete metabolic systems.

1.1 Prokaryotic genome diversity in the era of metagenomics

Since the sequencing of the first microbial genome, that of *Haemophilus influenzae Rd KW20* in 1995 (Fleischmann et al., 1995), from nowadays, the amount of genome sequences deposited in public databases have increased exponentially each year. The advances produced in recent years in sequencing technologies with the development of ultra-high-throughput sequencing methods like Roche-454 LifeSciences (www.454.com), Solexa-Illumina (www.illumina.com), or ABI-SOLiD (www.appliedbiosystems.com), capable of generate a vast amount of sequences in the order of hundreds of megabases in a single run (Mardis, 2008b; Mardis, 2008a), together with the development of bioinformatic tools and specialized databases for data mining and functional annotation (Medigue and Moszer, 2007), have reduced the time and economic cost of whole genome sequencing in such a way that large sequencing centers like the Sanger Institute of the Welcome Trust or the Joint Genome Institute of the US department of Energy are able to sequence a prokaryotic genome within a day, a fact that have converted

the sequencing of a cultivable organism a routinely task. At March of 2009, there were 959 bacterial genomes completely sequenced, with 2397 more bacterial genomes in a process of ongoing sequencing (http://www.genomesonline.org/gold.cgi).

These advances have generated a change in the scope of prokaryotic genome sequencing projects, from the sequencing of representative isolates culturable bacteria, mainly important human pathogens, to metagenomic studies in which the objective is the sequencing of entire microbial communities of a given environment providing a way to study the structure of microbial communities in terms of species diversity and richness distribution and their functional potential in terms of metabolic capabilities (Tringe and Rubin, 2005; Tringe et al., 2005). These metagenomic studies have allowed to characterize at DNA sequence level organisms and environments that were previously thought to be inaccessible, including obligate pathogens and symbionts that cannot be cultured outside their hosts, environmental microorganisms that are not able to growth in pure culture, or ancient DNA extracted from fossil record. These studies have revealed the enormous variability in terms of prokaryotic diversity and community structure that is present in different natural environments. For example, the analysis of the microbial community structure of acid mine drainage biofilms shows a microbial ecosystem with an extremely low diversity in terms of prokaryotic species, with a dominant bacterial lineage that corresponds to *Leptospirillum* group II followed in abundance by an archaea belonging to a new *Ferroplasma* group II, and with minor presence of three other bacterial lineages. The low species diversity of this environment, explained by their extreme biochemical composition, allows the near-complete assembly of the genomes of the most abundant community members, the bacterium *Leptospirillum* group II and the archaea *Ferroplasma* type II, together with the partial assembly of the 3 other species. In addition, the near completion of the two most abundant genomes alows to carry out metabolic reconstructions that delineates the role of each individual organism in the acid mine drainage environment (Tyson et al., 2004). By contrast, the metagenomic analysis of the microbial ecosystem of the Sargasso sea revealed the presence of at least 1800 individual species in the samples analyzed covering almost all bacterial phyla together with more than 1.2 million genes, highlighting the microbial diversity revealed by environmental metagenomics in comparison with genome sequences obtained from pure microbial cultures (Venter et al 2004). In these complex environments, the genomic sequencing of DNA samples did not result in assembled genomes, and the objective is to characterize the species diversity of the microbial community and to identify and annotate gene families that are important for species survival in the corresponding environment and, comparing results from different samples, identify gene functions that are over or underrepresented in a particular microbial community (Tringe and Rubin, 2005).

The extensive intraspecies variations revealed by metagenomic studies and whole genome projects have also challenged the traditional classification schemes of prokaryotic organisms. The principles of prokaryotic taxonomy arisen along the twentieth century, starting with the use of metabolic and physiological features together with morphological traits determined by light microscopy to characterize bacteria. This constituted the principles of the early classifications such as those found in the early version of the *Bergley´s Manual of Determinative Bacteriology* (Breed et al., 1957). Posterior technical innovations like electron microscopy for examine microbial cells fine structure and DNA sequence-based methods modify the principles of taxonomic classification, and DNA hybridization became the basis for a bacterial species definition that is still used today. This consist on considering as strains of the same prokaryotic species those organisms whose genomes shows a 70% of re-association in DNA hybridization experiments (Wayne et al., 1987; Stackebrandt et al., 2002; Stackebrandt et al., 2002). However, the most important breakthrough in prokaryotic and also eukaryotic taxonomy was the introduction of phylogenetic analysis based on molecular data to study species evolution, which constitutes the basis of molecular evolution discipline that started with the work of Zuckerkandl and Pauling in the 1960´s whom showed that the amino acid sequences of several proteins like cytochrome C and globins were conserved even between distantly related species, proposing also the concept of molecular clock, a relatively constant rate of sequence evolution characteristic of each protein coding gene in the absence of functional change (Zuckerkandl and Pauling, 1965). This culminated with the work of Woese and coworkers, who postulated that the conservation of the sequences of several molecules, mostly ribosomal RNA, in all living organisms allows prokaryotic classification based on sequence similarity by using a universally distributed trait like the noncoding RNA gene of the small ribosomal subunit (Woese, 1987; Woese et al., 1990), leading to a change in bacterial systematics in which classic phenotypic criteria is replaced by criteria based on molecular sequence data. These are the basis of the polyphasic prokaryotic classification system employed in the *Bergley´s Manual of Systematic Bacteriology*, a modified version of the determinative manual referenced above, that is considered the best approach to bacterial classification widely accepted among microbiologist and that is based on the phylogenetic analysis of small subunit rRNA genes (16S rRNA) complemented with classical microscopically and biochemical observations (Brenner et al., 2005). In order to cope molecular sequence data with traditional taxonomy, a prokaryotic species is defined as a group of strains characterized by some degree of phenotypic consistency, with 70% of DNA-DNA binding in reassociation studies and with at least 97% of identity at 16S ribosomal RNA sequence level (Stackebrandt et al., 2002). This means that if genomic DNA of an unknown strain shows less than 70% of hybridization with the genomic DNA of the reference strain of a given species, can be named as a different species if it is accompanied by a description of as many phenotypic properties as possible for their precise delineation in what has been named a polyphasic approach to species definition (Vandamme et al., 1996; Gevers

et al., 2005).However, DNA hybridization approach have several disadvantages, in the sense that the 70% threshold is artificially derived to cope with the traditional prokaryotic taxonomy derived exclusively from phenotypic characters without taking into account any evolutionary consideration. In addition, it relies on the availability of a single reference strain to compare that has to be maintained in pure cultures, so the prokaryotic diversity in natural environments cannot be characterized by this approach. These disadvantages together with the advent of 16S ribosomal RNA gene sequencing convert this last methodology as the primary technique for bacterial identification, being widely used to characterize prokaryotic diversity in environmental microbiology studies (DeLong and Pace, 2001) or to assign unculturable organisms to new species (Hugenholtz et al., 1998). However, although 16S gene sequencing is very useful to assign strains to species level, is of marginal value at the genus level and of no value above the genus level due to the absence of enough sequence divergence at these taxonomic levels (Brenner et al., 2005). In addition, their specific character as slowly evolving molecule makes that it lacks enough resolution power to clearly distinguish strains belonging to related species (Fox et al., 1992). Another problem associated with bacterial classification based on single genes, including 16S sequences, is the phenomenon of homologous recombination among similar species, which results in the replacement of a segment of the chromosome of the recipient bacterium with the corresponding fragment from a different strain of the same species or from closely related species. Homologous recombination is believed to be common among bacteria, although their efficiency decreases when increases sequence divergence (Feil and Spratt, 2001). However, there is natural evidence that replacements take place between species differing at 5-25% in their nucleotide sequences (Hanage et al., 2006a; Hanage et al., 2006b).

The described difficulties associated with DNA-DNA hybridization approach and the limits of 16S based classification to clearly discriminate between closely related species, together with the difficulties associated with any single-gene similarity approach due to the role of horizontal gene transfer and recombination masking phylogenetic signal, have led to the development of multigene-based approaches for bacterial classification. These approaches can be used to characterize microbial populations at intraspecific level based on allelic profiles of different individuals defined from a small set of housekeeping genes, a methodology known as multilocus sequence typing (Cooper and Feil, 2004; Maiden, 2006). A modification of this approach for species definition has been proposed by Gervers named multilocus sequence analysis (MLSA) based on the phylogenetic analysis of the concatenated nucleotide sequences of several marker genes (Gevers et al., 2005). MLSA has been used to establish phylogenetic position of new species (Christensen et al., 2004; Holmes et al., 2004), but also to analyze species limits among large numbers of strains of closely related species, like a work of Godoy and collaborators, where MLSA of a large number of strains corresponding to *Burkholderia pseudomallei*, *B. mallei*, and *B. thailandensis*, initially typed by

MLST, shows a clear separation of all strains corresponding to *B. pseudomallei* and *B. thailandensis* in two different clusters supporting their separation in different species despite their low divergence at sequence level, whereas strains corresponding to *B. mallei*, all identical by MLST, clusters together with strains of *B. pseudomallei* (Godoy et al., 2003). This is a case of "species within species" in which strains of different species appears clustered together by MLSA but have very different ecological profiles that justify their classification in different species. This is also reflected in their large differences in their genome sizes, with *B. mallei* having a genome one megabase smaller than *B. pseudomallei*, being associated with a narrower ecological niche as obligate parasite of horses, mules and donkeys with no other known natural reservoir (Godoy et al., 2003; Nierman et al., 2004). MLSA have provided other similar examples of bacterial species considered as such because their association to different human or animal diseases that at MLSA are "clones" with distinctive ecology and biology derived from other "mother" species, like different strains of species from the *Bacillus cereus* complex (*B. anthracis*, *B. thuringensis*, *B.cereus*) that appears grouped in 8 different lineages by MLSA (Priest et al., 2004).

The emergence of whole genome sequencing and the availability of completely sequenced genomes from closely related species and even from different strains of the same species allowed to study the genetic and functional relationships between organisms at the whole-genome level, and provided a novel approach to compare evolutionary relationships inferred from 16S ribosomal RNA, MLSA, or phenotypic characters with relationships inferred from whole genome data. This was carried out by Konstantinidis and Tiedje in 2005 through gene content comparisons between completely sequenced genomes using two different indices of genome similarity, average nucleotide identity (ANI) and average of amino acid identity (AAI) among shared genes between two genomes (Konstantinidis and Tiedje, 2005b; Konstantinidis and Tiedje, 2005a). Both measures of genome similarity shows great correlation with distances inferred from 16S rRNA sequence data, and in the case of ANI, also shows a great correlation with DNA hybridization values reported in the literature, estimating that a value of ANI of 94% corresponds to the traditional 70% of DNA-hybridization experiments for species definition, being a robust measure of the genetic and evolutionary relatedness between closely related strains. In contrast, AAI was a robust descriptor of genetic relatedness at higher taxonomic levels, with phylogenetic trees inferred from AAI values being very congruent in terms of tree topology with traditional phylogenies based on concatenate sequence data. However, the dynamic nature of bacterial genomes is also reflected in this analysis by the fact that there is extensive overlap between taxonomic ranges for a fixed AAI value (e.g. a given AAI value detected within bacteria and between bacteria and archaea), whereas strains with an ANI values above of 94% cutoff for species definition shows differences in gene content of 5 to 35% of the genomes. This reveals a probable continuum of genetic diversity in the prokaryotic world that is not reflected

by clear boundaries fixed by traditional taxonomy, and postulates that a more natural definition of bacterial species should be flexible enough to take into account the ecology of the organisms, in order to better predict the phenotype and ecological properties of the species (Konstantinidis et al., 2006).

This dynamic nature of prokaryotic genomes in terms of genome sizes and gene content at all taxonomic levels was one of the most important features revealed by comparative genomics as soon as the first completely sequenced genomes were available. Nowadays, with 959 completely sequenced bacterial genomes, there is a range of genome size variation that spans from the 0.16 kilobases of the obligatory endosymbiont of psyllids *Carsonella rudii* (Nakabachi et-al 2006) to the 13.33 megabases of the free-living mixobacterium *Sorangium cellulosum* (Schneicker et-al 2007). An analysis of the distribution of genome sizes among different bacterial taxonomic groups shows that species with large and small genome sizes coexist in taxonomic lineages like the γ-proteobacteria or the firmicutes (Silva and Latorre 2008), and these variations in genome sizes are clearly related to the ecological behavior and environmental niche of the bacterial species. The smallest prokaryotic genomes tends to belong to organisms restricted to stable ecological niches with minor fluctuations in their environmental conditions, often in association with a eukaryotic host in different types of ecological associations that range from mutualistic symbiosis to pathogenic interactions (Ochman and Moran, 2001; Andersson et al., 2002; Mira et al., 2002; Moran, 2002), whereas genomes at the opposite end of genome size range corresponds to organisms that lives in highly complex and variable environments such as soil (Casjens, 1998; Bentley and Parkhill, 2004; Fraser-Liggett, 2005). However, despite the wide range of genome sizes, prokaryotic genomes show similar levels of coding density, with approximately one gene per kilobase of DNA in contrast to what is observed in eukaryotic genomes, so there is a clear correlation between genome size and gene content in prokaryotic genomes (Figure 1.1). Bacterial genome size is consequence of the sum of different genetic events like gene duplication and horizontal gene transfer, that leads to increases of gene content, and lineage-specific gene loss, that generates reduction in genome sizes, and is not a good indicator of evolutionary lineage due to the great variability observed at all taxonomic levels (Bentley and Parkhill, 2004; Koonin and Wolf, 2008). In this context, there is a significant difference in the functional gene categories that are over and underrepresented with increasing genome sizes. This was revealed in a comparative study carried out by Konstantinidis and Tiedje in which whole genome comparizons of 115 prokaryotic organisms revealed that larger genomes were significantly enriched in genes involved in regulation and secondary metabolism, that are almost absent in small genomes of intracellular symbionts and pathogens (Konstantinidis and Tiedje, 2004). This is associated with a broad metabolic diversity accompanying to bacterial species with large genomes, that explains also the increase in the proportion of regulatory genes that are needed for a successful control of their metabolic activities

under different environmental conditions, although an unanticipated complexity in the transcriptome of *Mycoplasma pneumoniae* has been revealed by transcriptomics despite their highly streamlined genome and their low number of regulatory genes (Guell et al., 2009). In contrast, genes involved in informational processes like protein translation, DNA replication, or cell division are significantly depleted in larger genomes because most of these processes are encoded by a fixed number of genes, so their proportion diminishes with increased genome size (Konstantinidis and Tiedje, 2004). This indicates that genome size and content appears largely dictated by environmental pressures, and that species living in complex environments with highly fluctuating conditions tends to have highly versatile metabolism with large number of regulatory elements in order to be able to adapt to changing conditions, whereas these genes are almost absent in small genomes because these organisms are associated with highly stable environments like the intracellular cytoplasm of eukaryotic host cells in intracellular bacterial endosymbionts and pathogens.

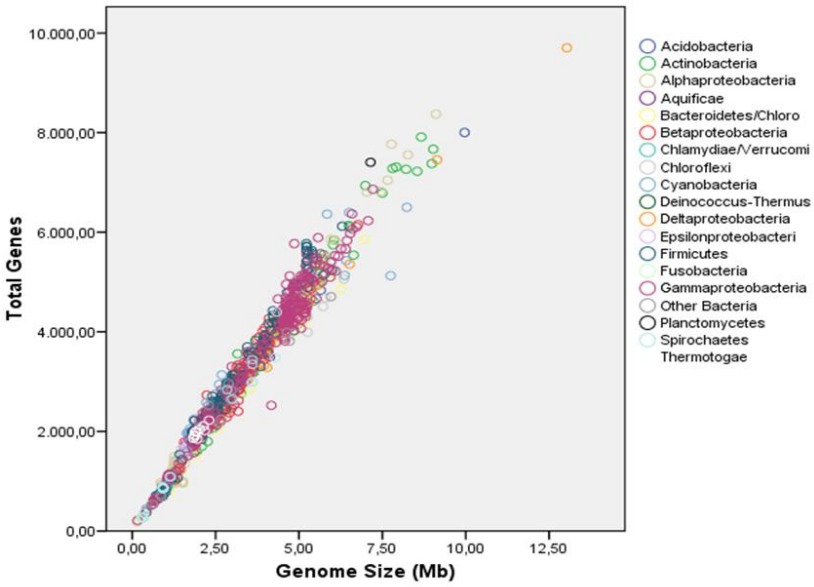


***Figure 1.1**: Graphical relationship between genome size and gene content of the 959 bacterial genomes sequenced up to Juny of 2009. Each color represents different species of the same taxonomic group following taxonomic classification in the NCBI*

However, genome size variation is not restricted to higher taxonomic levels, and significant variability is also observed even between different strains of the same

species. This was revealed as soon as the genomes of closely related organisms were sequenced and their gene complements were compared, like with the comparison of laboratory strains of *Escherichia coli* with uropathogenic and enterohaemorraghic strains, that revealed that gene content differs by more than 30% between pairs of genome sequences, with less than 40% of the total gene complement being shared by these three strains (Welch et al., 2002). In many cases, strain-specific genes allows bacterial strains to exploit a different environment or ecological niche than their close relatives, like differences in gene content among strains of photosynthetic soil bacteria *Rhodopseudomonas palustris* that allows the colonization of different microenvironments with different conditions of light, oxygen, and nutrient availability (Oda et al., 2008). This is also the case of differences among high-light and low-light adapted ecotypes of marine photosynthetic bacterium *Prochlorococcus marinus* consequence of differences in genes involved in adaptation to different light intensities together with strain-specific genes involved in cell surface features (Kettler et al., 2007). However, all cellular organisms share common mechanisms for genetic information transmission and expression as well as similar metabolic and catabolic capabilities to obtain energy from a limited set of sugars and amino-acids, and these similarities are also reflected in the gene content of genome sequences. Therefore, in an individual genome it is possible to distinguish a core of essential genes named the *endogenome* that is responsible for basic functions of cell metabolism (replication, transcription, translation, basic biosynthetic and catabolic capabilities) and that is shared by almost all closely related genomes, and the *exogenome*, that is composed by those genes responsible of the uniqueness of a given organism (Casjens, 1998). The *exogenome* can be also divided in genes responsible for ecological and phenotypic properties of a given strain like antibiotic-resistance genes that can be of great value under particular environmental circumstances, and genes without any beneficial effect on the organism such as parasitic elements like transposases or phage genes.

The concept of a genomic core of genes conserved across different organisms is an important research topic in comparative and evolutionary genomics due to their implications in the reconstruction of ancestral genomes and minimal cell reconstructions, and has been also widely used in phylogenetic analysis to overcome limitations associated with phylogenetic inferences based on single genes. Its characterization started as soon as the first complete bacterial genomes were available, with the pioneering work of Musheguian and Koonin in 1996 comparing the gene complements of *Haemophilus influenzae* and *Mycoplasma genitalium* in which a core set of 256 genes shared by the two genomes were identified, being postulated as essential gene functions that must be present in an hypothetical minimal genome due to their conservation across broad evolutionary range (Mushegian and Koonin, 1996). However, posterior approaches including more genomes have revealed that the size of the core set is highly dependent on the number of genomes compared and their evolutionary distance. For prokaryotic

species for which genomes of several strains have been completely sequenced, the identification of orthologous genes corresponding to the species core set is relatively straightforward due to their relative conservation at both sequence level and genome position (Welch et al., 2002; Gil et al., 2003; Fuxelius et al., 2007; Stinear et al., 2008), but with more divergent species, the core becomes progressively smaller and more elusive because of high levels of sequence divergence, high rearranged genomes, and problems in recognizing paralogy (Koonin, 2000; Charlebois and Doolittle, 2004). The paucity of truly universal genes is consequence of the combined action of several evolutionary events that shapes the evolution of bacterial genomes. First, at long evolutionary distances, orthologous genes might have accumulated so many differences that their homology is no longer detectable by computational approaches based on sequence similarity. Second, each lineage may have adopt new genes or molecular strategies to accomplish the same cellular functions by means of horizontal gene transfer, generating functional homologies non-detected by sequence comparisons in a process known as non-orthologous gene displacement (Koonin et al., 1996). Finally, massive gene loss in the evolution of the genomes of parasite and endosymbiotic bacteria generates highly streamlined core gene sets when these reduced genomes are considered. However, despite these difficulties, there is so much interest in the reconstruction of deep core sets, specially the universal core genome comprising bacteria, archaea and eukarya, because it may serve as initial approach to the reconstruction of the gene composition of the last universal common ancestor (Lazcano and Forterre, 1999; Kyrpides et al., 1999; Ouzounis et al., 2006). From a phylogenetic point of view, the identification of core genes conserved by distantly related organisms that have evolved by vertical inheritance from a common ancestor are useful in the resolution of deep nodes of phylogenetic trees that are difficult to be solved by single gene approaches, allowing phylogenomic approaches to the study of bacterial evolution that allows more precise characterization of the evolutionary relationships among taxonomic groups and better assesment of different evolutionary process like horizontal gene transfer that determines bacterial gene content (Makarova et al., 1999; Nesbo et al., 2001; Daubin et al., 2002; Lerat et al., 2003;).

In contrast with the concept of core genome or endogenome, the "accessory" or "auxiliary" set of genes that compose the exogenome contains both genes presents in a subset of the compared genomes and strain-specific genes unique of a single genome. The exogenome is responsible of the species diversity at genome level because might encode supplementary biochemical pathways and functions that are not essential for bacterial cell growth but which confer selective advantages in terms of adaptation to a different ecological niche, antibiotic resistance or colonization of a new host (Medini et al., 2005; Abby and Daubin, 2007; Tettelin et al., 2008). This plasticity in genetic repertories of closely related genomes is consequence of different processes of gene gain and gene loss that shapes the gene complement of a given strain and that makes single genomes as incomplete representatives of the total

gene complement of prokaryotic species; a genome sequence provides the gene complement of a particular prokaryotic strain adapted to a given environment or ecological niche, but this strain is only a single representative of their corresponding species, that can present high levels of variability in terms of genome sizes and gene content that can only be revealed by the comparison of their complete genome sequences. Differences in gene complements among closely related strains allow the colonization of different ecological niches as consequence of different metabolic capabilities associated with different gene inventories in each genome, but also are consequence of the proliferation of different types of mobile genetic elements like insertion sequences, prophage elements, or different genomic islands consequence of horizontal gene transfer events, one of the main evolutionary forces shaping prokaryotic evolution (Figure 1.2).

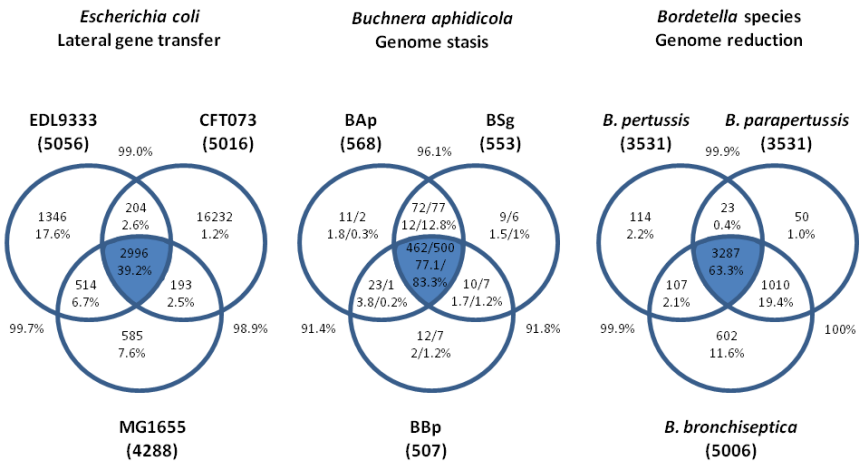


Figure 1.2: *Comparison of gene content in three bacterial lineages under different evolutionary preassures using "Venn diagram" representation. The lineage pan-genome is encompassed within the Venn diagram. The numbers inside the diagram boxes represents the number of genes (and the percentage of the total) found to be shared among the indicated genomes. Core genome of each lineage is represented in the shaded region of the diagram. Numbers outside the diagram represents the total gene number of each genome and the percentage identity at 23S genes. For Buchnera, the numbers reflect analysis with/whithout pseudogenes. Adapted from Lawrence and Hendrickson 2005.*

Based on this genomic plasticity, it have long speculated with the existence of pools of genetic material available to different organisms living on a particular ecological niche, a concept known as "pan-genome" (Lawrence and Hendrickson, 2005; Lapierre, 2008). The pan-genome is defined as the total gene repertoire of a specific group of organisms, normally a set of strains of a given species, and includes all gene families present in the genomes of the group, both "core" gene

families present in all compared genomes and strain-specific genes present in a single genome. One of the first attempts to define the pan-genome of a specific species was performed on eight strains of *Bacillus anthracis* and eight strains of the group B of *Streptococcus agalactiae* by Tettelin and collaborators in 2005 in order to answer the question of how many genomes are needed to fully describe the gene complement of a bacterial species (Tettelin et al., 2005). For *S. agalactiae* a pan-genome of 2713 genes was characterized, composed of a "core" set of 1806 genes and an "auxiliary" set of 907 genes. Each strain contains an average of 1806 genes common to all the other strains and an additional set of 439 "auxiliary" genes that are absent in one or more strains. Mathematical modeling based on the eight compared genomes reveal that unique genes will continue to emerge as even hundreds of new genomes are added at an estimated rate of 33 new genes with each newly sequenced strain, leading to what is known as "open" pan-genome in which a very large number of genome sequences would be needed to completely characterize the entire gene repertoire to which the species have access. In contrast, in species like *B. anthracis*, the number of new genes added to the pan-genome was found to rapidly converge to zero after the addition of only four genomes, forming a "closed" pan-genome in which four genome sequences are enough to completely characterize this species (Tettelin et al., 2005). The differences in the size and content of pan-genomes between species reveal differences in their ecological profiles, with open pan-genomes being typical of species that colonize multiple environments and have multiple ways to exchange genetic material like the *Streptococci*, *Meningococcal*, *Salmonellae* and *Escherichia*. The increases of pan-genome size at each added genome reveals a vast amount of genomic diversity that remains undiscovered (Lefebure and Stanhope, 2007; Schoen et al., 2008; Rasko et al., 2008). By contrast, species living in more restricted and stable environments with limited access to the global microbial gene pool and with low capability to acquire foreign genes tends to have a closed pan-genome, like strains of *B. anthracis*, *Mycobacterium tuberculosis*, or *Chlamydia trachomatis*, being an extreme example of this genomic stability the situation of maternally transmitted endosymbionts like the primary endosymbiont of aphids *Buchnera aphidicola*, with an extremely stable closed pan-genomes reflecting their isolated niche in their host environment with little opportunity for lateral acquisition of foreign DNA (Tamas et al., 2002; Medini et al., 2005), reflecting the existence of a fossil gene content that was present in the last common ancestor of the *B. aphidicola* strains (vanHam et al 2003) with differences between strains due to a few gene losses, rearrangements involving plasmids and chromosome and gene duplications (Silva et al 2003). Recently, the concept of the pan-genome was also applied to the bacterial branch of the tree of life by Lapierre and Gogarten analyzing 573 completely sequenced bacterial genomes and concluding with the infinite character of bacterial pan-genome due to the massive presence of genes presents in a specific group of genomes together with genome-specific genes, with genes present in almost all bacterial genomes representing only the 8% of a typical bacterial genome (Lapierre and Gogarten, 2009). One of the

main contributors to this high level of genome plasticity observed between prokaryotic genomes are differences in the umber and types of mobile genetic elements like bacteriophages, transposons, or genomic islands,

1.2 Mobile genetic elements as mediators of genome plasticity

A significant fraction of the variable pool of genetic material of prokaryotic genomes corresponds to different types of mobile genetic elements, which constitute the most important source of genomic variability in prokaryotic evolution. Mobile genetic elements are defined as segments of DNA that encode enzymes and other proteins that mediate their own transfer both within genomes and between bacterial cells. The set of mobile genetic elements of a genome has been designed as mobilome (Frost et al., 2005; Binnewies et al., 2006), and consist basically of insertion sequences and transposons, bacteriophages, plasmids, and genomic islands. Traces of the activity of mobile genetic elements are evident in all prokaryotic genomes at different levels. The existence of different copies of a given IS element or prophage gene in a genome gives rise to the possibility of generate chromosomal rearrangements by homologous recombination between copies of the repeated element. These types of chromosomal rearrangements by homologous recombination are strongly dependent of the degree of sequence homology between recombining sequences, and in the case of IS elements, their high levels of sequence similarity between copies of the same genome favors this kind of rearrangements by homologous recombination. The orientation of the repeated sequences determines the type of rearrangement that takes place; if repeated sequences are in the same orientation, recombination leads to the deletion of the DNA segment comprised between them, whereas if they present inverted orientation, recombination leads to the inversion of the DNA segment comprised between them (Mahillon and Chandler, 1998; Chalmers and Blot, 1999; Brussow and Hendrix, 2002; Frost et al., 2005). This has been observed for example in whole genome comparisons of different strains of the plant pathogen *Xylella fastidiosa*, in which five of the six inferred recombination sites responsible for three genome inversions were located on duplicated prophages, whereas in the closely related species *Xanthomonas campestris*, different types of insertion sequences proliferates in different strains (Van Sluys et al., 2003; Monteiro-Vitorello et al., 2005). Another nice example is found in whole genome comparisons of Japanese and American strains of *Streptococcus pyogenes*, in which a large genome inversion is consequence of homologous recombination across two prophage sequences (Nakagawa et al., 2003). However, most of genome rearrangements are eliminated by natural selection due to their detrimental effect over organismal fitness because they are able to eliminate essential genes or alter the gene expression pattern of the genes included in the rearrangement, being purged by natural selection. In addition, mobile genetic

elements are vectors for the horizontal transfer of genes, being responsible for many specific activities of a given strain or bacterial species. This occurs when mobile genetic elements carry genes encoding for specific activities that can mediate ecological adaptation, like antibiotic-resistance genes, or genes associated with pathogenicity or bacterium-host interactions. This is the case of the ability of nitrogen fixation of plant symbiotic bacteria of the genera *Rhizobium*, in which large symbiosis islands carry genes responsible for plant root-cell invasion and nitrogen fixation to ammonia (Sullivan and Ronson, 1998). In many cases, the difference between a harmless commensal or soil bacterium and a deadly pathogen resides in the presence of a toxin-encoding plasmid, like in the case of the choleric toxin phage CTXφ that carries the cholera toxin genes responsible for the virulent character of *Vibrio cholerae* (Waldor and Mekalanos, 1996; Karaolis et al., 1999). Cyanophages of photosynthetic marine cyanobacteria of the genera Prochlorococcus are also responsible for photosynthetic capabilities of these bacteria that allow their surveillance under nutrient-poor conditions (Sullivan et al., 2005). In the next sections, I describe the principal features of the three main groups of mobile genetic elements that resides in prokaryotic genomes (genomic islands, transposons and insertion sequences, and bacteriophages), specially focused on the different forces that govern their origin and evolution, together with their impact in the evolution of different prokaryotic lineages.

1.2.1 Genomic Islands

Genomic islands (GIs) are segments of DNA between 10 and 100 kilobases in length that harbors phage or plasmid related sequences like integrases or insertion sequence elements involved in their horizontal transmission between prokaryotic organisms. These blocks of DNA are most often inserted inside tRNA genes, a feature shared with many bacteriophages, and may be unstable due to the presence of flanking direct repeats frequently homologous to phage attachment sites that promotes their integration and excision from the prokaryotic genomes (Hacker et al., 1997; Buchrieser et al., 1998). However, in addition to genes involved in their motility, GIs also carry gene clusters encoding specific functions that modify the phenotype of their prokaryotic host in different ways (Hacker and Carniel, 2001). GIs were first identified in uropathogenic strains of *E. coli* as large and unstable genomic regions that contains virulence associated genes like fimbrial adhesins or haemolysins that are absent from closely related non-pathogenic strains. By such motif, they were initially designated as pathogenicity islands (PAIs) (Knapp et al., 1986; Hacker et al., 1990). However, the study of larger number of genomes have revealed that similar structures are present in non-pathogenic bacteria encoding gene functions useful for the survival and transmission of their prokaryotic hosts, leading to the inclusion of PAIs as a subgroup into the much broader group of GIs (Hacker and Kaper, 2000; Hentschel and Hacker, 2001). In addition to the presence of integrases and transposases that mediates their transmission, these GIs share some

sequence and structural features consequence of their mobile character and their prevalent expansion by horizontal gene transfer that allows their distinction from the rest of the prokaryotic genome. These characteristics are a biased sequence composition in comparison to the rest of the genome in terms of GC content, oligonucleotide frequencies, or codon usage of the encoding genes that allows their identification by computational methods (Yoon et al., 2005; Vernikos and Parkhill, 2006; Vernikos and Parkhill, 2008). From an evolutionary point of view, the acquisition of these GIs confers a selective advantage to the prokaryotic organisms under specific environmental conditions, and as consequence, genomic islands increasing the fitness of their prokaryotic hosts can be also named fitness islands (Preston et al., 1998). These GIs or fitness islands can be subdivided in different subgroups depending on the lifestyle of the prokaryotic host rather than by their gene composition. For example, GIs that helps prokaryotic organisms to live in a given environment can be considered as "ecological islands", whereas GEIs involved in the survival of prokaryotic organisms as saprophytes in a host can be considered as "saprophytic islands"(Hacker and Carniel, 2001). In the context of host-dependent bacteria, can be also defined "symbiotic islands" as GIs that helps symbiotic bacteria to positively interact with their eukaryotic hosts, and the originally defined PAI as those GIs that enhances the virulent phenotype of their pathogenic host (Hentschel and Hacker, 2001; Dobrindt et al., 2004). This classification is not exclusive, and a given GI may act as a ecological islands when the prokaryotic host is an environmental bacteria and as pathogenicity island when is harbored by a pathogenic bacteria, like in the "high pathogenicity island" (HPI) of virulent species of the genera *Yersinia*, that is also present in other non-pathogenic enterobacteria increasing their survival in iron-limited environments(Carniel et al., 1996; Bach et al., 2000). In other cases, GIs may increase the adaptability of a prokaryotic organism under certain environmental conditions, allowing exploiting novel ecological niches: For example, a metabolic island in *Salmonella seftenberg* involved in the transport and metabolization of sucrose increases the metabolic versatility of the bacteria (Hochhut et al., 1997). GIs may also carry genes involved in antibiotic resistance, like the "SCCmec" islands presents in different strains of *Staphylococcus aureus* conferring resistance to methicillin, enhancing their survival both in soil environments with other antibiotic-producing organisms and in hospitals with strong antibiotic pressure (Ito et al., 2001; Hiramatsu et al., 2002).

In the context of prokaryotic organisms that lives in close associations with eukaryotic hosts, many of these bacteria, both endosymbiotic and pathogenic, depend on different types of secretion systems that allow the bacterial attachement to eukaryotic host cells, their invasion, and the interaction with host cell activities through different sets of effector proteins that are secreted through bacterial membrane to the host cell cytoplasm in a molecular "crosstalk" that allows bacterial survival in the intracellular host environment (Hueck, 1998; Pugsley et al., 2004; Galan and Wolf-Watz, 2006; Gerlach and Hensel, 2007). In pathogenic bacteria,

PAIs encoding type III and type IV secretion systems are responsible of the interaction with eukaryotic host cells and the delivery of different sets of effector proteins involved in the pathogenic interaction like toxins, invasins, adhesins, or modulins that interferes with host-cell functions (Hueck, 1998; Dobrindt et al., 2004). These PAIs encoding secretion systems has been extensively studied in bacterial pathogens like *Salmonella enterica*, where two different type III secretion systems are involved in different stages of the host cell infection, with one required for the initial interaction and invasion of epithelial host cells whereas the other is required for the systemic infection once the bacteria has gained access to the host cell cytoplasm (Ochman and Groisman, 1996; Ochman et al., 1996; Galan, 1999; Galan, 2001; Kuhle and Hensel, 2004). These two PAIs encoding type III secretion systems are also present in other non-pathogenic bacteria like *Sodalis glossinidius*, which retains both type III secretion systems with the same structure and conserving most of the effector proteins as in *S. enterica* (Dale et al., 2002; Dale et al., 2005). *S. glossinidius* is one of the main research points of this thesis, and represents a nice example of the plasticity of GIs depending on the ecological context of their bacterial host, given that the same type III secretion systems that in *S. enterica* are involved in pathogenic associations with eukaryotic hosts, in non-pathogenic symbiotic organisms like *S. glossinidius* acts as ecological islands that favors the survival of the symbiotic organism through analogous processes.

1.2.2 Bacteriophages

Bacteriophages are viruses that infect bacteria, and are considered the most abundant and diverse entities in nature, with more than 10^{30} tailed phages estimated in the biosphere, typically outnumbering prokaryotic cells by 10-fold in environmental samples (Ashelford et al., 2003; Edwards and Rohwer, 2005; Comeau et al., 2008). They are not an homogeneous group, and are classified on the basis of their genome type (as double or single stranded DNA or RNA) and their structural morphology into 13 different phage families, with great diversity on genome sizes that ranges from 2.3 kilobases to more than 300 kilobases (NCBI prophage database July 2009). The minimal genome of a tailed phage encodes the genes needed for DNA packaging, head and tail fiber biosynthesis, DNA replication and transcription regulation, together with lytic genes responsible for bacterial cell death by phage induction. As the phage genome increases in size, the virion morphology becomes more complicated and the phage interferes with more cellular activities (Brussow and Hendrix, 2002; Canchaya et al., 2003). About 96% of all bacterial viruses correspond to double stranded DNA tailed bacteriophages, which can be divided into lytic and temperate depending on their lifecycle inside bacterial cells. Lytic bacteriophages multiply inside the bacterial cells without integrating in the chromosome, generating a progeny of virions that will kill the cell, being released to extracellular media and invading other bacterial cells. In contrast, temperate bacteriophages, although they are able to propagate lytically under certain

conditions, can be also integrated inside bacterial chromosome as prophage elements that are replicated in concert with the bacterial chromosome. During this association, lytic genes whose activation would be detrimental to the bacterial host are not expressed, whereas other prophage genes named lysogenic genes are able to alter the phenotypic properties of the host bacterium in different ways, since protecting against further prophage infections to increasing the virulence of a pathogenic host. Under certain environmental conditions, stably integrated prophages can experiment a process known as induction by which lytic genes are activated generating bacteriophage virions that will be released to the extracellular media after bacterial kill (Casjens, 2003; Brussow et al., 2004). Whole genome sequence analysis has revealed that a large proportion of bacterial genomes carry prophage elements integrated inside their genome sequence, that in some cases represent until 20% of the prokaryotic genome and more than 50% of the strain-specific DNA in several important pathogens(Casjens, 2003; Hatfull, 2008), although the proportion of these retaining lytic activity is unclear. From an evolutionary point of view, prophage genomes seems to be only transient sequences on bacterial chromosomes that leads ultimately to the death of bacterial host cells because of their lytic activity despite any beneficial effect that can introduce in the short term, and as consequence different mutational events will lead to the inactivation of prophage genes and their ultimate disappearance by deletional events (Lawrence et al., 2001; Canchaya et al., 2004). Several observations supports this view, like the inability of lytic induction of most prophages of enterohaemorraghic strains of *E. coli* (Brussow et al., 2004), the absence of lytic activity in all prophage sequences of two different species of *Lactobacillus* (Ventura et al., 2003), or the frequent observation of prophage remnants in bacterial genomes that can be easily explained by an ongoing process of gene decay (Lawrence et al., 2001). However, this prophage inactivation is not a universal process and it has been also detected bacterial species in which all their prophage sequences could be induced (Banks et al., 2003).

Bacteriophages play also a key role in the evolution and virulence of many bacterial pathogens and are important vehicles of horizontal gene exchange between different bacterial species, being responsible for a large proportion of strain specific sequences of bacterial genomes. For example, whole genome comparison between enterohaemorraghic *E.coli* 0157:H7 strain Sakai with the laboratory strain *E. coli* k12 revealed a conserved backbone of 4.1 megabases between both genomes together with a large amount of strain-specific sequences that in 0157:H7 strain includes 18 complete and partial prophage elements that constitutes 12% of the whole genome sequence (Ohnishi et al., 2001), whereas in *Streptococcus pyogenes* M3 MGAS 315 six prophage elements constitutes about 12% of the chromosome (Beres et-al 2002). Comparative prophage genomics have also revealed an enormous genetic diversity in bacteriophage populations, with almost no sequence similarity between bacteriophages from non-overlapping host ranges (Beres et al., 2002), and even with many bacteriophages within the same bacterial host range sharing little or

no sequence similarity. This diversity together with the limited number of prophage genome sequences available compared with the enormous phage diversity revealed by viral metagenomic analysis suggests that a significant proportion of bacteriophage diversity remains unexplored (Breitbart et al., 2002; Hatfull, 2008).

However, despite the great sequence diversity, bacteriophages belonging to the same family share common genome organization and transcription patterns, and similar lifestyles. For example, in phages with siphoviral morphotypes the genes encoding for head and assembly proteins share a clear synteny, arranged together with head genes in 5' to the tail genes, and this genomic structure is conserved in phages with no sequence similarity at both nucleotide and amino acidic level, probably reflecting an ancestral genomic feature in bacteriophage evolution indicative of a very ancient divergence (Casjens, 2005). Comparisons of prophage genomes have also revealed an extensive mosaicism in their genome structure, with different gene modules that shows different evolutionary origins within the same prophage genome (Hendrix et al., 1999; Pedulla et al., 2003). This mosaic structure constitutes the basis of a modular theory of phage evolution proposed by Suskin and Botstein in which homologous recombination between related prophages at repeated linker sequences flanking genetic modules would be responsible for the module exchange, like short regions of sequence similariry between gene modules in lamboid coliphages (Susskind and Botstein, 1978; Botstein, 1980; Clark et al., 2001). Posterior genomic analyses have revealed that non-homologous recombination events have also taken place at nearly random points across the genome, generating a wide range of prophage genomic structures, most of which are non functional as consequence of recombination events that take place within coding regions. Natural selection will purge these non-functional prophages leaving the gene-boundary recombinants in which recombination breakpoints have taken place in intergenic regions do not disrupting functional modules (Juhala et al., 2000; Hendrix, 2002). In addition, comparative genomics has also revealed the presence of genes of unknown function within conserved genetic modules that are typically flanked by their own promoter and terminator sequences in an arrangement that ensures their autonomous replication, even from a repressed prophage. These genetic elements have been named morons and, in many cases, their nucleotide composition differs substantially from that of the adjacent genes, arguing for a recent acquisition of the moron from an outside source (Brussow et al., 2004). For many of these morons, their precise genetic function is unknown, although there is other cases in which the moron is expressed from a repressed prophage providing functions that appear beneficial for the bacterial host cell, like morons encoding virulence factors that change the phenotype or fitness of the bacterial host, allowing adaptation of pathogens to new hosts or the emergence of novel pathogens or epidemic clones (Waldor and Mekalanos, 1996; Mirold et al., 1999; Mirold et al., 2001; Akerley et al., 2002). Under these circumstances, the presence of the moron increases the maintenance of prophage sequence by increasing host fitness. In other

situations, moron sequences increases the phage fitness increasing their lytic cycle functions; in this context, a moron sequence will be retained in phage genome because it increases the fitness of the phage during the lytic growth providing an alternative view of prophage evolution by moron-accretion in which prophages evolve by gradual acquisition of morons, which provide selective benefits to the prophage genome. In this context, it is possible that essential phage regions like genes encoding integrases, lysozymes, or immunity regions could originally entered the genome as morons that becomes stably integrated into the genome of the prophage during its evolution (Hendrix et al., 2000).

1.2.3 Transposable elements (transposons and insertion sequences)

Transposable elements occur naturally in nearly all species of prokaryotes. They are defined as specific DNA segments that can repeatedly insert into one or more sites in one or more genomes (Campbell et al., 1979b; Campbell et al., 1979a). Two different types of transposable elements have been defined in bacteria, transposons and insertion sequences. Transposons are DNA segments that contain genes involved in their transposition together with other genes involved in different functions, like antibiotic or heavy metal resistance genes (Campbell et al., 1977; Blot, 1994). Insertion sequences (IS) are the simplest mobile genetic elements in prokaryotic genomes; are small sequences that ranges in size from 600 to more than 3000 base pairs and contain one or more open reading frames (ORFs) coding for the transposase proteins needed for their own transposition; the whole IS element are delimited by terminal inverted repeats that contain sequences that are recognized and bound by the transposase in the first steps of the transposition reaction. This structure appear flanked by short direct repeats of the insertion site that are generated by the IS element during the transposition process (Mahillon and Chandler, 1998; Chalmers and Blot, 1999). IS elements are classified into 20 different families on the basis of various shared features. An IS family is defined as a group of IS elements with related transposases with strong conservation of their catalytic sites, with conserved structural organization, and with similar inverted repeats (Siguier et al., 2006a). Transposition activity of prokaryotic IS elements is often tightly regulated and takes place at very low levels, usually restricted to *cis* activity in which the transposition of a given IS element can be only promoted by their own transposase, in contrast to what occurs in eukaryotes, where many non-functional copies of transposable elements can passively proliferate in *trans* activity using transposases from functional copies (Nagy and Chandler, 2004). In addition to their own activity as independent mobile DNA, two IS elements can form composite transposons that mobilizes the genomic segment comprised between them. Examples of composite transposons includes Tn10 and Tn5, which are mobilized by copies of IS10 and IS50 respectively and are important mediators of the mobilization of antibiotic resistance determinants (Kleckner et al., 1975; Berg, 1977; Reznikoff, 1993). Their mobilization across the prokaryotic genome can proceed by

two different mechanisms. In replicative transposition, the transposable element is copied and transposed to a novel genomic site, increasing the copy number, whereas in conservative transposition the transposable element is mobilized to other genome location without replicating itself (Galas and Chandler, 2009).

Several studies have focused on the evolutionary factors that govern the proliferation and evolution of transposons and IS elements across bacterial genomes, trying to answer the question of why this type of mobile DNA is maintained in prokaryotic genomes. Some authors postulate that transposable elements are mantained in prokaryotic genomes as selfish DNA replicating itself at the expense of their prokaryotic host, and that are maintained over evolutionary time because they reach an equilibrium, maintaining transposition rates compatible with the viability of their prokaryotic host (Doolittle and Sapienza, 1980; Orgel et al., 1980; Campbell, 1981). Other authors postulates that this equilibrium in not enough to ensure their maintenance over long evolutionary timespans, and that transposable elements are retained because they confer some type of beneficial effect to the host, for example by mobilizing additional genes during their transposition, or conferring temporary benefits that promote their maintenance in bacterial populations over the short term in laboratory conditions, mainly promoting genome rearrangements that affects gene expression patterns by mutations in regulatory regions, or inducing the expression of genes located near the insertion site by promoters included in the IS element (Blot, 1994; Schneider et al., 2000; Edwards and Brookfield, 2003; Schneider and Lenski, 2004). However, to clearly determine if this type of mobile DNA persist due to their beneficial effect to their host is necessary to study their dynamics on evolutionary timescales. The study of abundance and sequence divergence of different copies of a given IS element both within and between genomes reveals that sequence divergence between copies of a given IS within a prokaryotic genome are in most cases extremely low, with identical copies in many cases, in sharp contrast with a significantly higher level of sequence divergence observed between genomes, that indicates very recent expansions of these IS elements in prokaryotic genomes that does not adjust with a selective maintenance over long evolutionary time spans (Sawyer et al., 1987; Lawrence et al., 1992; Chalmers and Blot, 1999). In addition, the distribution and abundance of different types of IS elements between prokaryotic chromosomes are highly skewed, with most genomes having very few or no copies of a given IS element whereas a small number of genomes have large copy numbers, and this patchy distribution is observed even between closely related genomes, with strong variation in copy numbers of a given IS element between different strains of the same species, suggesting an evolutionary scenario in which IS elements would have evolved mainly as selfish DNA by successive cycles of expansions and extinctions mediated by horizontal gene transfer between closely related genomes, with no time enough in each cycle for IS sequence divergence that explains their high levels of sequence identity between IS copies within a genome (Wagner, 2006; Wagner et al., 2007; Wagner and de la Chaux N., 2008). Under this scenario,

massive IS proliferation occurs after the transfer of a given IS element, that takes place mainly between closely related species, until a point in which this is detrimental for host fitness. In this point, natural selection together with down regulation of the transposition activity and excision leads to the elimination of this IS element, that could be re-introduced by horizontal transfer event sometimes thereafter (Wagner, 2006). This hypothesis does not exclude possible beneficial effects of IS proliferation under evolutionary conditions in which higher genome plasticity would be required, but over long evolutionary time is higher the deleterious effect of their proliferation over functional capabilities of the bacteria.

With the advent of whole genome sequencing, it has also been possible to characterize the full set of transposons and IS elements present in different prokaryotic lineages to analyze the factors that govern their proliferation and abundance within prokaryotic genomes and the consequences of their massive proliferation over genome structure and dynamics. It has been observed great heterogeneity in the numbers of IS elements present in different prokaryotic genomes even at close taxonomic ranges, with a massive proliferation of IS elements associated with prokaryotic organisms that have experienced recent transitions to specialized or restricted ecological niches compared with their more ecologically generalized parental strains (Moran and Plague, 2004; Frost et al., 2005; Bordenstein and Reznikoff, 2005; Mira et al., 2006). One of the best examples is found in the genomes of three species of the genus *Bordetella* in the β subdivision of proteobacteria, *B. bronchiseptica*, *B. parapertusis*, and *B. pertusis* (Parkhill et al., 2003). *B. bronchiseptica*, a pathogen that colonizes the respiratory tract of a wide range of mammalian hosts possesses the largest genome (5.34 megabases), with no IS elements presents in the genome although have several prophages. Phylogenetic analysis suggests that is the ancestral *Bordetella* genome structure from which the other two evolve. In comparison, *B. parapertusis*, which infects respiratory tract of humans and sheep's, with a smaller genome size of 4.77 megabases, has lost all prophages but have acquired 22 copies of IS1001 and 90 copies of IS1002. This expansion of IS elements is even higher in *B. pertusis*, a specialized human pathogen that possesses the smallest genome of the three *Bordetella* genomes with 4.1 megabases, and in which has been characterized more than 260 IS elements, with a massive proliferation of IS481 (238 copies). Whole genome comparisons suggest a decisive role of IS elements in the generation of whole genome rearrangements and deletions responsible for the genome size reduction from the ancestor, being also involved in massive gene inactivation with 358 genes being inactivated in *B. pertusis* by insertion of IS elements (Parkhill et al., 2003). A similar massive expansion of IS elements has been also found in the genomes of different species and strains of human pathogenic bacteria like *Shigella* (Jin et al., 2002; Wei et al., 2003; Yang et al., 2005), *Yersinia* (Parkhill et al., 2001; Deng et al., 2002; Chain et al., 2004), or *Burkholderia* (Nierman et al., 2004), but also in plant pathogens like different species of the *Xanthomonas* genus (da Silva et al., 2002; Lee et al., 2005)

and in certain symbiotic bacteria like the nematode symbiont *Photorhabdus luminescens* (Duchaud et al., 2003). By contrast, the reduced genomes of most obligatory intracellular bacteria, specially bacterial endosymbionts of insects with long evolutionary associations with their hosts, have very few, if any, traces of mobile genetic elements in their genomes (Shigenobu et al., 2000; Tamas et al., 2002; Akman et al., 2002; van Ham et al., 2003; Gil et al., 2003; Degnan et al., 2005; Perez-Brocal et al., 2006). However, exceptions to this general tendency also exist, like the genome of the obligatory intracellular pathogen *Mycobacterium leprae* that contains no IS elements despite have experienced a massive gene inactivation process with more than 1000 pseudogenes (Cole et al., 2001), or the genome of the arthropod parasite *Wolbachia pipientis* that, despite having genomic features common to anciently host-restricted bacteria like small genome size and low GC content, retains very large numbers of mobile genetic elements (Wu et al., 2004).

In order to gain insights into the factors that govern IS abundance in prokaryotic genomes, Touchon and Rocha carried out a exhaustive characterization of the IS elements presents in 262 prokaryotic organisms, comparing the number of IS elements in each genome with different variables like genome size, prokaryotic lifestyle, or levels of horizontally transferred genes (Touchon and Rocha, 2007). The results of this work reveals a strong correlation between the number of IS elements presents in a genome and their size, that is explained by the changes in the density of highly deleterious insertion sites with genome size, in the sense that as the genome becomes larger, the fraction of genes whose inactivation leads to very high fitness cost to the bacteria becomes smaller, allowing a better tolerance of IS transposition than small genomes in which most of the genes become essential. In addition, there was also a strong correlation between the number of IS elements and the number of horizontally transferred genes in a genome, indicating the importance of horizontal gene transfer in the evolution of IS elements also postulated by Wagner (Wagner, 2006). However, the presence of most of IS elements out of horizontally transferred regions also suggests that transposition activity is a major factor determining IS abundance. Finally this study revealed that the pathogenic character of the organism does not determines the number of copies of IS elements present in the genome, but its lifestyle. In this context, it was revealed a strong correlation between the frequency of IS elements and the facultative character of ecological association (Touchon and Rocha, 2007), being coincident with other studies in which bacteria that establish facultative association with eukaryotic hosts was estimated to contain four-fold more mobile DNA than obligatory intracellular bacteria, with composition more similar to free-living bacteria, whereas in the small genomes of obligatory mutualistic endosymbionts of insects, mobile genetic elements appear completely absent (Bordenstein and Reznikoff, 2005).

In the context of this thesis, the secondary endosymbiont of the tsetse fly *Sodalis glossinidius* are one of the few genomes corresponding to a facultative

endosymbiotic bacteria in the very beginning of the transition to a host-dependent lifestyle, with a genome of 4.3 megabases more similar in size to closer free living enterobacteria like *E.coli*. However, a massive process of gene inactivation is taking place in the genome of *S. glossinidius* consequence of the relaxed selective pressures over large genome regions that become superfluous in the context of the host-restricted environment, yielding a genome with more than 900 pseudogenes with a large proportion of prophage genes (Toh et al., 2006), providing an excellent case to evaluate the relative role of mobile genetic element proliferation in the initial stages of these genome reduction process. One of the objectives of this thesis is to characterize the different types of mobile genetic elements presents in the genome of *S. glossinidius* and to evaluate the impact of their proliferation in their genome structure, as well as their relative role in the gene inactivation process

1.3 Mechanisms of bacterial evolution

Gene content plasticity in prokaryotic genomes, observed at all evolutionary levels through comparative genomic studies, is consequence of the major role of gene gain and gene loss in prokaryotic evolution. Genomic changes by loss or acquisition of genetic material allow rapid and drastic changes in the functional capabilities of a prokaryotic cell, allowing rapid adaptation to a given environment. This explains the evolutionary success of prokaryotic organisms in the colonization of most natural environments. In this context, comparative genomics has brought major insights into the mechanisms responsible of the fluid evolution of prokaryotic genomes; evolutionary genomics have revealed that prokaryotic adaptation and speciation are determined by acquisition of selectively valuable genes by means of gene duplication or horizontal gene transfer together with gene loss of non-essential genes during periods of relaxed selection as consequences of mutation, deletion, and genetic drift in small sized bacterial populations, with the gene content of prokaryotic genomes being a reflect of their life histories and evolutionary pressures (Snel et al., 2002; Kunin and Ouzounis, 2003b; Gogarten and Townsend, 2005; Lawrence and Hendrickson, 2005; Rocha, 2008b). However, tracing the evolution of gene content across evolutionary lineages and the specific contribution of these different processes in shaping prokaryotic genomes is still a largely open question, suggesting that gene content of prokaryotic genomes is largely dependent of their evolutionary history, genome size, and functional selection, with as many scenarios for genome evolution as the number of possible ecological niches for bacterial adaptation. The specific role of each of the process involved in gene content evolution of specific genomes has been documented, like important role of gene duplication in the evolution of *Vibrio cholerae* (Heidelberg et al., 2000), the massive gene loss affecting genomes of symbiotic and pathogenic bacteria (Andersson and Kurland, 1998; Cole et al., 2001; Nakabachi et al., 2006; Perez-Brocal et al., 2006),

or the massive horizontal gene transfer in the evolution of *Xanthomonadales* (Comas et al., 2006) and in the evolution of different *E.coli* strains (Perna et al., 2001). In all this processes of genome evolution repeated sequences plays a pivotal role due to their essential role as hotspots for recombination events necessary for the insertion and deletion of genetic material in the genomes, generating not only changes in gene content but also alterations in the chromosome structure by rearrangement events like inversions and translocations (Rocha et al., 1999; Huynen and Snel, 2000; Rocha, 2003b; Sankoff and Nadeau, 2003; Rocha, 2004). Comparative genomics have revealed a strong correlation between the presence of mobile genetic elements and the degree of genome stability of prokaryotic genomes that can be explained in terms of evolutionary pressures associated with prokaryotic lifestyles. At one extreme, obligatory bacterial endosymbionts have the smallest genomes and the highest levels of genome stability with absence of genome rearrangements and innovation, whereas at the other extreme, free-living bacteria with larger genomes undergo high levels of gene order and gene content variability between and within species due to gene transfer events and recombination at repeated sequences favored by the presence of high number of mobile genetic elements dispersed across the genome (Casjens, 1998; Bentley and Parkhill, 2004; Fraser-Liggett, 2005). In the next sections of this introduction, the role of horizontal gene transfer, gene duplication, and gene loss will be analyzed in the context of gene content and genome structure evolution.

1.3.1 Genome evolution by gene gain: Gene evolution and horizontal gene transfer as generators of genome variability

The main evolutionary events leading to expansions in gene content in bacterial genomes are intra-genomic gene duplication and lateral acquisition of exogenous genes from other genomes through lateral gene transfer, including different types of mobile genetic elements like prophages and insertion sequences that can be also generated by intragenomic expansions, increasing the gene number. Together with gene loss, these evolutionary events are responsible for the dynamic nature of prokaryotic genomes. The clear identification of homologous genes evolved by speciation (orthologous genes) from those generated by gene duplication (paralogous genes) and by horizontal transfer events (xenologous genes) are essential to understand prokaryotic genome evolution (Zhaxybayeva et al., 2005; Koonin, 2005; Lapierre, 2008; Fournier et al., 2009). Gene duplication is an important evolutionary process for gene innovation, which facilitates the adaptation of organisms to changing environments faster than by gradual mutation. Their impact in evolution as main source of genetic innovation responsible for novel gene functions dates with the seminal work of Ohno in 1970 *Evolution by gene duplication* where postulates that gene duplication is the main process responsible for the emergence of functional novelty during evolution (Ohno, 1970). In the model postulated by Ohno, the duplicated gene is free to accumulate mutations due to

relaxed selection on gene function given that an original gene copy retains the essential ancestral task. In this context, three possible evolutionary outcomes of the duplicated copy arisen: the maintenance of the original sequence and function by concerted evolution, the inactivation by accumulation of degenerative mutations (and frequently its removal from the genome by deletion events), or the emergence on novel gene functions in a process called neofunctionalization.

In the neofunctionalization model, gene copies need sufficient mutations to derive a novel gene function from their parents, assuming that the duplicates are initially neutral and as consequence free to accumulate mutations. Subsequently, this model of gene duplication has been challenged by evolutionary analysis of sequence evolution rates on duplicated genes that demonstrates that duplicated genes do not seems to have experienced any extensive period of neutral evolution(Lynch and Conery, 2000; Lynch and Conery, 2003a; Lynch and Katju, 2004), suggesting a two stages evolutionary model for gene duplication in which duplicated genes are retained and subjected to purifying selection during earlier phases of their evolution due to the benefits of gene dosage effect under a given environmental condition, whereas in later stages of their evolution are likely to create new gene functions by relaxed selection (Kondrashov et al., 2002). Another model for gene innovation by duplication is the duplication-degeneration-complementation model proposed by Force and collaborators in 1999 by which both paralogous genes becomes selected and retained by losing separate functions from an ancestral multifunctional gene, in a process named subfunctionalization (Force et al., 1999; Lynch and Force, 2000). In addition, not only gene duplication but also whole genome duplication contributes to the increase of genome complexity along evolutionary scale, being responsible for the origin of major eukaryotic groups (Wolfe and Shields, 1997; Meyer and Schartl, 1999; Dehal and Boore, 2005).

While in eukaryotes the role of gene duplication in generation of novel gene functions is undoubted due to the limited role of horizontal gene transfer in their evolution, in prokaryotes horizontal transfer events of genetic material between organisms is also an important source of genetic novelty. Several studies have analyzed the role of gene duplication in the evolution of individual bacterial genomes, revealing that families of paralogous genes represent a significant fraction of different prokaryotic genomes (Brenner et al., 1995; Koonin et al., 1995; Labedan and Riley, 1995); duplicated gene families have been detected in almost all prokaryotic genomes sequenced and have been essential to understand patterns of protein sequence evolution and protein family evolution over broad taxonomic ranges in protein families presents in different prokaryotic organisms. The comparative analysis of different prokaryotic genomes soon revealed a clear positive correlation between genome size and the number of duplicated genes. This was first detected by Jordan and collaborators in 2001 in a comparative genomic study with 21 complete bacterial genomes with the objective of characterize lineage-specific

gene expansions by gene duplication, concluding also with the recent emergence of most lineage-specific gene families due to the small size of most prokaryotic gene families and the highest sequence similarity of small-sized families compared with larger ones (Jordan et al., 2001). More extensive comparative genomic studies including more bacterial genomes confirms the strong correlation between genome size and the number of duplicated genes, with small prokaryotic genomes of obligatory intracellular bacteria having the smallest gene family sizes whereas in large-sized genomes like that of *Streptomyces coelicor*, paralogous genes can represent until 50% of the whole genome sequence (Gevers et al., 2004; Pushker et al., 2004). In addition, these comparative genomic studies reveal that not all protein coding genes are equally likely to be as duplicated copies in the genomes, and that most lineage-specific gene family expansions are correlated with the evolutionary pressures acting on each bacterial lineage, being responsible for a significant proportion of the phenotypic and genotypic differences among closely related strains (Saier, Jr. and Paulsen, 1999; Hooper and Berg, 2003a; Hooper and Berg, 2003b; Gevers et al., 2004; Pushker et al., 2004). In this context, a study of protein superfamilies over 56 bacterial genomes carried out by Ranea and collaborators in 2004 identifies two main different types of protein families depending on their degree of correlation with genome sizes. At one side there was protein families involved in essential cellular functions like protein biosynthesis and replication that are presents in almost all bacterial genomes representing ancestral cellular functions conserved across evolution but that do not contributes to the diversity of genome sizes observed between bacteria. At the other side, there were size-dependent protein families that can be subdivided into families with linear correlation relative to genome size and those with non-linear correlation, both with different functional associations. Linearly distributed superfamilies were enriched in functions of the cellular metabolism, and are presents in all bacteria including those of obligatory intracellular bacteria, whereas non-linearly distributed superfamilies are mainly associated with regulation of gene expression and tends to be absent from bacteria with small genomes, increasing in number and complexity in bacteria with larger genomes (Ranea et al., 2004). The fact that larger sized genomes correspond to free-living bacteria living in complex environments explains gene family expansions of regulatory proteins as a mechanisms to regulate gene expression under changing conditions.

The biological role of gene amplification in short-term adaptation of bacterial strains to environmental changes has been long postulated as a mechanisms of over expression of gene products needed for bacterial survival under a given condition, allowing rapid reversion of the duplicated state when the selective pressures changes through gene duplicate inactivation (Rigby et al., 1974; Anderson and Roth, 1977; Romero and Palacios, 1997). Several examples of natural gene amplification support this hypothesis, like amplification of chromosomal regions involved in antibiotic resistance (Koch, 1981; Matthews and Stewart, 1988; Nichols and Guay, 1989),

adaptation for growth under nutrient scarcity conditions (Sonti and Roth, 1989) or high temperatures (Riehle et al., 2001), and increased virulence in pathogenic bacteria(Mekalanos, 1983). Even in small sized genomes of endosymbiotic bacteria gene amplification is present associated to different aspects of host-symbiont interaction, like amplification of genes encoding anthranilate synthase associated to a plasmid in *Buchnera aphidicola* to overproduce tryptophan for aphid host species with high demands for this amino acid (Rouhbakhsh et al., 1996), with alternative arrangements of these tandemly duplicated genes in other aphid species with low demand for tryptophan (Lai et al., 1996). However, comparisons of gene family sizes between strains evolving under reductive genome evolution with larger relatives also reveals that gene family inactivation is a common trend in all processes of genome reduction that can proceed by gene inactivation as consequence of mobile element proliferation like in the evolution of *Shigella flexneri* from *E.coli* strains (Jin et al., 2002; Wei et al., 2003) or by massive gene inactivation consequence of mutational events like in the evolution of *M.leprae* since their divergence from *M. tuberculosis* (Cole, 1998) or in *Ricketsia* lineages (Andersson and Andersson, 1999a; Andersson and Andersson, 2001), revealing that not only genome expansions but also genome reduction can be partially explained by the parallel growth or simplification of gene families.

However, despite the importance of gene duplication in prokaryotic evolution, the capability of prokaryotic organisms to transfer genetic material both within and between species is probably the main factor responsible of the dynamic nature of prokaryotic genomes. This process, known as horizontal gene transfer (HGT) or lateral gene transfer (LGT) has been subject of an intense debate area in evolutionary biology concerning to their relative role shaping prokaryotic evolution at all phylogenetic levels and their implications in the traditional view of prokaryotic species definition (Lawrence, 1999; Gogarten et al., 2002; Boucher et al., 2003; Brown, 2003; Daubin et al., 2003; Kurland et al., 2003; Bapteste et al., 2004). The exchange of genetic material between prokaryotic organisms can be produced by three different mechanisms named transduction, conjugation, and transformation. By transduction, the exchange of genetic material is mediated by bacteriophages that incorporates into their capside structure DNA from the donor genome, that will be integrated in the recipient genome during the infection, whereas by conjugation, the DNA from the donor genome is integrated into plasmids that will be transferred to the recipient genome, allowing the integration of longer DNA segments than by transduction (Ochman et al., 2000; Lawrence and Hendrickson, 2003; Gogarten and Townsend, 2005). By contrast, in transformation, free DNA from the environment is taken up by pathways encoded by the recipient genome itself, and requires the recipient cells to be in a competent state to allow the integration, that is normally induced by environmental conditions like starvation (Dubnau, 1999). Once transferred, foreign DNA must be integrated in the recipient chromosome by recombination that can proceed through two different ways. By homologous

recombination, the foreign DNA replaces a homologous sequence in the recipient genome in a process that is strongly dependent on the degree of sequence divergence between recombining sequences, whereas by illegitimate recombination exogenous DNA without homology with the recipient genome may integrate anywhere in the recipient genome (Gogarten et al., 2002; Ochman et al., 2005). The significance of HGT in prokaryotic evolution was recognized long time before the emergence of modern genomics, although initially there was considered a minor phenomenon that acts only under specific evolutionary pressures and that was marginal in early models of prokaryotic evolution, adaptation and speciation. This early view of prokaryotic evolution and speciation postulated that prokaryotic lineages evolve asexually by binary fission in a clonal model in which the exchange of DNA between bacterial cells occurs at much lower rate than in eukaryotes with sexual reproduction, with periodical selection periods that allow the fixation of strains carrying beneficial characters generated by mutation, generating a bottleneck effect that reduces the genetic variability in bacterial populations to a group of nearly identical individuals (Levin, 1981; Ward, 1998; Cohan, 2001). This model was reinforced by analyses of genetic variation on *E. coli* lineages based on multilocus enzyme electrophoresis, which showed extensive linkage disequilibrium between *E.coli* loci coincident with the expectations of a clonal mode of evolution with minor role of genetic exchange. (Ochman et al., 1983; Whittam et al., 1983a; Whittam et al., 1983b; Ochman and Selander, 1984). This traditional view of prokaryotic evolution changed with the increasing availability of nucleotide sequences that revealed that different genes showed different evolutionary histories in phylogenetic reconstructions (Milkman and Crawford, 1983; DuBose et al., 1988; Dykhuizen and Green, 1991), a feature that is difficult to reconcile with a clonal view of bacterial evolution. It suggested a more important role of gene exchange by homologous recombination in bacterial evolution than previously recognized. Initial estimates of homologous recombination rates in *E.coli* indicated that it was strongly similar to mutation rate (Guttman and Dykhuizen, 1994), which means that is equally probable that an allele has been arisen by mutation or by homologous recombination (ratio 1:1 between both processes), and this was further reinforced with multilocus sequence typing studies that increased these estimates to a ratio of 20:1 and 50:1 for *E. coli* and even higher for other bacterial lineages (Feil and Spratt, 2001). However, this model of gene exchange by homologous recombination is strongly dependent of the degree of sequence divergence between the donor and the recipient genomes, with different recombination barriers to foreign DNA integration by homologous recombination like the mismatch repair systems of bacterial genomes capable of distinguish foreign DNA based on sequence divergence eliminating DNA too divergent from that of the recipient genome (Vulic et al., 1997; Majewski and Cohan, 1998; Cohan, 2001), or restriction-modification systems that eliminated sequences without the methylation patterns of the recipient genome (Rocha et al., 2001). In addition, homologous recombination alone does not explain the enormous variability of gene content in complete genomes of closely related species and even

strains of the same species that has been revealed with the emergence of modern genomics through the publication of the complete genomes of hundreds of prokaryotic organisms that allows the evaluation of the impact of HGT at different evolutionary levels. Examples of the massive presence of horizontally transferred genes can be found in nearly all genomes completely sequenced, although their relative fraction respect the whole genome sequences depends strongly on the ecological context and evolutionary pressures affecting the prokaryotic lineages. HGT is responsible for the 30% of the gene content of enterohaemorraghic strain 0157:H7 of *E. coli*, comprising more than 1387 genes distributed in strain-specific clusters also known as pathogenicity islands responsible for their pathogenic character, being also highly variables even among different isolates of the same strain with different virulent phenotypes (Perna et al., 2001; Welch et al., 2002), whereas in *Salmonella enterica*, the vast majority of virulence genes responsible for pathogenic phenotype appears associated to two different pathogenicity islands horizontally acquired that allows *Salmonella* cells to invade epithelial cells and cause a systemic disease (Groisman and Ochman, 1997). In hyperthermophilic bacteria *Aquifex aeolicus* and *Thermotoga maritima*, massive horizontal transfers with archaea are responsible up to 25% of their gene content (Aravind et al., 1998; Nelson et al., 1999). It has been also demonstrated the essential role of ancient horizontal gene transfer in the evolution of *Xanthomonadales* group by phylogenetic analysis, revealing high levels of mosaicism with different genes showing different affinity with gamma, beta, and alpha proteobacteria after phylogenetic reconstructions (Comas et al., 2006). In contrast, the genomes of intracellular bacterial endosymbionts and parasites show very few, if any, genes horizontally transferred due to their ecological isolation in the stable environment of their host cytoplasm (Ochman et al., 2000; Ochman and Moran, 2001; Silva et al., 2003), although horizontal gene transfer is also involved in evolution of symbiotic associations like in nitrogen-fixing symbiosis of *Mesorhizobium loti* with leguminous plants through the acquisition of symbiotic island containing the genes for nitrogen fixation and nodule formation together with cofactor biosynthesis genes (Kaneko et al., 2000). Many of these transfers are originated by illegitimate recombination mechanisms from distant related sources rather than by homologous recombination, and these transfers from distant sources increase the heterogeneity of prokaryotic genomes in terms of sequence composition and phylogenetic signals (Jain et al., 2002; Lawrence and Hendrickson, 2003). This atypical feature of horizontally transferred genes has been used for their identification by both phylogenetic methods and compositional methods. Phylogenetic methods are based on the identification of genes that are grouped with genes from otherwise unrelated taxa in phylogenetic reconstructions based on sequence data under appropiate model of evolution or genes with highly restricted phylogenetic distributions restricted to a given linage but excluded from their closer relatives (Doolittle, 1999a; Doolittle, 1999b; Eisen, 2000). Although phylogenetic reconstructions are powerful methods for the detection of lateral gene transfer events, they are not free from the inherent

problems of this methodology, like the potentially fast rates of sequence evolution of certain species that may produce wrong phylogenies, the problems associated with the resolution of deep diverging nodes, or the problems associated with erroneous orthologous identifications and incomplete taxon sampling (Lecointre et al., 1993; Moreira and Philippe, 2000; Philippe et al., 2005). By contrast, compositional methods do not depend on sequence comparisons with other genomes, but rather by the identification of compositional deviations of the horizontally transferred regions compared with the average composition of the recipient genomes in parameters like nucleotide composition, dinucleotide patterns, or codon usage biases (Karlin and Burge, 1995; Lawrence and Ochman, 1998; Mrazek and Karlin, 1999). However, these compositional methods have also caveats in the recognition of horizontally transferred genes in the sense that only transferences from distant sources with large differences in sequence composition will be effectively detected by those methods, whereas transfers from closer relatives with similar nucleotide composition will not be detected. In addition, selective pressures over certain genes can originate compositional deviations from the average genome does not originated by horizontal gene transfer, leading to genes with atypical nucleotide composition that can lead to erroneous inferences of laterally transferred genes (Lafay et al., 1999); finally , once integrated in the genome, laterally transferred genes with base composition and codon usage patterns of the donor genome start to adapt to the mutational patterns of the recipient genomes in a process known as amelioration by which the composition of transferred genes becomes more similar to the average composition of the whole genome, making difficult the detection of ancient transfers (Lawrence and Ochman, 1997). Despite this caveats, this compositional methods provides estimates of the acquisition time through the extent of deviation in compositional parameters of laterally transferred genes from the characteristic values of the recipient genome, like deviations in the average nucleotide compositions at first and second codon positions or deviations in the codon usage patterns. This kind of analysis was carried out by Lawrence and Ochman over the genome of *E.coli* strain MG1655, identifying 755 horizontally transferred genes representing 17.6% of the genome, and estimating a rate of transfer of 16 kb per million years since their divergence from *S. enterica* (Lawrence and Ochman, 1998). Comparative genomics have also revealed a strong heterogeneity in the presence of horizontally transferred genes in bacterial genomes, ranging from virtually none in the reduced genomes of obligatory intracellular bacteria like *Ricketsia prowaletzkii*, *Mycoplasma genitalium*, or *Borrelia burgdorferi* to nearly 17% of the genome in *Synecocystis PCC6803* (Ochman et al., 2000). Functional analysis of horizontally transferred genes have also revealed that not all genes are equally prone to be transferred and that informational genes encoding proteins involved in transcription, translation, and replication are much less prone to be transferred than operational genes encoding metabolic enzymes, transporters, and other gene functions. This was explained by the higher number of cellular complexes in which are involved the products of informational genes compared with operational genes, like the transcriptional and

translational complexes that involves until 100 different gene products. This leads to Jain and collaborators to postulate the complexity hypothesis by which informational genes are restricted to be transferred due to their strong level of coadaptation with their interacting genes, that makes very unlikely their replacement by foreign genes, whereas this restriction is smaller in operational genes allowing their continuous transfer among prokaryotic lineages (Jain et al., 1999). However, several examples of horizontal gene transfer events affecting informational genes have been detected, like aminoacyl tRNA synthetases (Wolf et al., 1999; Woese et al., 2000), ribosomal proteins (Brochier et al., 2000), or even genes encoding subunits of the RNA polymerase complex (Iyer et al., 2004), indicating that no gene appears completely immune to lateral gene transfer.

With all these observations, is it clear that horizontal gene transfer plays a pivotal role in prokaryotic evolution, leading to mosaic genomes that contain genes with different evolutionary histories. However, while the presence of horizontal gene transfer in almost all prokaryotic genomes is not questioned, there is an intense discussion with respect to their global impact over prokaryotic evolution in the sense of the validity of the traditional prokaryotic classification following the Darwinian paradigm of bifurcated evolutionary trees. At one side, there are several authors that consider a single bifurcated tree as an incomplete representation of the true evolution of prokaryotes due to the prevalent presence of laterally transferred genes at all taxonomic ranges and the few gene families putatively free of lateral transfer events, making any phylogenetic reconstruction based on a single gene an incomplete description of prokaryotic evolution (Doolittle, 1999b; Gogarten et al., 2002; Boucher et al., 2003; Bapteste et al., 2004). Several studies have revealed substantial incongruence in phylogenetic trees reconstructed from single orthologous genes (Woese et al., 2000; Nesbo et al., 2001), significant variations in GC content and codon usage across bacterial genomes (Medigue et al., 1991; Guerdoux-Jamet et al., 1997; Ragan, 2001), and substantial variations in gene content composition between closely related genomes, that has lead to some authors to purpose a more natural classification scheme, that some refers as "synthesis of life" (Gogarten et al., 2002), that copes with the essential role of horizontal gene transfer in prokaryotic evolution in a quantitative way. Under this framework, prokaryotic evolution is better represented as a phylogenetic network in which vertical branches of prokaryotic tree are horizontally interconnected with different strength depending on the degree of connectivity between branches based on their corresponding transferred genes (Bapteste et al., 2004; Kunin et al., 2005; Dagan and Martin, 2006; Doolittle and Bapteste, 2007).

However, other studies suggest that, despite the prevalence of horizontal gene transfer in the evolution of prokaryotic genomes, a minimal fraction of transferred genes are fixed in the evolutionary lineages, being possible to identify a core of genes sharing a common history of vertical inheritance, reflecting the species

phylogeny. In addition, a significant fraction of lateral gene transfer inferences based on phylogenetic incongruence on gene tree comparisons can be explained by failures or inconsistencies of the phylogenetic reconstruction methods and lack of clear phylogenetic signal of certain gene sequences (Kurland, 2000; Kurland et al., 2003). Phylogenetic analyses with quartets of bacterial species with completely sequenced genomes at different taxonomic levels show that only a minor fraction of orthologous genes shows discordant topologies from species phylogeny based on small subunit ribosomal RNA. This vertical signal of orthologous genes contrasts with the high levels of gene acquisition determined from gene content comparisons between complete genomes resulting in two major groups of genes in prokaryotic genomes, orthologous genes conserved among genomes with low levels of gene transfer among species and laterally acquired genes that corresponds principally with strain specific genes of unknown function and genes corresponding to mobile genetic elements like bacteriophages or insertion sequences, being possible high levels of horizontal gene transfer affecting non-essential genes with the existence of a core of orthologous genes with essential function that can be useful to define bacterial species in phylogenetic analysis (Daubin et al., 2003; Ochman et al., 2005; Kuo and Ochman, 2009). Similar conclusions were observed in a phylogenetic study with 13 gamma proteobacterial genomes in which 203 of the 205 orthologous genes present in the 13 genomes showed concordant phylogenies with the species phylogeny inferred from ribosomal sequences and from concatenated alignment of the 205 orthologs (Lerat et al., 2003). This behavior was also observed in genes absent from one or more of these genomes, with very few of these genes displaying statistically supported incongruencies with the organismal phylogeny inferred from ribosomal or concatenated genes, despite the major role of horizontal gene transfer reflected in the large number of gene families restricted to one or two genomes (Lerat et al., 2005). Recently, a mixed model has been proposed in which predominant vertical inheritance is combined with pathways of gene sharing between closely related organisms, like between cyanobacteria strains of the genera *Prochlorococcus* and *Synechococcus* or between distantly related organisms that live in similar ecological niches (Beiko et al., 2005).

In conclusion, the observed gene content variability between prokaryotic genomes is a clear indicative of the important role that horizontal gene transfer is playing over prokaryotic evolution. Horizontal gene transfer of single genes or even gene clusters in the form of genomic islands or pathogenicity islands allows a rapid adaptation of prokaryotic lineages to a changing environment, conferring the functional capabilities needed to exploit a novel ecological niche or allowing changes to pathogenic lifestyle, being also responsible for strain-specific traits traditionally employed for species definition. However, it is also clear that a significant fraction of these transfers corresponds to genes of unknown function and genes corresponding to mobile genetic elements like bacteriophages or insertion sequences whose effects over the recipient genome are far from being beneficial, so

in the long term a minor fraction of the transferred genes will be stably maintained on the prokaryotic chromosomes. In addition, the relative role of horizontal gene transfer is highly variable across different taxonomic groups, depending on the ecological constraints and evolutionary pressures affecting bacterial lineages, so any kind of generalization with regard to a global quantitative effect of horizontal gene transfer over all prokaryotic kingdoms will be meaningless. Finally, a realistic model for prokaryotic evolution and speciation has to consider not only the effects of horizontal gene transfer and gene duplication increasing the gene content of prokaryotic genomes, but also the effects of gene loss in the evolution of many prokaryotic lineages, with the balance between gene gain and gene loss being probably the key evolutionary parameter to understand prokaryotic genome evolution (Snel et al., 2002; Kunin and Ouzounis, 2003a; Kunin and Ouzounis, 2003b; Mirkin et al., 2003).

1.3.2 Genome evolution by gene loss: Reductive evolution in obligatory intracellular bacteria

I have discussed in the previous section the importance of gene duplication and, specially, horizontal gene transfer in the generation of novel biological capabilities that allows prokaryotic organisms to rapidly adapt to changing environments and novel ecological niches. However, the opposite phenomenon, gene loss, is equally important in prokaryotic evolution, and is especially relevant in the evolution symbiotic associations between prokaryotes and eukaryotes. One of the clearest patterns revealed by comparisons of genome sizes across different prokaryotic lineages is that species that establish obligatory associations with eukaryotic hosts have smaller genomes than those of their free-living relatives (Wernegreen, 2002; Moran, 2007; Moya et al., 2008; Silva and Latorre, 2008). Because of the high degree of genomic compactness observed in most prokaryotic genomes, variations in genome size are strongly correlated with variations in gene content, in contrast to what is observed in eukaryotic organisms in which no correlation is observed between gene number and genome size, what has been called the "C-value paradox" (Thomas, Jr., 1971). The knowledge of prokaryotic genome sizes, initially by means of analytic techniques like pulse field electrophoresis and, in recent times, by genome sequencing, together with the reconstruction of their evolutionary relationships based on well supported phylogenetic trees have revealed that genome reduction is a recurrent process that has taken place several times in different prokaryotic lineages across evolution in response to different evolutionary pressures. Extreme genome reduction has been documented in different bacterial groups like gram-positives, *chlamydiae*, *spirochetes*, and different lineages within alpha and gamma subdivisions of proteobacteria (Silva and Latorre, 2008). Initially, some authors proposed that these small sized genomes are the natural state of ancestral genomes, and that larger genomes were derivate from smaller ones through gene acquisition or duplication events (Wallace and Morowitz, 1973), although now it has

long been recognized that reduced genomes are derived features evolved from ancestral larger genomes through a process of genome reduction consequence of a drastic change in the ecological niche and evolutionary pressures associated to a given prokaryotic lineage (Andersson and Kurland, 1998; Andersson and Andersson, 1999b; Klasson and Andersson, 2004). This process of genome reduction have occurred several times in different taxonomic lineages, for example in parasitic bacteria of the genera *Wolbachia* and *Ricketsia* in the alpha subdivision of proteobacteria (Andersson and Andersson, 1999a; Ogata et al., 2001; Sallstrom and Andersson, 2005; Werren et al., 2008), the class mollicutes, that includes several small sized genomes like that of *Mycoplasma genitalium*(Fraser et al., 1995), the class actinobacteria, that includes *Mycobacterium leprae* and *Tropheryma whipplei* (Cole et al., 2001; Raoult et al., 2003), in pathogenic bacteria of the class *Chlamydiae* (Horn and Wagner, 2004; Horn, 2008), in the gram positive intracellular plant pathogen *Phytoplasma asteris* (Oshima et al., 2004), or in parasitic spirochetes such as *Borrelia burgdorferi* (Fraser et al., 1997). Extreme genome reduction has also been found in the hyperthermophilic nanoarchaeon *Nanoarchaeum equitans*, that lives in symbiotic association with the crenarchaeon *Ignococcus* (Waters et al., 2003), and also in free living cyanobacteria *Prochlorococcus marinus*, in which variations in genome sizes of one megabase are found between different strains associated to different ecotypes adapted to high and low light environments in the marine medium (Dufresne et al., 2005; Kettler et al., 2007). But among all prokaryotic lineages, the gamma subdivision of proteobacteria includes the most numerous and extreme cases of genome reduction described to date, mostly associated to endosymbiotic bacteria that lives in obligatory intracellular association with insect hosts (Wernegreen, 2002; Moran, 2003; Baumann, 2005; Silva et al., 2007), and is the group in which the dynamics of genome reduction has been more extensively studied together with obligatory intracellular parasitic alpha proteobacteria of the genera *Ricketsia* and pathogenic intracellular bacteria of the genera *Mycobacterium*, due to the availability of complete genome sequences from different strains of the same species together with complete genome sequences of closer relatives with larger genomes. In addition, although in both parasitic and endosymbiotic bacteria the process of genome reduction is associated with a complete host-dependent lifestyle due to a massive loss of gene functions that impedes the autonomous survival of bacterial cells in a free-living stage, the outcomes of the process are different in terms of the ecological relationship established with their hosts. On the one hand, in bacterial endosymbionts of insects, the host cells become completely dependent on products provided by the bacterial endosymbiont for their survival, establishing a mutualistic relationship in which none of the partners can survive without each other. Most of these associations have a nutritional purpose, in which bacterial endosymbionts provides insect hosts with an additional supply of a given metabolite or nutrient that is absent from their unbalanced diets and that the insect hosts cannot produce by themselves, like amino acid supply of *Buchnera aphidicola* to their aphid hosts,

cofactors and vitamins supply of *Wigglesworthia glossinidia* to their tsetse fly host, or urea recycling and amino acid supply of *Blochmannia floridanus* to their ant hosts, retaining genes encoding functions more directly beneficial for their hosts rather than for the endosymbiotic bacteria itself (Akman et al., 2002; van Ham et al., 2003; Gil et al., 2003; Perez-Brocal et al., 2006). By contrast, in parasitic and pathogenic associations like that of the genera *Ricketsia*, *Mycoplasma*, *Borrelia*, *Treponema*, or *Chlamydia*, the bacterium does not provides any beneficial effect to the host, and retains mainly genes involved in cellular interactions, pathogenesis and defense strategies to ensure their survival inside host cells, with many of these genes presents also in non-pathogenic symbiotic bacteria, where it carries out a different function depending on the ecological context (Groisman and Ochman, 1997; Hentschel et al., 2000; Ochman and Moran, 2001; Lawrence, 2005)

Despite variations in the ecological role of the intracellular bacteria, both symbiotic and parasitic or pathogenic bacteria share common features in terms of drastically reduced genomes encoding a minimal streamlined metabolism, with high rates of DNA sequence evolution and strong nucleotide compositional biases, and with lower levels of genome flux by horizontal gene transfer, indicating a common mechanisms of reductive evolution. In this reductive process, is it possible to distinguish the initial stages in which a free-living bacterium starts its association with an eukaryotic host from the last stages characterized by a highly streamlined and stable genomes consequence of a massive gene loss process associated with a strict host-dependent lifestyle. The initial transition from a free-living to an intracellular environment supposes an important change in the selective pressure over the entire gene repertoire of the prokaryotic organism. In a transition to a host associated lifestyle, host tissues provide much more stable environment with a more or less constant supply of metabolic intermediates than in the free-living environment. In this context, the selective pressure to retain most biosynthetic gene functions is strongly reduced and as consequence, inactivating mutations are not effectively eliminated by purifying selection, leading to the massive accumulation of pseudogenes over the prokaryotic chromosome. In addition, this lack of selective pressure over most parts of the chromosome allows the massive proliferation of different types of mobile genetic elements, like insertion sequences and prophages that can produce further gene inactivation if they insert inside a gene sequence, being also an important source of genome rearrangements by recombination events between these repeated sequences (Mira et al., 2002; Moran and Plague, 2004; Silva and Latorre, 2008). Gene loss is also enhanced by changes in the structure of intracellular bacterial populations, which are characterized by a drastic reduction in the effective population size compared with their free living relatives consequence of the drastic population bottlenecks that these intracellular bacteria experienced due to their strict transmission by vertical inheritance from mothers to their descendents (Andersson and Kurland, 1998; Mira and Moran, 2002). This supposes an increase of random genetic drift and a decrease in the efficiency of selection that leads to the

accumulation of slightly deleterious mutations in genes that are non-essential but beneficial. In addition, these gene loss events could not be restored by horizontal gene transfer due to the restricted intracellular environment associated with this bacterium, so any gene loss is, in principle, irreversible, leading in some cases to a decrease in the fitness of the bacterial population by the fixation of these slightly deleterious mutations that could not be reverted by recombination, a process known as Muller´s Ratchet (Muller, 1964; Felsenstein, 1974; Andersson and Kurland, 1998). As consequence, the initial stages of the genome reduction process are associated with a massive proliferation of pseudogenes and repetitive elements, as it is observed in the genomes of pathogenic *Mycobacterium leprae*, with more than 1000 pseudogenes compared with their closer relative *M. tuberculosis* (Cole et al., 2001), or in genomes of the genera *Bordetella*, in which genome reduction and insertion sequence proliferation are strongly correlated with an increased host-restricted lifestyle (Parkhill et al., 2003). However, this massive gene inactivation has to be followed by gene loss in order to produce the observed genome reduction associated with the last stages of the process, like in the different strains of the primary endosymbiont of aphids *B. aphidicola* (Shigenobu et al., 2000; Tamas et al., 2002; van Ham et al., 2003; Perez-Brocal et al., 2006), or in the extreme reduction of the primary endosymbionts of psyllids *Carsonella rudii* (Nakabachi et al., 2006). In this context, two different aspects of the reductive evolution process have been specially studied by evolutionary biologists in recent years, named the dynamics of the genome reduction in terms of size and content of the deletional events and the evolutionary forces that govern the process of genome reduction.

Concerning with the first point, the genome downsizing can proceed by gene-by-gene inactivation and progressive gene disintegration or by large deletional events that allow the simultaneous removal of many genes in a single event. In order to distinguish between both possibilities it would be necessary to analyze the gene content and order of the ancestral genome before the reductive evolutionary process started, which in most cases dated back to hundreds of millions of years ago. In this context, comparative genomics offers the possibility to reconstruct the gene content of these ancestral genomes based on a parsimony criterion in which, given a well supported phylogeny and a minimal number of three genomes (including one genome as outgroup, the reduced genome under study, and a closer relative with large genome sizes), the gene content of the hypothetical ancestor of the reduced genome can be inferred based on the orthologous genes present and absent in the reduced genomes compared to their closer relatives with large genomes. Based on this approach, the minimal gene content of the ancestral genome the endosymbiotic bacteria *B. aphidicola* from the aphid *Acyrthosiphon pisum (BAps)* was reconstructed by comparison with the genomes of *E. coli* and *V. cholerae* considering as ancestral genes those orthologous genes presents in both *B. aphidicola* and *E. coli* together with orthologous genes presents in *E.coli* and *V.cholerae* and absent in *B. aphidicola*, with this last set being genes that has been

lost during the reductive evolution process, leading to an ancestral gene content of 1818 genes (Silva et al., 2001). This ancestral gene content was increased in another study that consider additional genomes in the comparison, leading to a gene content of *BAps* ancestor of 2425 genes by the inclusion of *Y. pestis* as possible outgroup specie together with *V. cholerae* (Moran and Mira, 2001), reflecting the importance of the compared genomes in this type of analysis. These two studies proposed alternative views of the genome reduction process, and whereas Moran and Mira postulates that the large number of gene losses located between synthenic fragments in *BAps* and *E. coli* are indicatives of the predominant role of large deletional events in the initial stages of genome reduction, Silva and collaborators postulates that the presence of a major proportion of single-gene deletions within synthenic fragments between *BAps* and *E. coli* are consequence of the major role of gene inactivation and progressive degradation in the reductive evolution process. With the availability of whole genomes of different strains of *B. aphidicola* from different aphid hosts, it has also been possible to reconstruct the gene content and order of their last common ancestor (last symbiotic common ancestor or LCSA), as well as to map specific gene loss events in different branches of the phylogenetic tree leading to actual reduced genomes and to infer rates of DNA loss (Tamas et al., 2002; Silva et al., 2003; Gomez-Valero et al., 2004a). This reveals rates of DNA loss enough high to produce the almost complete gene disintegration of inactivated genes in a short evolutionary time (Gomez-Valero et al., 2004a). Similar studies has carried out in other prokaryotic lineages with reduced genomes, like in the alpha subdivision of proteobacteria comparing 13 genomes that includes both small-sized genomes of obligatory intracellular parasites like bacterial of the genera *Ricketsia* and *Wolbachia*, facultative intracellular bacteria with larger genomes like *Bartonella* and *Brucella*, together with the largest genomes of soil-borne plant symbionts and pathogens like *Shinorhizobium*, *Agrobacterium*, and *Bradyrhizobium* (Boussau et al., 2004).In this study, a parsimony criterion were employed to reconstruct the gene content of the ancestor of the 13 genomes together with the most parsimonious scenario of genome evolution in each of the lineages, leading to an ancestral genome containing between 3000 and 5000 genes, with massive gene losses that occurred twice independently in the evolution of alpha proteobacteria associated to the ancestor of the obligatory intracellular lineages *Ricketsia* and *Wolbachia* and in the ancestor of the facultative intracellular lineages *Bartonella* and *Brucella*, whereas massive gene expansions that takes place in the lineages corresponding to plant – associated *Rhizobiales* with larger genomes, specially in *Mesorhizobium loti* and *Bradyrhizobium japonicum* (Boussau et al., 2004). A similar study was carried out specially focused in the genome evolution of seven different species of the genus *Ricketsia*, revealing the prevalent role of gene loss as main cause of genome diversification across different species, with highly variable rates of gene loss, genome rearrangement and sequence evolution in different *Ricketsia* lineages associated with adaptations to their different intracellular niches corresponding to different arthropod hosts (Blanc et al., 2007). In addition, the identification of large

number of pseudogenes in each genome indicates that genome reduction is still an ongoing process in Ricketsia evolution, coinciding with previous studies (Andersson and Andersson, 1999b; Andersson and Andersson, 1999a) and with the observation of intense duplication of transposase genes in different lineages like *R. felis* or *R. bellii* (Blanc et al., 2007). The process of genome reduction has been also analyzed in the intracellular pathogen *Mycobacterium leprae*, the causative agent of leprosy, that have a reduced genome (3.2 megabases) compared with other mycobacterial species, with more than 1000 pseudogenes indicative of a massive process of gene decay since their divergence from the species of the *M. tuberculosis complex* (Cole et al., 2001; Brosch et al., 2001). Gene content comparisons of orthologous genes and pseudogenes have allowed the reconstruction of the gene content and order of the ancestor of *M. leprae*, revealing a total of 1537 gene losses in the lineage leading to *M. leprae* of which 1149 are presents as pseudogenes and 408 genes has been completely lost from *the* genome (Gomez-Valero et al., 2007). The reconstruction of the ancestral gene order reveals that the majority of gene losses are produced by gene-by-gene deletional events, detecting only a single block of 37 contiguous genes lost in block in *M. leprae*, and the dating of the gene inactivation events based on nonsynonymous substitutions shows that pseudogenes are originated from a massive gene inactivation event dated back 20 million years ago (Gomez-Valero et al., 2007).

With regard to the evolutionary forces that govern the process of genome reduction, it has been proposed different explanations that can be generally grouped into mutational explanations and selectionist explanations. The mutational hypothesis postulates that the compact nature of reduced genomes is consequence of a mutational bias towards deletions over insertions in prokaryotic genomes (Mira et al., 2001). This mutational bias is counterbalanced by natural selection on gene function, and the balance between this opposite forces determines the final genome size. In free-living bacteria, natural selection for gene function together with the acquisition of exogenous genes through horizontal gene transfer prevents massive genome reduction, but in intracellular bacteria, the relaxed selective pressure over many gene functions in the nutrient-rich intracellular environment and the increased effect of genetic drift associated to reduced effective population sizes increases the fraction of non-functional DNA in the genome, that will be eliminated as consequence of the deletional bias by either large deletions covering several genes or by gene-to-gene disintegration by small deletional events (Mira et al., 2001; Silva et al., 2003). This mutational bias towards deletions has been detected in both free-living and intracellular bacteria analyzing the size and frequency of insertions and deletions in bacterial pseudogenes compared with functional orthologs (Andersson and Andersson, 1999a; Andersson and Andersson, 2001). It has been also proposed that the deletional bias would prevent the accumulation of detrimental mobile genetic elements like transposons, insertion sequences, and specially bacteriophages due to the predominant role of horizontal gene transfer in bacterial evolution,

(Lawrence et al., 2001). Under this scenario, high deletion rates will be selectively favored in free-living bacteria because can prevent the accumulation of deleterious genetic material in the form of genomic parasites like bacteriophages or insertion sequences, but in obligatory intracellular bacteria, their isolated ecological niche with no influx of genetic material by horizontal gene transfer leads to reduced deletion rates that, together with the increasing effect of random genetic drift, generates the increasing accumulation of pseudogenes that is observed in different intracellular bacteria like *M. leprae* or different *Ricketsia* spp. (Andersson and Kurland, 1998; Cole et al., 2001). However, the mutational bias towards deletions associated to bacterial genomes would progressively remove these pseudogenes, leading to the highly streamlined genomes associated to long-term endosymbionts like *Buchnera aphidicola* and *Carsonella rudii* (Lawrence et al., 2001).

Is has also been argued that the compact nature of prokaryotic genomes is consequences of their selective advantage over organismal fitness in terms of faster replication rates and less energy allocation to replication of non-essential DNA, that leads to the elimination of almost all non-coding DNA that is observed in most prokaryotic genomes in comparison with eukaryotes (Cavalier-Smith, 2005; Giovannoni et al., 2005). Under this hypothesis, it is assumed that the cost of replicating and maintaining a segment of few nucleotides is significant enough to be detected by natural selection, and those genomes with smaller amounts of non-coding DNA will be selectively favored in natural populations with effective population sizes larger enough to avoid the effects of random genetic drift. In many prokaryotes, the large effective population sizes will enhance the effect of natural selection, but this explanation does not apply to the reduced genomes of intracellular bacteria, in which effective population sizes are significantly reduced compared to that of free-living bacteria (Mira and Moran, 2002; Daubin and Moran, 2004), although the polyploidy observed in many of these intracellular bacteria can increase the effective population size if genomes are the unit of selection (Komaki and Ishikawa, 1999; Komaki and Ishikawa, 2000). In addition, many rapidly growing organisms like *E.coli* have large chromosomes, whereas some slow growing organisms like *Borrelia burgdorferi* have small genomes; It is also clear that multiple replication forks can be maintained simultaneously in most actively replicating cells, what indicates that the rates of cell growth are not limited by replication rates, confirmed by the lack of correlation between genome sizes and doubling times in different prokaryotic organisms (Bergthorsson and Ochman, 1998; Mira et al., 2001).

Finally, an explanation based on population genetic principles was proposed by Michael Lynch and collaborators that tries to explain the increases in genome size and complexity observed from prokaryotic to eukaryotic genomes (Lynch and Conery, 2003b; Lynch et al., 2006; Lynch, 2006). This hypothesis assumes that nearly all forms of excess DNA in terms of non-coding DNA constitutes a

mutational burden to genomes and that minimization of genome size would be favorable in selective terms because reduces this mutational burden, being the population genetics environment defined in terms of effective population sizes what determines the lineages that will eliminate this excess of DNA more efficiently (Lynch, 2006). Under this hypothesis, an insertion of a DNA segment imposes a mutational burden to the organisms that will be determined by the number of positions (*n*) that needs to be unchanged in order to be stably maintained in the genome; given a rate of mutation μ and an effective gene size *Ng*, that in prokaryotic organisms is equal to effective population size, Lynch and collaborators estimate that a species in which $2*Ng\mu >> 1/n$ will be essentially immune to the fixation of nonfunctional DNA segment because it will be eliminated by natural selection. This supposes that organisms with large effective gene sizes or high mutation rates will tend to have smaller genomes. Under this assumption, the compact nature of prokaryotic genomes compared with eukaryotic ones is explained by their much higher effective population sizes, and estimates of $2Ng\mu$ in prokaryotes and higher eukaryotes shows that this parameter is ten folds higher in prokaryotes (Lynch, 2006). However, although this hypothesis explains genome size variations between eukaryotes and free-living prokaryotes, fails to explain the situation of intracellular bacteria, in which highly streamlined genomes are associated to a significantly reduction in effective population sizes (Daubin and Moran, 2004). In order to explain the evolution of intracellular bacteria by this mutational burden hypothesis, it has been argued that polyploidy associated to these organisms means that their effective gene size is much higher than their effective population size (Lynch and Conery, 2004).

1.4 Genomes and systems biology: Filling the gap between the genome and the phenotype of an organism

Comparative genomics has been essential in the study of prokaryotic evolution revealing different patterns of gene gain and loss across different evolutionary lineages and the importance of horizontal gene transfer in the evolution of prokaryotic genomes, but to understand the effects of genome structure and organization over the functional capabilities of the organisms is necessary a global integration and analysis of the whole set of molecular components that defines a living cell. In this context is widely know that most biological characteristics arises from complex interactions between cellular constituents, like the interaction between transcription factors with regulatory sequences to control gene expression, the interaction between constituents of different protein complexes, or the integration of biochemical reactions in metabolic pathways that describes cellular metabolism (Barabasi and Oltvai, 2004; Albert, 2005; Almaas, 2007). This kind of interactions can be represented in different types of graphs or networks, and the integration

between these different networks is what determines the final behavior of any living cell. We can reconstruct protein-interaction networks where the nodes are proteins and the links between them are defined by molecular interaction data inferred from genomic, transcriptomic, and proteomic data (Rain et al., 2001; Butland et al., 2005; Arifuzzaman et al., 2006). In transcription regulatory networks the nodes are transcription factors and their regulatory targets and the links are regulatory interactions between them characterized experimentally by chip on chip protocols that allow the purification of the protein-DNA complexes and the characterization of the regulatory interactions (Babu et al., 2004; Luscombe et al., 2004; Ma et al., 2004; Salgado et al., 2006b; Salgado et al., 2006a). In this context, regulatory networks can be viewed as direct graphs in the sense that the direction of the interaction can be established from the transcription regulator to their target genes (Seshasayee et al., 2006). But perhaps the most studied biological networks are the metabolic networks, that represent the set of biochemical reactions that occur in the cells, and that allow a computational approach to the physiology of the whole bacterial cell in terms of its growth capabilities under different media, its nutrient needs for growth, or its efficiency in the production or consumption of different metabolites (Feist et al., 2009; Durot et al., 2009). In metabolic networks, different graph representations are possible, the most common of which is a substrate graph where the nodes represents cellular metabolites that are connected by edges that corresponds to metabolic reactions in which they are involved. This is the scheme most used in all metabolic network reconstructions (Schilling and Palsson, 2000; Reed et al., 2003; Borodina et al., 2005; Puchalka et al., 2008; Oberhardt et al., 2008), although alternative representations are also possible, like reaction graphs where the nodes correspond to metabolic reactions that are connected if they share at least one metabolite (Albert, 2005; Almaas, 2007).

Theoretical studies on the topological organization of biological networks have revealed that despite the diversity of networks existing in nature in terms of components and type of interactions, their structure and organization is governed by common organizational principles, also shared with other complex networks of social interest like social networks or the world wide web network, the most important of which is their power-law distribution in their connectivity degrees that defines these complex networks as scale free (Barabasi and Albert, 1999; Jeong et al., 2000; Wagner and Fell, 2001). Biological networks with a power-law connectivity distribution are highly non-uniform, and their global structure is maintained by a few nodes or "hubs" with very high number of connections whereas the rest of the nodes have a few links. This structure means that there is high diversity of node degrees, defined as the number of connections of a given node, and this makes not possible to define a typical node in the network that could be used to characterize the rest of the nodes, being this the reason to name this networks as "scale-free" networks (Barabasi and Oltvai, 2004; Albert, 2005). This contrast with initial models of complex networks as "random networks", where most of the nodes

had the same number of links due to the random distribution of connectivity between any pair of nodes (Erdos and Renyi, 1960). First evidence of the scale-free character of biological networks came from metabolic networks. The analysis of 43 different metabolic networks from both eukaryotes and prokaryotes revealed common topological organization in which most of the metabolites are involved in very few reactions whereas few metabolites such as pyruvate or coenzyme A acts as metabolic hubs, being involved in large number of reactions (Jeong et al., 2000; Wagner and Fell, 2001). The main consequence of this power-law distribution of connectivity is the high robustness of biological networks to random perturbations like gene inactivation or loss, most of which having little effect to organismal fitness, because these random deletions will affect with higher probability to low connected nodes and their deletion will not disturb network integrity from a topological point of view. This increases the adaptability of the systems, but makes them more vulnerable to direct attacks due to the reliance on hubs for the maintenance of network structure (Edwards and Palsson, 2000d; Wagner, 2000; Albert et al., 2000; Jeong et al., 2001). This tolerance against random mutations is in agreement with the results obtained from systematic mutagenesis experiments in organisms like *Saccharomyces cerevisiae* and *E. coli*, where it has been detected a wide tolerance to the deletion of substantial number of individual genes from their genomes (Ross-Macdonald et al., 1999; Gerdes et al., 2003). Another consequence of the scale-free topology is that any two nodes of the network can be connected by only a few links, a feature that has been named a "small-word effect" that is present in all complex networks, and that in the context of metabolic networks has been explained as a way to minimize transition times between metabolic states in response to environmental changes, allowing the metabolism to rapidly react to external perturbations (Watts and Strogatz, 1998; Fell and Wagner, 2000).

The origin and evolution of scale-free organization of complex networks has been hypothesized as consequence of network growth by preferential attachment of new nodes to highly connected nodes of the network (Barabasi and Albert, 1999). In biological systems, this has been explained by increasing network complexity by gene duplication, specially modeled in biological networks of eukaryotes (Pastor-Satorras et al., 2003; Wagner, 2003; Teichmann and Babu, 2004). However, this model of network evolution by gene duplication needs to be modified for prokaryotic organisms, where horizontal gene transfer plays a major role in their evolution. Several studies have focused in the effects of horizontal gene transfer over network evolution in prokaryotic organisms and how horizontally acquired genes are integrated in the whole structure of the biological network, revealing a minor effect of gene duplication in metabolic network structure and a major presence of horizontally acquired genes, that are integrated by preferential attachment to peripheral nodes of the network, having initially few links with other nodes (Pal et al., 2005a; Pal et al., 2005b). Studies with regulatory and protein-interaction networks have also revealed similar patterns, with recently acquired

genes having fewer regulatory and physical interactions than the average genes of the network (Wellner et al., 2007; Price et al., 2008; Lercher and Pal, 2008), and although the number of interaction increases with evolutionary time since the transfer event, the average connectivity remains significantly lower than that of non-transferred genes (Lercher and Pal, 2008). The process of reductive genome evolution of intracellular bacteria has been also studied in the context of biological networks, revealing different evolutionary patterns in different types of biological networks, with protein-interaction networks having evolved by eliminating peripheral nodes with few interactions while preserving network hubs, that generates a reduction of modularity of the network compared with that of free-living bacteria (Ochman et al., 2007; Tamames et al., 2007), whereas in regulatory networks, network hubs has been almost completely removed in highly reduced genomes like *Buchnera aphidicola* consequence of the highly stable host environment that makes precise modulation of gene expression non-necessary (Wilcox et al., 2003; Ochman et al., 2007).

This kind of topological analysis of biological networks has important intrinsic limitations in the sense that although important conclusions can be extracted from topological structure of the whole network, does not provide information about the functional profile of the biological system. This is especially true in the context of metabolic networks, where the topology of the network alone does not provide information about the metabolic phenotypes that will be expressed by the cell under specific environmental conditions. Metabolic phenotypes can be defined in terms of flux distributions through a metabolic network, that represent the amount of substrates that are converted to products by each metabolic reaction within a unit of time (Edwards et al., 2002; Barabasi and Oltvai, 2004). Their prediction and interpretation requires mathematical modeling and computer simulation over metabolic network with the final objective of develop dynamic models for the complete simulation of whole cell metabolism (Durot et al., 2009). However, detailed dynamic models of cellular metabolism require information about enzyme kinetics and regulation, that for many reactions are unknown, difficult to measure and possibly context-dependent, a fact that has limited this kinetic modeling to metabolic systems much smaller than genome-scale metabolic networks (Bailey, 2001; Smallbone et al., 2007). However, in absence of kinetic information is still possible to infer accurately the physiological capabilities of a metabolic network by an approach known as constraint-based analysis, which reduces the range of achievable flux distributions that a metabolic network can display by several physiological constraints that restricts the flux values over network reactions (Covert et al., 2001; Price et al., 2004; Becker et al., 2007). The most important is the mass-balance constraint that assumes that the metabolic system is operating at steady-state, with no accumulation or depletion of metabolites in the network, so the rate of consumption of any metabolite is equal to the rate of production and becomes only determined by the stochiometric coefficients of the metabolites in the different

reactions. Additional constraints are the bounds or limits defined by the reversibility of the reactions or additional upper and lower bounds over reactions defined by experimental measures (Schilling et al., 2000a; Palsson, 2006). Constraint-based analysis limits the allowable functional states of the metabolic network to those that satisfy the imposed constraints, and in mathematical terms this is represented as a solution space named the "convex space" that represents the phenotypic potential of an organism (Covert and Palsson, 2003; Price et al., 2003). Different mathematical methods have been developed to explore the solution space of flux distributions over the metabolic network and analyze specific properties of metabolites and reactions consequence of flux constraints. One possibility consists in uniform sampling the set of possible flux distributions to obtain an overview of the functionality of the metabolic network at steady state, being possible to evaluate these fluxes in the context of different environmental conditions (Almaas et al., 2004; Wiback et al., 2004). Alternatively, the diversity of achievable flux distributions can be determined locally for each reactions by Flux VariabilityAnalysis, that allows to identify the maximum and minimum flux achievable by each reaction in all possible metabolic states (Mahadevan and Schilling, 2003), whereas metabolic pathway analysis by methods that compute elementary modes or extreme pathways allows to determine the set of elementary and independent metabolic routes that can occur in the metabolic model based on flux distributions that respect all assumed constraints and that are minimal in the sense that cannot be decomposed in smaller elementary routes (Schuster et al., 2000; Klamt and Stelling, 2003; Palsson et al., 2003). Other approaches like Flux Coupling Analysis identify all pairs of reactions whose fluxes are always coupled at steady state (Burgard and Maranas, 2003), that in the context of bacterial evolution forms physiological modules that tend to be gained and loss together during evolution (Pal et al., 2005a). But perhaps the primary goal of genome-scale metabolic models has been the prediction of growth phenotypes under different environmental conditions by Flux Balance Analysis (FBA), which allows to determine the production capabilities and the systemic properties of a metabolic network (Bonarius et al., 1997; Schilling et al., 2000a; Palsson, 2006). This approach uses linear optimization techniques to determine the optimal flux distributions within a network that allows to maximize or minimize a particular objective function, that is normally defined in terms of a biomass production equation that reflects all metabolic compounds that the cell need to produce in order to survive under specific environmental conditions that defines the boundaries of the system in terms of energy source and respiratory conditions (Varma and Palsson, 1994a). Gene deletion analysis over the metabolic network has been also used to identify essential and non-essential genes under different growth conditions, and alternative approaches to FBA has been proposed to predict the metabolic behavior of viable gene deletions like minimization of metabolic adjustment (MOMA), or regulatory on/off minimization (ROOM), that are based on the assumption that the metabolism in a knockout system operates as close as possible to the original

network, minimizing the flux distribution distances (Segre et al., 2002; Shlomi et al., 2005).

In the context of bacterial evolution, metabolic network analysis has been used to examine the process of genome reduction in minimal genomes of bacterial endosymbionts of insects, revealing that is possible to predict the gene content of reduced genomes with high sensitivity from the analysis of the viability of multiple gene knockouts in terms of biomass production over functional metabolic network of close free-living relative by FBA, although there is also variability in the space of possible reduced networks for a given environmental condition (Pal et al., 2006b). Similar analysis over a minimal metabolic network has revealed significant differences in the network robustness when topological structure and functional capabilities are compared, with an apparent robustness of minimal networks from a topological point of view that contrast with a significant fragility when the same deletions are analyzed in terms of their effect over the elementary modes of the network (Gabaldon et al., 2007). FBA have been also used to evaluate the fitness contribution of horizontally transferred genes in bacterial metabolic networks, revealing that the contribution of these genes is mostly environmental-specific, allowing bacterial adaptation to new environments rather than optimization in fixed environments (Pal et al., 2005a).

In the context of this thesis, the process of genome reduction will be studied through a systems-biology approach using as model organism *Sodalis glossinidius*, in order to quantify the effects of gene inactivation over the metabolic capabilities of the bacteria by comparing its ancestral metabolic network with its functional one. In addition, simulation experiments of genome reduction will be carried out to determine possible pathways of future evolution in the context of the specific environment of their tsetse host. The recent transition of *Sodalis glossinidius* to an endosymbiotic lifestyle and its close relationship with *E. coli*, one of the best characterized organisms for which different reconstructions of their metabolic networks are available, provides an opportunity to evaluate the complete evolutionary process of reductive evolution from the free-living ancestor with a completely functional metabolism capable of growth under nutrient-limited conditions to an extremely reduced bacteria with highly streamlined functional capabilities that depends on the nutrients provided by their eukaryotic host for their survival. Simulated reductive evolution allows to characterize essential and non-essential genes needed for system survival and to study if the essential character of a given gene in terms of its role in the functional profile of the system is reflected in their patterns of sequence evolution.

2. Objectives

Objectives

The present thesis is focused in the study of microbial genome evolution, specially centered in the reductive evolution process experienced by bacterial endosymbionts of insects at different stages of the genome reduction process. The general objectives of this thesis deals with three main evolutionary aspects related with the reductive evolution process: 1) Evolution of gene order and genome rearrangements in γ-proteobacterial genomes including bacterial endosymbionts at different stages of the genome reduction process, 2) study of gene inactivation and mobile genetic elements proliferation in initial stages of the genome reduction process and 3) systems biology approach to the complete process of genome reduction since a hypothetical free-living ancestor to minimal metabolic systems. These general objectives are accomplished through the combination of different bioinformatic approaches based on comparative genomics and network analysis to analyze specific objectives that are:

1) Reconstruct gene order phylogenies based on a shared gene set between γ-proteobacterial genomes and comparison with phylogenetic reconstructions based on sequence data.

2) Analyze the patterns of gene order evolution across different evolutionary lineages including bacterial endosymbiotic lineages at different stages of the genome reduction process.

3) Characterization of pseudogenes and mobile genetic elements in the genome of *Sodalis glossinidius*, the secondary endosymbiont of tsetse flies, and evaluation of their impact over genome structure and functional capabilities of this recently stablished symbiotic association

4) Evaluation of the functional capabilities of *Sodalis glossinidius* at different stages of the genome reduction process since their transition from a free-living bacteria to a host-dependent lifestyle by metabolic network reconstruction and Flux Balance Analysis

5) Prediction of the possible future evolution of *Sodalis glossinidius* in the context of the reductive evolution process by means of reductive evolution simulations over their functional metabolic network and analysis of the correlation between gene essenciality and the patterns of sequence evolution.

3. Genome rearrangement distances and gene order evolution in γ-proteobacteria

Chapter 3

3.1 INTRODUCTION

The availability of genomic data has allowed the study of different features concerning both chromosomal organizational structure and the generation of genetic variability in bacterial populations across evolution. The organization of bacterial chromosomes are determined by different constraints that has been revealed by comparative genomics, like the organization of functionally related genes into polycistronic units named operons allowing the development of different mechanisms for the regulation of gene expression in functionally related genes (Jacob and Monod, 1961), although alternative theories for their maintenance postulate that their conservation across evolution is consequence of a selfish model for operon evolution where the clustering of functionally related genes results in their easier spread by horizontal gene transfer (Lawrence and Roth, 1996). In addition, co regulation of gene expression and conservation of gene organization in terms of conservation of gene neighborhoods over evolutionary time is also observed at higher order than operons, suggesting that a proper description of prokaryotic genome organization must include also supra-operonic organization (Lathe, III et al., 2000; Korbel et al., 2004). The organization of bacterial chromosomes, with a single origin of replication, generates also a gene dosage gradient from the origin towards the terminus of replication that allows over expression of genes located near the origin of replication, a feature that is amplified in fast-growing bacteria where multiples rounds of replication are generated during exponential growth (Schmid and Roth, 1987; Sousa et al., 1997). Comparative genomics have also revealed that genes are also more often coded in the leading than in the lagging strand in bacterial chromosomes, that has been explained as a mechanism to avoid head-on collisions between the replication fork and the RNA polymerase, and this gene strand bias is more pronounced in the case of essential genes, due to the fact that head-on collisions leads to truncated transcripts that results in non-functional proteins and complexes, a fact that is particularly important for essential genes and that explains the fact that 94% and 76% of essential genes from *B.subtilis* and *E.coli* respectively are coded in the leading strand (Rocha and Danchin, 2003b; Rocha and Danchin, 2003a).

Opposite to this organizational constraints in bacterial chromosome structure, bacterial genomes possess hundreds or thousands of repeated elements that are able to recombine by means of homologous or illegitimate recombination (Achaz et al., 2003; Rocha, 2003a). These repeats are part of duplicated genes, regulatory elements or insertion sequences and are constantly created by means of recombination, horizontal gene transfer, and transposition, and deleted by further recombination or by accumulation of point mutations (Achaz et al., 2003). The compact nature of bacterial chromosomes suggests that only selective pressures or self-replication of repeats allows their stable maintenance. Positive selection on duplicated genes can result from gene dosage effects that allows faster cell growth or

from the generation of genetic variability (Romero and Palacios, 1997; Klappenbach et al., 2000; Kresse et al., 2003). In addition, intra-chromosomal recombination appears associated with adaptation strategies to heterogeneous stresses in changing environments, generating genetic variation by the generation of new chimerical genes or new genome architectures (Rocha, 2004). This creates an evolutionary conflict or trade-off between the conservation of genome organization and the generation of genetic variability by means of recombination processes, that depend on ecological, genetical, and physiological characteristics of bacterial strains in terms of lifestyles, population sizes, recombination mechanisms, and growth rates (Rocha, 2004).

The construction of detailed genetic maps in several bacterial species soon revealed that the overall gene order was not conserved over a long evolutionary time scale. The sequencing of the complete genome in many bacterial species and strains clearly showed that closely related species had accumulated fewer rearrangements than the distant ones. However, these tendencies presented exceptions, with some phylogenetic lineages showing remarkable conservation and others extensive genome rearrangements (Nadeau and Sankoff, 1998).

There are four types of changes that may affect the order of the genes on the bacterial genome. First, inversions and translocations are frequently detected when the genome of closely related species are compared (Hughes, 2000). Inversions are frequently symmetric around the axis of DNA replication (Tillier and Collins, 2000; Eisen et al., 2000), while translocations may be intra- or interchromosomal (in bacteria like *V. cholerae*, with two chromosomes). Second, genes may be removed in a single event or as a consequence of a process of progressive disintegration, which produces gaps when the genomes of two species are compared (Andersson and Andersson, 2001; Silva et al., 2001; Moran, 2002). Third, horizontal gene transfer (HGT), considered one of the major forces shaping the evolution of prokaryotic genomes (Koonin and Galperin, 1997; Garcia-Vallve et al., 2000; Ochman et al., 2000; Boucher et al., 2003), may produce insertions throughout the genome. The extension of this phenomenon has been reported to be as high as 20-30% within species in genome comparisons, with many inserted foreign DNA segments in the genome, like in uropathogenic strains of *Escherichia coli* (Welch et al., 2002). However, HGT varies considerably among species with some of them being completely refractory to this phenomenon, like the reduced genomes of bacterial endosymbionts of insects (Ochman et al., 2000). Finally, partial duplications of the genome may produce redundant genomic segments. Genome rearrangements have been studied in several bacterial groups. Gamma (γ)-Proteobacteria is one of them, with inversions as one of the most frequent rearrangement type in interspecies whole genome comparisons (Hughes, 2000). The study of this bacterial group is also very interesting because more than 200 genomes have been sequenced with different degrees of relatedness at taxonomic level.

Between them, the small genomes of the bacterial endosymbionts of aphids *Buchnera aphidicola* from different host species (Shigenobu et al., 2000; Tamas et al., 2002; van Ham et al., 2003; Perez-Brocal et al., 2006; Moran et al., 2009), of primary and secondary endosymbionts of tsetse flies *Wigglesworthia glossinidia* and *Sodalis glossinidius* (Akman et al., 2002; Toh et al., 2006), of primary endosymbionts of different species of carpenter ants (Gil et al., 2003; Degnan et al., 2005), or the smallest bacterial genome corresponding to the primary endosymbiont of psyllids *Carsonella rudii* (Nakabachi et al., 2006) have been reported in recent years. The case of *B. aphidicola* has been specially studied. This bacterial endosymbiont is transmitted maternally from mothers to their descendents in their aphid hosts (Baumann et al., 1995), and the time of divergence of the first three sequenced strains has been proposed to be as long as 164 Myr (von Dohlen and Moran, 2000). During this period of time, an almost complete genome stasis was observed, with none HGT or gene duplication event detected and with only four small rearrangements that differentiated the genomes of *B. aphidicola* from *Baizongia pistaciae* from those of the strains of *Acyrthosyphon pisum* and *Schizaphis graminum* that corresponds to two small inversions affecting to one and six genes and two small translocations from two plasmids affecting to two and four genes (Tamas et al., 2002; Silva et al., 2003; van Ham et al., 2003). However, despite this genome stasis, at least 164 gene loss events had placed during the evolution of this three lineages (Silva et al., 2003; Gomez-Valero et al., 2004a). This period of genome structural stability contrasts with the remaining evolution of the *B .aphidicola* lineage after its divergence from its free-living relative *Escherichia coli*, with chromosomal rearrangements and more than one thousand lost genes by large and small deletion events (Moran and Mira, 2001; Silva et al., 2001).

During recent years, several attempts have been made to produce a γ-proteobacterial phylogeny and to establish the relationships among the different bacterial endosymbionts and the remaining γ-proteobacterial genomes, specially the *Enterobacteriaceae*, that contains their closer free-living relatives. Clustering of the primary endosymbionts of aphids, tsetse flies and carpenter ants has been proposed based on 16S rDNA phylogeny (Sauer et al., 2000) or in a concatenated phylogeny inferred from 61 protein coding genes (Gil et al., 2003). The monophyly of *B. aphidicola* strains and *W. glossinidia* has been also reported based on whole genome phylogenies (Daubin et al., 2003; Canback et al., 2004). On the other hand, 16S phylogenies that rejects the monophyly of the three endosymbiotic species have been also reported (Charles et al., 2001).

Symbiotic lineages presents particular problems in phylogenetic reconstructions due to their accelerated sequence evolution since the establishment of the endosymbiotic association (Moran, 1996; Itoh et al., 2002) and their biased nucleotide and amino acid compositions (Moran, 1996; Clark et al., 1999; Shigenobu et al., 2000; Palacios and Wernegreen, 2002; Rispe et al., 2004). In fact,

the production of conflicting topologies is frequent, as with the phylogenetic analysis of *B. aphidicola* genes, where more than two thirds of them did not support their sisterhood with *E. coli* and became basal to the γ-proteobacterial phylogeny (Itoh et al., 2002; Canback et al., 2004). Several attempts may be made to solve these problems associated to gene-based phylogenies by using genome-based approaches in several phylogenomic studies (Wolf et al., 2002; Bapteste et al., 2004). This studies can be based on sequence information such as concatenated alignments (Hansmann and Martin, 2000; Brown et al., 2001), on supertree approaches (Sicheritz-Ponten and Andersson, 2001; Bininda-Emonds, 2004), or phylogenies inferred from a core of shared genes putatively free of HGT (Daubin et al., 2002), as well as on other comparative genome data such as gene content (Fitz-Gibbon and House, 1999) or gene order (Suyama and Bork, 2001).

The main objectives of this chapter are, firstly, the estimation of genome rearrangement distances based on two different measures of genome stability such as inversions and breakpoint number between pairs of completely sequenced γ-proteobacterial genomes. These distances are inferred from a subset of genes that are shared by all the genomes under study, which putatively do not contain genes originated by HGT events. This is because the objective is to analyze the movement of the genes that evolve slowly at genome rearrangement level. For that reason, genes involved in HGT events are not selected, and only genes shared by every genome were chosen. These shared genes are probably essential or functionally important, and their changes of position in the genome may be deleterious. Secondly, the lineages that had evolved faster or slower at genome structure level are determined based on gene order distances, specially focused in the lineages corresponding to bacterial endosymbionts of insects at different levels of the endosymbiotic relationship. And finally, gene order distances are used to obtain a gene-order based phylogeny and to compare these phylogenetic reconstructions with traditional sequence-based phylogenies in order to determine the monophyly of bacterial endosymbionts based on gene order evolution.

3.2 MATERIAL AND METHODS

3.2.1 Table of orthology

The first step to estimate the genome rearrangement distances between genome pairs is the reconstruction of a table of orthology. In this study, thirty-one genomes of γ-proteobacteria has been selected for the analysis that includes different bacterial endosymbionts at initial and final stages of the adaptation to intracellular lifestyle like the genome of the secondary endosymbiont of tsetse flies *Sodalis glossinidius* or three different strains of the primary endosymbiont of aphids *Buchnera aphidicola* respectively (Table 3.1). The aim of the table of orthology is to include only those

orthologous genes that were presents in all the genomes, either as gene or pseudogene, and to remove any gene acquired by horizontal gene transfer in at least one of the compared genomes. Pseudogenes were included into the table of orthology because in order to estimate rearrangement distances between genomes the only important thing to consider is to known the position of the gene (or pseudogene) and their transcriptional orientation, independently of their functional status.

The analysis starts with the tables of orthology that were obtained from two previous studies. In the first, the genomes of *B. aphidicola Bap*, *E.coli* K12, and *Vibrio cholerae* were compared, removing paralogous genes originated by duplication and xenologous genes originated by horizontal gene transfer (Silva et al., 2001). In the second, the genomes of five insect bacterial endosymbionts were compared to detect the orthologous genes (Gil et al., 2003). The presence of those genes detected in the seven previous genomes as either genes or pseudogenes is searched for in the remaining γ-proteobacterial genomes. This analysis has been carried out in the Microbial Genome Database for Comparative Analysis (MBGD) (Uchiyama, 2003) with a maximum BLAST score of 0.0001 and a phylocut value of 0.4. The gaps, defined as absence of a gene, detected in several genomes were treated in several ways to confirm the absence of an orthologous gene or pseudogene, and in the case of absence confirmation the corresponding gene were removed from the orthologous table. Genomes were searched for a similar sequence to the absent gene with their corresponding amino acid encoded sequence by using the TBLASTN algorithm (Altschul et al., 1997) in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2004). In some cases, the detection of a sequence with significant similarity indicated incorrect annotations while, in others, the presence of an unannotated pseudogene. Both situations were analyzed in detail before taking the final decision to maintain or remove the gene from the table.

When the MBGD comparisons rendered more than one gene in any of the genomes, the phylogenetic tree and the genomic context of the duplicated genes were analyzed in order to decide which was the orthologous or paralogous gene and whether the gene was to be maintained in the table or orthology. In the case of two true orthologous genes being present in a genome as consequence of recent duplication, one of the copies at random was retained and the other was removed.

Tree	Species name	Acc.Numb.
BAp	*Buchnera aphidicola str. APS (Acyrthosiphon pisum)*	NC_002528
BBp	*Buchnera aphidicola str. Bp (Baizongia pistaciae)*	NC_004545
BSg	*Buchnera aphidicola str. Sg (Schizaphis graminum)*	NC_004061
bfl	*Candidatus Blochmannia floridanus*	NC_005061
ecc	*Escherichia coli CFT073*	NC_004431
eco	*Escherichia coli K12*	NC_000913
ecs	*Escherichia coli 0157-H7*	NC_002695
ece	*Escherichia coli 0157:H7 EDL933*	NC_002655
hdu	*Haemophilus ducreyi 35000HP*	NC_002940
hin	*Haemophilus influenzae Rd KW20*	NC_000907
pae	*Pseudomonas aeruginosa PAO1*	NC_002516
ppu	*Pseudomonas putida KT2440*	NC_002947
pst	*Pseudomonas syringae pv. tomato str. DC3000*	NC_004578
pmu	*Pasteurella multocida Pm70*	NC_002663
sfl	*Shigella flexneri 2a str. 301*	NC_004337
sfx	*Shigella flexneri 2a str. 2457T*	NC_004741
son	*Shewanella oneidensis MR-1*	NC_004347
stm	*Salmonella typhimurium LT2*	NC_003197
stt	*Salmonella enterica subsp. enterica serovar Typhi Ty2*	NC_004631
sty	*Salmonella enterica subsp. enterica serovar Typhi str. CT18*	NC_003198
vch	*Vibrio cholerae O1 biovar eltor str. N16961*	NC_002505
vpa	*Vibrio parahaemolyticus RIMD 2210633*	NC_004603
vvu	*Vibrio vulnificus CMCP6*	NC_004459
wgl	*Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis*	NC_004344
xac	*Xanthomonas axonopodis pv. citri str. 306*	NC_003919
xcc	*Xanthomonas campestris pv. campestris str. ATCC 33913*	NC_003902
xfa	*Xylella fastidiosa 9a5c*	NC_002488
xft	*Xylella fastidiosa Temecula1*	NC_004556
ype	*Yersinia pestis CO92*	NC_003143
ypk	*Yersinia pestis KIM*	NC_004088
sgl	*Sodalis glossinidius str. Morsitans*	NC_007712

Table 3.1: *γ-proteobacterial genomes included in the study*

The decision to remove a gene when a putatively horizontal gene transfer event had taken place in at least one genome was difficult because, after more than 600 My of evolution in the lineage of the γ-proteobacteria, some conflicting phylogenies in fast evolving genes were not related to horizontal gene transfer events but to problems associated with the phylogenetic methods employed to infer the actual topology.

In order to detect possible events of horizontal gene transfer, all genes included in the table of orthology were searched among the putatively horizontal gene transfer events detected in the Horizontal Gene Transfer Database (HGT-DB) (Garcia-Vallve et al., 2003). Most of the genes in the table of orthology were not included in the HGT-DB, and only 4 genomes (*B. aphidicola BSg*, *Pasteurella multocida*, *Pseudomonas aeruginosa* and *Xylella fastidiosa*) showed more than two genes in these lists. The phylogenetic trees of these genes were reconstructed and their genomic context was looked at MBGD but any of them were able to be removed with great confidence. In fact, the six genes observed in *B. aphidicola BSg* are an artifact associated to the special base composition of this species, because *B. aphidicola* is refractory to horizontal gene transfer events due to their isolated environment and their strict vertical transmission from the mother to their descendents (Tamas et al., 2002; Silva et al., 2003). To additionally confirm the absence of horizontal gene transfer events in our table of orthology, the included genes were searched in the results of a recent analysis where candidate xenologous genes were detected with several criteria including phylogenetic validation (Medrano-Soto et al., 2004). The final table of orthology contains 244 genes presents in the thirty-one analyzed γ-proteobacterial genomes, and was putatively free of xenologous genes consequence of horizontal gene transfer events, although it is not impossible that a very small number of them were not detected.

3.2.2 Breackpoint, inversions and amino acid substitution distances

Two genome rearrangements distances were estimated between genome pairs. The disruption of gene order, named breakpoint, was introduced early into computational studies of genome evolution (Nadeau and Taylor, 1984). This concept served to estimate the breakpoint distance (BP), defined as the number of pairwise gene adjacencies present in one genome but absent in the other. Finally, BP distances were used to reconstruct phylogenies by several distance based methods. In this study, for each specie or strain, its genome can be represented as a small circular molecule composed of 244 genes in which the order and transcriptional orientation is the same as that observed in the real complete genome. When comparing two genomes, represented as circular molecules composed by the 244 genes common to the thirty one genomes, one of them is taken as reference genome (genome A) with its genes ordered and compared with the other (genome B). The

first step in the comparison is to change the sign of genes in genome B when they are not in the same transcriptional orientation as the reference genome. The next step is to search for whether the adjacent genes in the genome A are also maintaining their adjacency in the genome B. Taking two adjacent genes *g1* and *g2* in the genome A, a breakpoint is considered to have occurred when they do not appear consecutively in genome B as either the pair (*g1,g2*) or (*-g2,-g1*). All the breakpoints between two genomes were counted and the BP distance was inferred as the number of breakpoints transforming one genome into the other. This distance may be normalized by dividing this number by the total number of genes (244 genes).

The second rearrangement distances is the inversion distances (INV). INV distances are defined as the minimum number of inversions (reversals) needed to transform one genome into the other. INV distances were estimated in the genome rearrangement web server GRIMM (Tesler, 2002). The genomes were considered as unichromosomal circular signed permutations. INV distances can be also normalized by dividing the inversion number between two genomes by the total number of genes (244 genes).

In order to compare genome rearrangement distances (INV distances and BP distances) with sequence-based distances, amino acid substitution distances were calculated assuming a specific empirical model of protein sequence evolution, using a substitution matrix with scores for all the possible exchanges of one amino acid for another inferred from the protein sequence dataset. For this purpose, the protein sequences encoded by 10 genes conserved in the 31 γ-proteobacterial genomes were used (*rpoC, rpoB, rho, rpoA, rpsC, nusG, rpsG, rplP,* and *rpsK*). These genes were selected by using the following protocol. First, slowly evolving proteins were searched by selecting those that presents more than 80% of amino acid identity between *B. aphidicola Bap* and *E.coli* K12. Second, those genes involved in information transfer were selected among those selected in the first step, and finally, the 10 genes with the largest amino acid sequence were finally selected. The amino acid sequences of these protein coding genes were obtained from MBGD and were aligned with the program CLUSTALX using the default parameters (Thompson et al., 1997). The resulting alignments were edited using G-BLOCKS program (Castresana, 2000) in order to select the most conserved sites of the alignment, removing highly variable sites and sequence gaps with the goal of selecting the amino acid positions with the greater phylogenetic information. Parameters were fixed in 19 identical residues for a conserved position, 22 identical residues for flanked position, 1 for the maximum number of contiguous nonconserved positions, and 10 for the minimum block sizes, so the program retrieves blocks of the alignment of a minimum size of 10 positions that are delimited by flanking positions and that contains conserved positions with a maximum of 1 contiguous nonconserved position.

The 10 alignments edited with G-BLOCKS were concatenated into a single one alignment composed of 3670 aminoacid positions, and this alignment is used to infer amino acid substitution distances by maximum likelihood (ML) by using TREEPUZZLE 5.2 (Strimmer and von, 1996; Schmidt et al., 2002) with the VT model of protein evolution (Muller and Vingron, 2000).

3.2.3 Relative inversion distances

To estimate whether the rates of rearrangements by inversions were constant among lineages, lineages were compared as follows. Considering two species A and B which diverged from a common ancestor O and an outgroup specie C, the ratio d_{AO}/d_{BO} were determined. To calculate this ratio, the pair-wise INV distance among species A, B, and C were used. To estimate if the lineage A had had a different inversion rate from lineage B since their divergence from their common ancestor O, the ratio d_{AO}/d_{BO} were estimated by the following formula:

$$\frac{(dAC - dBC + dAB)}{(dBC - dAC + dAB)}$$

Between all possible pairwise comparisons of A and B species, as species A were selected one representative of each endosymbiont specie, *Vibrionaceae* species, and *Pasteurellaceae* species (*BAp, bfl, wgl, sgl, vch, vvu, vpa, hdu, hin* and *pmu*), that were compared with the group of free-living enterics, that are considered as species B (*eco, sfx, sfl, stm, stt, sty, ype* and *ypk*), with the objective of determine whether the former lineages were evolving, on average, faster or slower than the latter. The genomes of *pae, ppu, pst* and *son* were used as outgroups (species C). The ratio d_{AO}/d_{BO} for each specie A is estimated as the average of those obtained with the four outgroup species. *BSg* and *BBp* were not included because the order of the 244 genes in their genomes was identical or almost identical to that from *BAp*. The genomes of *ece, ecs* and *ecc* were also not included because they were identical in gene order to the one from *eco.*

3.2.4 Phylogeny

The INV distance matrix between genome pairs obtained with GRIMM was used to reconstruct the phylogenetic tree of γ-proteobacteria using the Fitch-Margoliash (FM) (Fitch and Margoliash, 1967) and Neighbor joining (NJ) (Saitou and Nei, 1987) algorithms implemented in the FICH and NEIGHBOR programs respectively from the PHYLIP software package (http://evolution.genetics.washington.edu/phylip.html). The input order of species was randomized and global rearrangements were made to ensure that the optimum phylogenetic tree was obtained and that no species had fallen into a suboptimal region of the space of all possible trees.

Phylogenetic reconstruction with the BP distance matrix was carried out with the same methods and conditions as those implemented with the INV distance matrix.

To asses the reliability of the phylogenetic reconstruction with BP and INV distances is not possible to carry out a conventional bootstrap analysis because in this analysis we do not work with traditional sequence data where nucleotide or amino acid positions acts as independent characters that could be randomly selected as in a classical bootstrap analysis. When we are dealing with genome rearrangements, we have a single character that corresponds to the bacterial chromosome, and each of the possible combinations of genes in the chromosome represents different states of the character, so randomly altering the order of genes in the chromosome would render the data meaningless. In order to solve this problem, a Jackknife resembling method was applied instead of traditional Bootstrap that consisted of the random selection of 122 genes out of the initial 244, and the removal of the remaining genes from the genomes. Finally, new signed permutations for each genome are generated but this time for the randomly selected 122 genes instead the 244 original genes. Once the 31 genomes of 122 genes are obtained, the original analysis to obtain the INV and BP distances was carried out. A total of 100 Jacknife random samples were carried out, obtaining 100 pairwise distance matrices for inversions and breakpoints. These 100 matrices were loaded into the FITCH and NEIGHBOR programs to obtain the 100 breakpoints and inversion distance phylogenetic trees and finally, the CONSED program from the PHYLIP software package was used to obtain a majority rule consensus tree with the numbers of each node reflecting the percentage of times the clade defined by that node appears represented in the 100 Jacknife trees. These values were assigned to the nodes of the original breakpoints and inversions distances phylogenetic trees reconstructed from the original set of 244 genes.

The concatenated amino acid alignment described in the breakpoint, inversion, and amino acid substitution distances section was used to reconstruct a sequence-based phylogenetic tree for the 31 γ-proteobacterial genomes by ML using TREEPUZZLE 5.2 software and the quartet puzzling algorithm (Strimmer and von, 1996; Schmidt et al., 2002). Options included exact parameter estimation by quartet puzzling plus NJ, VT model of amino acid sequence evolution (Muller and Vingron, 2000), heterogeneity in the rates of aminoacid substitutions across the alignment (1 invariable and 8 gamma rates) and 1000 puzzling steps. Because a few nucleotide indeterminations were detected in the sequence of selected genes from *E. coli* CFT073 and *E. coli* 0157:H7 EDLP33, these genomes were removed from the phylogenetic analysis.

3.3 RESULTS

3.3.1 Orthologous genes shared by γ-proteobacterial genomes

The first step in the analysis was the reconstruction of a table of orthologous genes presents in the 31 γ-proteobacterial complete genomes under study. An ortholog of each gene must be present in every genome in order to be included in the table either as gene or pseudogene. We tried to remove genes that had been putatively acquired by horizontal gene transfer in any of the genomes, a very difficult task due to the fact that many genes can produce abnormal phylogenies for several reasons do not related to horizontal gene transfer. For this reasons we are very conservative, considering only as horizontally transferred those genes presents in specific database focused on horizontal gene transfer events and detected by both compositional and phylogenetic methods as horizontally acquired (Garcia-Vallve et al., 2003; Medrano-Soto et al., 2004).

3.3.2 Correlation of breackpoints and inversion distances

A matrix with the BP distances for the 31 γ-proteobacterial genomes was obtained by the protocol described in material and methods section. The order of the 244 in the four strains of *E. coli* included was identical, leading to BP distances equal to zero among these collinear genomes. The same situation arose when *BAp* and *BSg* strains of *B. aphidicola* were analyzed. In the opposite side of the distance range, the maximum BP distances were observed between *H. ducreyi* and *X. fastidiosa*, with 162 breakpoints, that corresponds to a normalized BP distance of 0.664. A matrix was also constructed for the INV distances between γ-proteobacterial genomes inferred from the genome rearrangement web server GRIMM (Tesler, 2002), with zero as the minimal INV distance, that is observed between the same pairwise comparisons as with BP distances, and with a maximum value for the INV distance between *H. ducreyi* and *X. axonopodis* with 159 inversions between them, that corresponds to a normalized INV distance of 0.652.

These two different types of genome rearrangements distances were expected to slightly underestimate the actual number of rearrangement events that are taking place in the evolution of the different γ-proteobacterial genomes because of the possibility of multiple breakpoints in the same place of the genome or because the optimal scenario to transform one genome into another via inversion events inferred from parsimony principle may provide a number of steps which are smaller that the actual one. However, several data points out that these differences will be practically negligible, because simulations comparing the estimated and the actual rearrangement distances between genomes show very similar values for the range of normalized distances included in this study (Bourque and Pevzner, 2002; Moret et al., 2002b). In fact, the expected underestimation for the maximum value of INV distance detected between *H. ducreyi* and *X. axonopodis* in the theoretical

simulations for INV distances was around three inversions (Bourque and Pevzner, 2002).

Finally, a great correlation was detected between BP and INV distances, with a correlation coefficient among them of 0.996 (Figure 3.1). This strong correlation indicated that inversions are the most common genome rearrangement event affecting the evolution of γ-proteobacterial genomes. Although other types of genome rearrangements such as transpositions may take place, their contribution would be very slight compared to that of inversions in this gene subset.

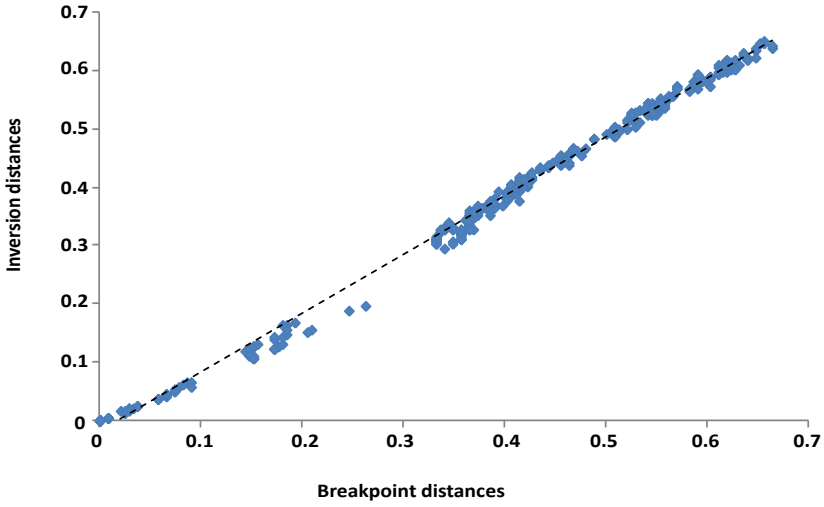


Figure 3.1: *Comparison between normalized Breakpoint and Inversion distances between pairs of γ-proteobacterial genomes. The discontinuous line represents the regression line between the two gene order distance measurements (correlation coefficient r = 0.998)*

3.3.3 Genome rearrangement versus amino acid substitution distances through γ-proteobacterial evolution

In order to detect whether genome rearrangement rates had been constant both between and within different evolutionary γ-proteobacterial lineages, a comparison was carried out between sequence-based distances and BP and INV distances. Pairwise ML-distances were computed between γ-proteobacterial genomes from an amino acid concatenated alignment of 10 protein coding genes comprised of 3670 amino acid positions, and the phylogenetic tree was reconstructed by ML (Figure 3.2).

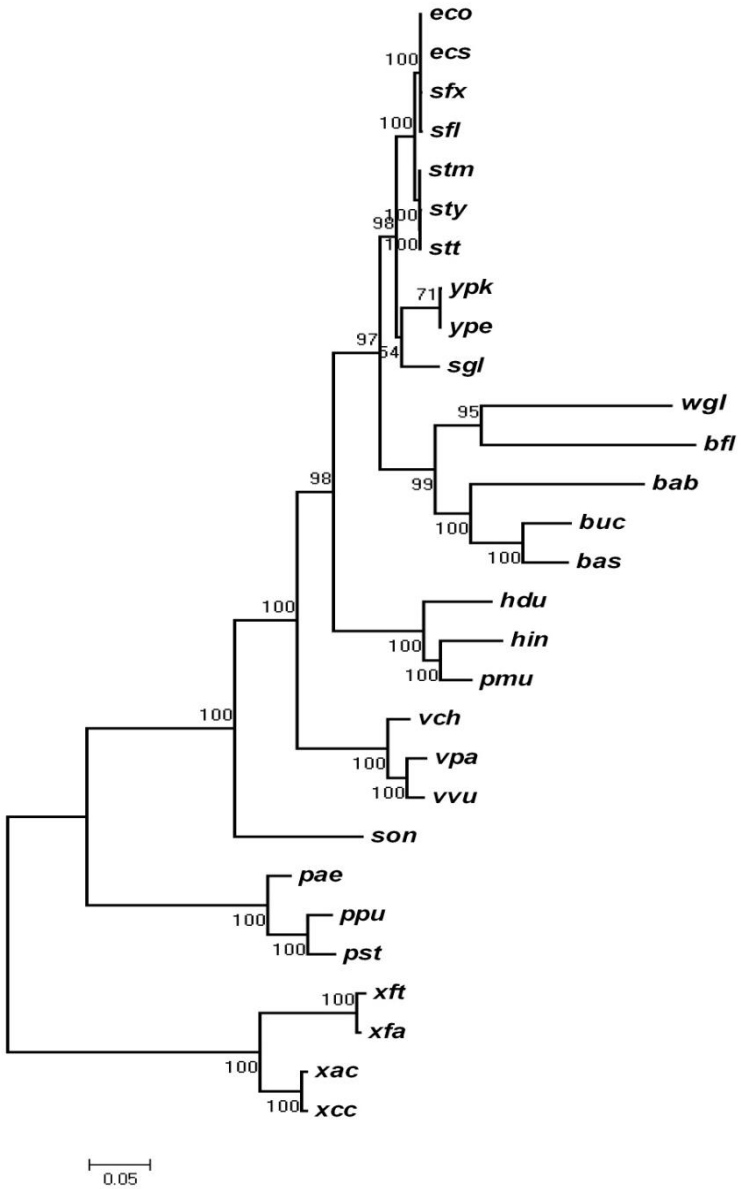


Figure 3.2: *Phylogenetic relationships between the 31 γ-proteobacterial genomes inferred by maximum likelihood using Tree-Puzzle 5.2 and a concatenated alignment of amino acid sequences from the proteins encoded by rpoC, rpoB, rho, rpoA, rpsC, rpsD, nusG, rpsG, rplP, and rpsK genes. The values in the nodes of the tree represent the proportion of quartets supporting the corresponding inner branch, as determined by the quartet puzzling method. Species abbreviations are described in table 3.1.*

The tree topology inferred from this concatenated alignment showed that the lineages corresponding to the primary endosymbionts of aphids, tsetse flies, and carpenter ants (*B. aphidicola* strains, *W. glossinidia*, and *B. floridanus* respectively) forms a well- supported monophyletic group. The clustering of these species in phylogenetic analysis was in agreement with previous observations (Sauer et al., 2000; Gil et al., 2003). The ML amino acid substitution distances were compared to BP and to INV distances (Figure 3.3; BP distance plot is not shown because it is almost identical to the INV distance plot, due to the great correlation between both genome rearrangement distances). In general, the progressive increase in sequence distances between genome pairs is associated with the increase in their genome rearrangement distances, although several groups of distances presented abnormal behavior compared with the central trend. In the first group, *H. ducreyi*, *H. influenzae*, and *P. multocida* showed significatively large BP and INV distances in all pair-wise genome comparisons, indicating that the lineage of the *Pasteurellaceae* is evolving at fast rearrangement rate than the rest of γ-proteobacterial genomes. The second lineage that appears evolving at slightly higher rearrangement rates than the central trend is *S. glossinidius*, the secondary endosymbiont of tsetse flies, that shows an apparent acceleration of rearrangement rates in comparison to the sequence evolution rates but maintaining the proportionality in both evolutionary distances in contrast to what happens with the lineage of *Pasteurellaceae*, where saturation in genome rearrangement distances have occur in light of their narrow range of INV distances (0.55-0.66 normalized INV distances) in comparison with the range of amino acid sequence distances (0.05-0.35). A third group was constituted by the distances between *Bl. floridanus* and free-living enterics. The reason for the abnormal position of these distances was the acceleration of the sequence substitution rate in the endosymbiotic bacterial lineage, which led to large ML distances. A fourth group was constituted by the distances among the three *B. aphidicola* strains that show little or no rearrangements but high amino acid substitution rates.

These results indicate that, for some species, is it possible to observe, for a specific group of protein coding genes, a constant increase in the number of rearrangements and amino acid substitutions over time. However, others behave heterogeneously, with situations of an almost null genome rearrangement rate like in the lineage of *B. aphidicola* together with other lineages with an extremely high rearrangement rate like the *Pasteurellaceae*. The same situation can be observed for the sequence substitution rates, which are extremely high in the case of endosymbiotic lineages.

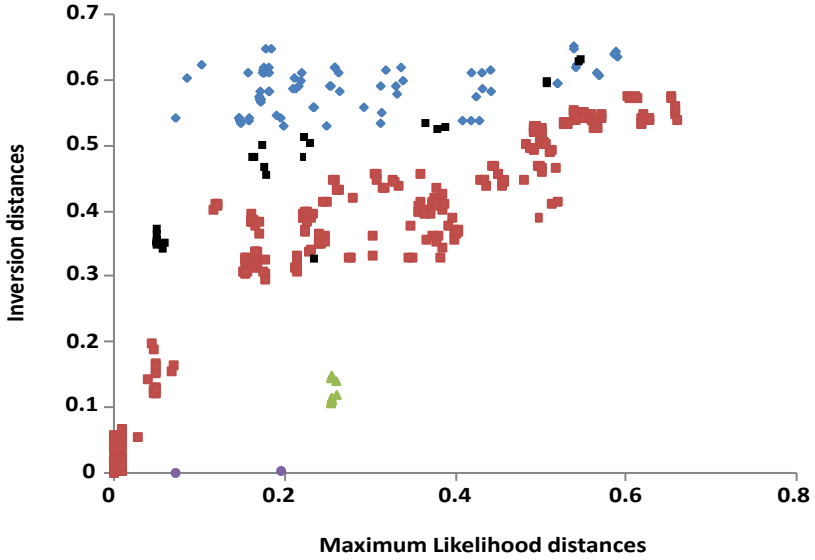


Figure 3.3: *Comparison of maximum likelihood distance to normalized Inversion distance between pairs of γ-proteobacterial genomes. Symbols represent the different groups observed in the graph. Blue rhombuses represent the distances from hin, hdu, and pmu to the rest of γ-proteobacterial genomes. Green triangles represent the distances from bfl to the genomes of free-living enterics (eco, ecs, sfx, sfl, stm, stt, sty, ype, and ypk). Purple circles are distances among the three B. aphidicola strains studied (BSg, BAp, and BBp). Black squares represent the distance from S. glossinidius to the rest of γ-proteobacterial genomes. Red squares represent the rest of the pairwise distance comparisons. Species abbreviations are specified in Table 3.1.*

3.3.4 Relative inversion distances

In order to confirm the faster evolutionary rearrangement rate of *Pasteurellaceae* and S. *glossinidius* and to characterize the situation of the three lineages of endosymbiotic bacteria compared with other free living *Enterobacteriaceae*, a relative rate test approach was carried out comparing the INV distance rates in the lineages of *B. aphidicola BAp*, *B. floridanus*, *W. glossinidia*, *S. glossinidius*, *V. cholerae*, *V. vulnificus*, *V. parahaemolyticus*, *H. ducreyi*, *H. influenzae* and *Pa. multocida* with those of the free living enteric bacteria *E. coli* K12, *S. thypimurium* LT2, *S. enterica* subsp. *Enterica* serovar Thipi str. CT18, *S. enterica* subsp. *Enterica* serovar Typhi Ty2, *S. Flexneri* 2a str. 301, *S. Flexneri* 2a str. 2457T, *Y. pestis* CO92 and *Y. pestis* KIM, after the divergence of the two clusters. *S. oneidensis* MR-1, *P. aeruginosa*, *P. putida*, and *P. syringae* were used as outgroup species (Figure 3.4).

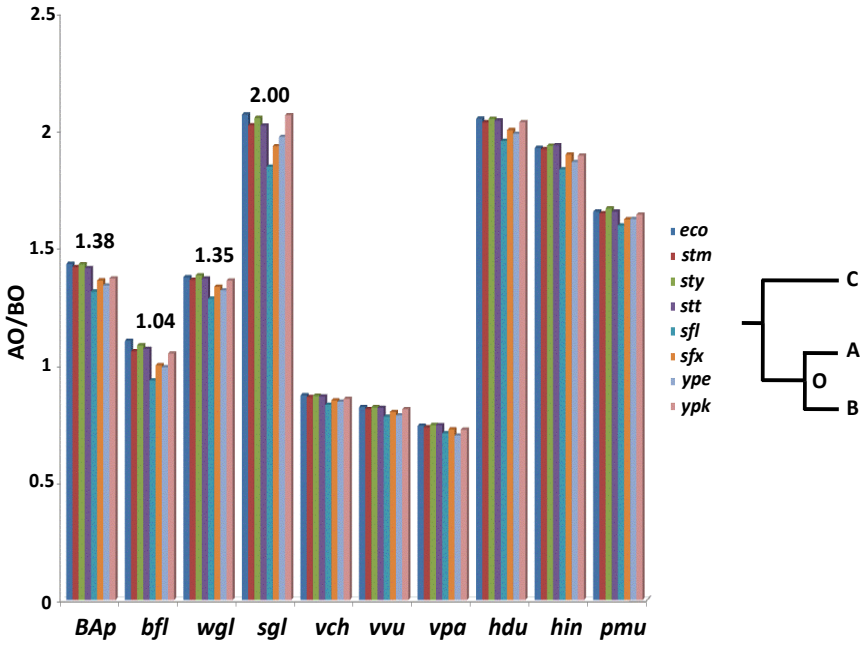


Figure 3.4: *Relative inversion distance rates. The inversion distance in the branch from the common ancestor O to the problem species A (d_{AO}) was compared with that of the branch from common ancestor O to the reference specie B (d_{BO}) by using the inversion distances from problem and reference species to the out-group specie C (see right tree). Problem species are BAp, bfl, wgl, sgl, vch, vvu, vpa, hdu, hin, and pmu. Reference species are defined in the colour legend, and out-group species are son, pae, ppu, and pst. The height of each column corresponds to the d_{AO}/d_{BO} ratio estimated as the average ratio of each A-B comparison over the four out-group species. The number over columns block of BAp, bfl, wgl, and sgl corresponds to the average d_{AO}/d_{BO} ratio after the comparison with the eight reference species and four out-group species. The species name for each abbreviation may be found in Table3. 1*

The results of these analyses showed that the branches leading to *H. ducreyi*, *H. influenzae* and *P. multocida* were evolving at significantly faster rearrangement rates than free living enterics, at average relative rates of 2.02, 1.90, and 1.64 respectively. This agrees with their abnormal position in the plot comparing sequence and genome rearrangement distances (Figure 3.3). A similar situation is observed for *S. glossinidius*, the secondary endosymbiont of tsetse flies, that are evolving at an average relative rate of 2.00, similar to *H. ducreyi* that also explains its acceleration in genome rearrangement rates in comparison with amino acid sequence distances compared with the rest of γ-proteobacterial genomes out of

Pasteurellaceae observed in Figure 3.3. The behavior of the three lineages of primary endosymbionts was not identical. While *B. aphidicola BAp* and *W. glossinidia* were evolving, on average, to a slight higher rate than free-living enterics (1.38 and 1.35 respectively), *B. floridanus* has evolved to almost the same rate as free-living enterics (1.04). Finally, *Vibrio* lineages, included as controls, have evolved to a slightly smaller rate than free-living enterics (*V. cholerae*, 0.86; *V. vulnificus*, 0.81; *V. parahaemolyticus*, 0.73).

To interpret the results of the relative genome rearrangements rate test correctly, it is important to bear in mind that the inferred rearrangement rates are averaging the total number of chromosomal rearrangements that have taken place after divergence from the free-living enteric cluster. In the case of the primary endosymbionts of aphids, it is known that during the last 100-150 My of evolution, the *B. aphidicola* genomes have experienced a minimal number of genome rearrangements (Tamas et al., 2002), what means that the average relative rate of 1.38 requires a more precise interpretation. Assuming that the divergence of *B. aphidicola* from the free-living enteric cluster occurred at some moment between 200 and 300 My ago, its possible to define two different periods of evolution in the lineage of *B. aphidicola* in terms of genome rearrangements. In a first phase of evolution, during the adaptation to endosymbiosis since their divergence from their free-living ancestor, the relative rearrangement rate would be much higher than the estimated 1.38, around 2.76 if we consider the same period of time between the divergence from *E. coli* and the divergence of the three strains of *B. aphidicola*, and since then to nowadays. This initial phase of accelerated genome rearrangement rates would be followed, in a second phase of genome stability that corresponds to the divergence of the different *B. aphidicola* strains, by a rate of genome rearrangements close to zero as a consequence of the inability to produce and fix new rearrangements in their genomes. This conclusions can be applied also to *S. glossinidius*, the secondary endosymbiont of tsetse flies, that with a genome of 4.2 megabases and a GC content of 54% appears more close to free-living enterics like *E.coli* K12 than to primary endosymbionts like *B. aphidicola* or *W. glossinidia*, the primary endosymbiont of tsetse flies, with ancient evolutionary associations with their insect hosts. The average relative genome rearrangement rate of 2.00 observed in *S. glossinidius* together with the conclusions observed for *B. aphidicola* indicates that this genome is in initial stages of adaptation to endosymbiotic lifestyle, and confirms the general trend of accelerated rearrangement rates in bacterial endosymbionts in initial stages of the transition to a host-dependent lifestyle from a free-living ancestor.

3.3.5 Phylogenetic reconstruction based on BP and INV distances (gene order phylogenies)

A phylogenetic reconstruction based on genome rearrangement distances was carried out with two main objectives. First, to detect periods of faster or slower evolutionary rates according to the length of the branches of phylogenetic tree and second, to use these distances to determine the relationship between the endosymbiotic species and their position within the γ-proteobacterial phylogeny. The NJ and the FM algorithms were used with BP distances to reconstruct the phylogeny of the 31 γ-proteobacterial genomes (Figure 3.5).

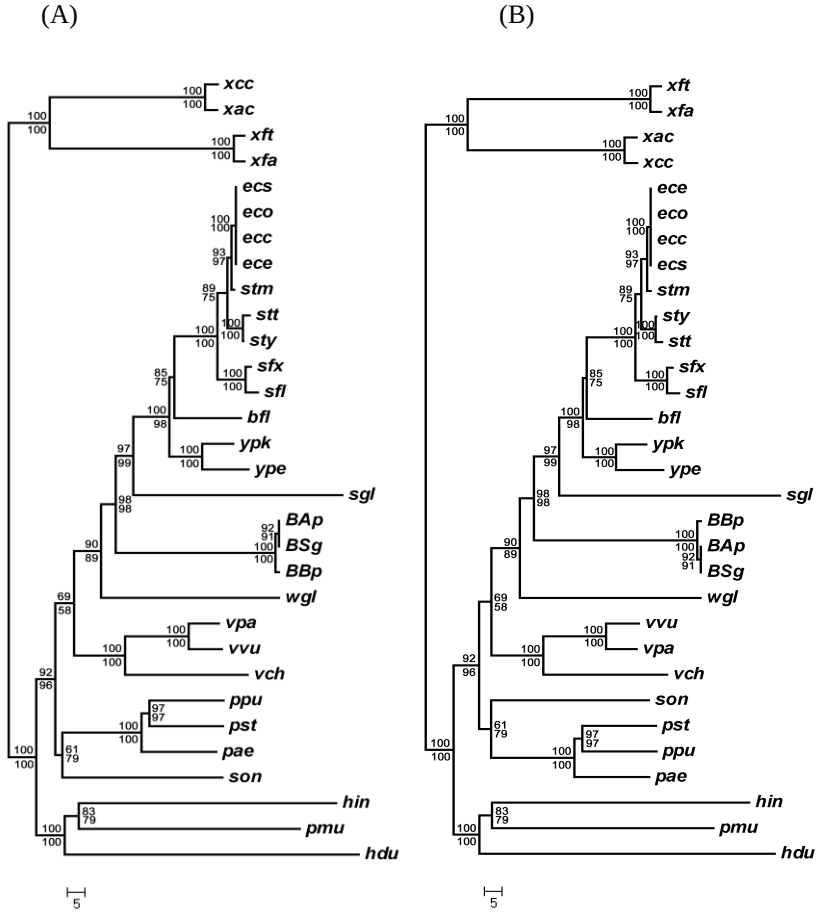


Figure 3.5: *Phylogenetic relationships between 31 γ-proteobacterial genomes inferred from breakpoint distances with Fitch-Margoliash (A) and Neighbor Joining (B) methods. Values at the nodes of the phylogeny represents the percentage of times the clade defined by that node are recovered in the 100 jackknife trees reconstructed with Fitch-Margoliash (upper values) and Neighbor Joining (lower values) methods. See species abbreviation in Table 3.1. The bar represents 5 inversions.*

Both methods inferred the same topology. In order to obtain supporting values for each node, BP distances were estimated after obtaining 100 random samples with genomes containing half the number of genes as the original set. The results of the Jacknife resampling method yields higher support of most of the nodes of the BP distance phylogenetic tree. The inferred topology obtained with BP distances was similar to that obtained based on amino acid sequences (Fig.3.2), but with several important differences.

The same approach was carried out with the matrix corresponding to INV distances (Fig 3.6), obtaining an even closer topology to the sequence-based one (Figure 3.2). In the INV distance phylogeny, the strains of *Shi. flexneri* move closer to strains of *E. coli*. However, in contrast to what is observed with BP distances, both methods of phylogenetic reconstruction gave approximately the same topology with the exception of *Sh. oneidensis* position, that in the phylogeny reconstructed with NEIGHBOR appears as outgroup of the *Pseudomonadales*, forming a monophyletic group in the same arrangement as in the phylogenetic trees reconstructed from BP distances. By contrast, in the phylogeny reconstructed with FITCH from INV distances, *Sh. oneidensis* appears in a similar location but without forming a monophyletic group with the *Pseudomonadales*, in a similar arrangement as observed in the phylogenetic tree inferred from sequence data (Figure 3.2). The results of the Jacknife resampling method assign a very small support to the position of *Sh. oneidensis* obtained by FITCH, and the majority rule consensus trees obtained with CONSENSE from 100 FITCH phylogenetic trees supports the monophy of *Sh. oneidensis* and *Pseudomonadales* in 81 and 69 of the Jacknife trees reconstructed by NEIGHBOR and FITCH respectively.

The most important differences between genome rearrangements and amino acid based phylogenies affected the position of the *Pasteurellaceae* cluster ad the position of the bacterial endosymbionts. In the gene order phylogenies, the *Haemophylus* spp. and *Pa. multocida* lineages acquired a basal position between the pseudomonads and the outgroup cluster. The second discordant result for the genome rearrangement phylogenies was the split of the monophyletic clade of the three lineages of primary endosymbiotic bacteria that is obtained in the sequence based phylogeny. The position of *Bl. floridanus* was, in addition, closer to *E. coli* than to *Y. pestis*. The *B. aphidicola* and *W. glossinidia* lineages maintained the position as outgroup of the cluster of enteric bacteria but as independent lineages instead as monophyletic group. The gene order phylogeny reconstructed the relations within the Vibrio and Pseudomonas genera well, although it was unable to generate a monophyletic group with the three Salmonella genomes. This feature is not surprising, due to the extremely low phylogenetic signal at gene order level observed between Salmonella and *E.coli* genomes, with a single inversion between the genomes of *E. coli* and *Salm. Thyphimurium LT2*.

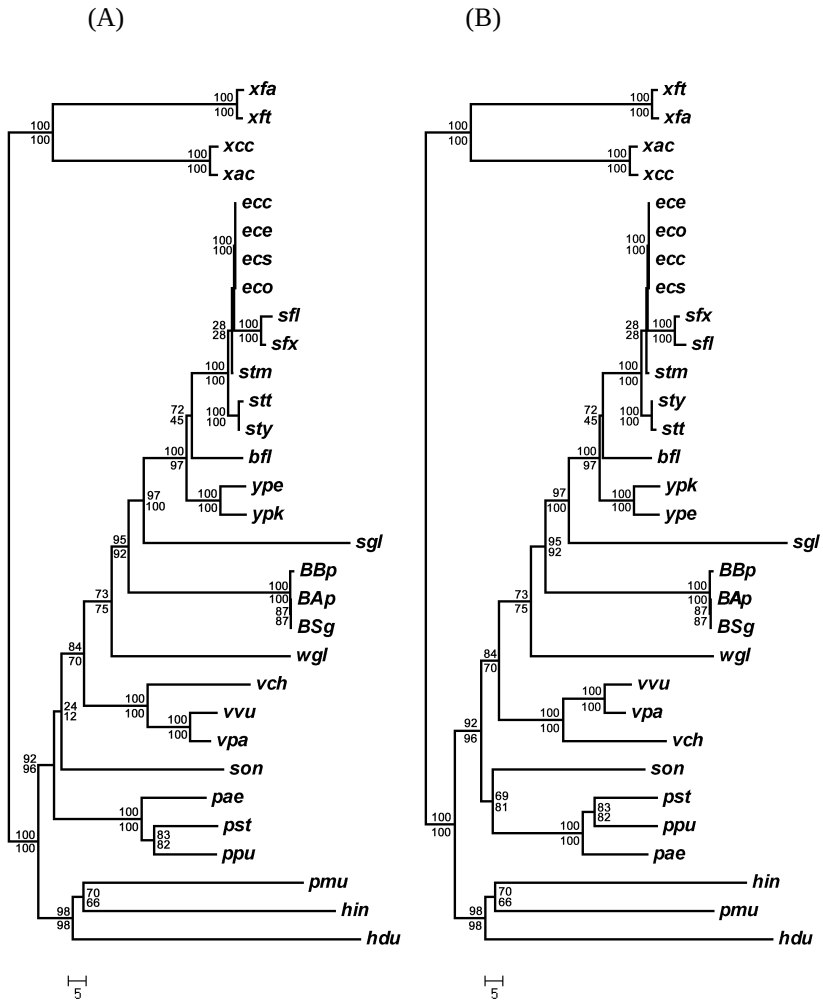


Figure 3.6: *Phylogenetic relationships between the 31 γ-proteobacterial genomes inferred from inversion distances with Fitch-Margoliash (A) and Neighbor Joining (B) methods. Values at the nodes of the phylogeny represents the percentage of times the clade defined by that node are recovered in the 100 jacknife trees reconstructed with Fitch-Margoliash (upper values) and Neighbor Joining (lower values) methods. See species abbreviation in Table3.1. The bar represents 5 inversions.*

The observation of the branch lengths in the phylogeny of the INV distances shows that the fixation of inversion events in the genomes is an irregular phenomenon during the evolution, combining periods of genome stability with others of acceleration of inversion rates. One example can be found in the high inversion rate observed in the *Shi. flexneri* lineage, with 9 and 11 inversions separating their genomes from the *E. coli* genome and even 4 inversions separating the genomes of *Shi. flexneri* 2a str. 301 and *Shi. flexneri* 2a str. 2457T. On the other hand, the genomes of *E. coli* and *Salm. Typhimurium* are separated by a single inversion, in spite of the estimated 100 My of divergence (Lawrence and Ochman, 1998). The two *Y. pestis* strains are another example of recent fast gene order evolution, with 14 inversions separating the two genomes in spite of the small ML distance between the two strains (0.0003) that are indicative of a very recent divergence. Another example is the large branches between the *Pasteurellaceae* species, which shows that they evolved and continue evolving at a fast genome rearrangement rate, and the large branch of *S. glossinidius* in gene order phylogenies in comparison to the amino acid sequence phylogeny, that can be explained by the accelerated genome rearrangement rates observed in the relative rate test.

The availability of three *B. aphidicola* strains has shown the existence of two very different periods of evolution, with an initial phase of fast rates of rearrangements corresponding to initial stages of adaptation to endosymbiotic lifestyle followed by a period of genome stasis during the divergence among *B. aphidicola* strains. This can be considered as a common feature among long-term established symbiotic associations in light of the strict collinearity observed between *bfl* and their close relative strain *Bl. pennsylvanicus* (data not shown), that points out to a similar evolution in terms of genome rearrangements. The acceleration in initial stages of endosymbiotic associations also explains the high rates of genome rearrangements observed in the secondary endosymbiont of tsetse flies *S. glossinidius*.

3.4 DISCUSSION

The availability of complete genome sequences of many γ-proteobacterial genomes makes this group very useful for the study of genome rearrangements through evolution. Several reasons make difficult a complete analysis of the changes in gene order, including the large divergence time of many γ-proteobacterial species, the different sizes and gene contents of their genomes, the presence of duplications and the high frequency of HGT events, and the existence of restrictions to gene-order change(Campo et al., 2004). To reduce the complexity of this study, in a similar way as is done with sequence data, a core of genes shared by all compared genomes that corresponds to slow-evolving genes that are putatively free of

xenologous genes generated by HGT has been selected for this genome rearrangement study. The restrictive criterion that the genes must be shared by the 31 γ-proteobacterial genomes analyzed mean that the selected 244 genes are close to the genome core (Jain et al., 1999; Charlebois and Doolittle, 2004).

To estimate the number of rearrangements between genomes, two genome rearrangement distances have been calculated. BP distances measure the number of breakpoints separating two genomes, which can be produced as a consequence of inversions and transpositions. Our BP distance measures the observed number of breakpoints between two genomes, a fact that can produce an underestimation of the real evolutionary BP distance because a BP may be reused by a new rearrangement event that would not be reflected in the observed BP distance, in a similar way as multiple substitutions in phylogenetic analysis based on sequence data. With sequence data, underestimation from observed distances can be corrected by the application of models of sequence evolution that allows the estimation of the true evolutionary distance between pairs of molecular sequences (Strimmer and von, 2003; Posada and Buckley, 2004). However, models of genome evolution have been much less developed, being limited to simple combination of the main rearrangement events like inversions and transpositions starting with the Nadeau and Taylor model where only inversions are considered with equal probability (Nadeau and Taylor, 1984), that has been extended subsequently in the GNT (Generalized Nadeau-Taylor) model to include both inversions and transpositions (Wang and Warnow, 2001). The second type of genome rearrangement distance is INV distance, that measures the minimal number of inversions needed to pass from one genome to another. Because INV distances do not consider transpositions, their use will only be appropriated when inversions are the most important genome rearrangement event through genome evolution. We consider that INV distance is appropriated to represent γ-proteobacteria genome evolution because the great correlation observed between inversions and BP distances in this study, that confirms previous results in other studies where inversions were detected as the main type of genome rearrangement observed in interspecific γ-proteobacterial genome comparisons (Hughes, 2000). As occur with BP distances, pairwise INV distances, as defined above, are underestimated and in general the true evolutionary distance is longer. Real INV distances may be got by analyzing a special mathematical problem called the Median Problem for signed genomes, in which, given three genomes, the objective is to produce a new genome that minimizes the sum of the distances from it to the other three, which allows also to reconstruct ancestral gene arrangements. Algorithms to try to solve these problems have been designed, although requires great computation capability and, for that reasons, these approaches to gene order evolution has been only applied with a limited number of genomes with a small number of genes, like mitochondrial or chloroplast genomes of eukaryotes (Bourque and Pevzner, 2002; Moret et al., 2002a; Tang and Moret, 2003).

The comparisons of rearrangement and sequence distances (Figure 3.3) and the observation of the branch lengths in the corresponding phylogenetic trees (Figures 3.5 and 3.6) shows that although rearrangement distances increase with evolutionary time, this increase occurs at heterogeneous rate with strong variations both between and throughout the evolution of γ-proteobacteria lineages. A great acceleration of the three *Pasteurellaceae* lineages has been detected. According to the results presented above, it has been estimated that, on average, the lineage of *Pasteurellaceae* are evolving at a relative rate of at least twice that of free-living enteric bacteria in terms of inversions. Natural competence observed between these species (Dubnau, 1999) is probably the responsible for this acceleration. The DNA uptake system in these bacteria requires the presences of a specific short sequence called the uptake signal sequence, for the binding and uptake of DNA fragments. The high number of copies of these uptake signal sequences in the genomes of *P. multocida* (927 copies) and *H. influenzae* (1471 copies) leads to the preferential uptake of DNA from close relatives (Bakkali et al., 2004). As consequence, these genomes will contain many genes product of HGT events that are difficult to identify due to the similar compositional characteristics of the donor and recipient species. Then, an important proportion of the *Pasteurellaceae* genes is possible to have been originated by HGT from close relatives, and should not have been included in this study, making this lineage unsuitable for performing gene order phylogenies.

Heterogeneity in phylogenetic position between sequence and gene order phylogenies has been also observed between endosymbiont lineages. The three *B. aphidicola* genomes included allows making a partition of the slightly high average relative rate of inversions (1.38) compared to free living enterobacteria into two extremely different periods of evolution in terms of gene order. These periods corresponds to the last 100 Myr. of almost complete genome stasis in terms of genome rearrangements and the initial period after the divergence from *E. coli* with a high rate of genome rearrangements (at least 2.76). These results confirm that the high number of rearrangements observed between *E. coli* and *B. aphidicola* occurred more than 100 Mya, most of them probably in a short period of time that corresponds to the adaptation of the ancestral genome to the endosymbiotic lifestyle due to the reduction of the restriction to the fixation of genome rearrangements. These conclusions are not restricted to the endosymbiotic lineage of *B. aphidicola*. For the bacterial endosymbionts of carpenter ants, the second genome published corresponding to *Bl. pensylvannicus* shows strict collinearity with the genome of *Bl. floridanus* included in this study, what indicates that the inferred average relative rate of inversions of 1.03 between Bl. floridanus and the free-living enterics can be also divided in the same two evolutionary periods as *B. aphidicola*, with the last 16-20 My corresponding to the divergence between *Bl. floridanus* and *Bl. pennsylvanicus* with complete genome stability and rearrangements accumulated during the initial phases of endosymbiotic adaptation from free-living ancestor at a

relative rate of 2.06 approximately given that the ancestral node of all ants known to have *Blochmannia* has been dated in 29.3-35.9 Mya (Degnan et al., 2004; Degnan et al., 2005). In the case of *W. glossinidia*, although only one genome sequence is available, it is possible that their behavior would be similar to that of *B. aphidicola* due to their similar average relative rate of inversions (1.35). Finally, in the recently established symbiotic association corresponding to the secondary endosymbiont of tsetse flies *S. glossinidius*, the average relative rate of 2.00 corresponds to an accelerated rate of genome rearrangement compared to free-living enterics that adjusts with the recent character of this endosymbiotic association in light of their genomic features and phylogenetic analysis (Chen et al., 1999; Toh et al., 2006). All this results indicates a common acceleration in the rates of genome rearrangements in bacterial endosymbionts during the initial transition to a host-dependent lifestyle from a free-living ancestor, that can be explained by the relaxation in the selection efficiency associated to this evolutionary transitions due to a relaxed selective pressure over large genome segments that becomes non-essential in the new host environment (Ochman and Moran, 2001; Wernegreen, 2004). This high rates of genome rearrangements can be favored by the massive proliferation of different types of mobile genetic elements during this initial stages of adaptation to host-dependent lifestyle, also consequence of the relaxation of the selective pressures, and that favors rearrangements by homologous recombination at repeated sequences across the genome (Moran and Plague, 2004). By contrast, at advanced stages of the evolutionary association a complete genome stability is observed probably due to the massive process of gene loss that includes genes involved in DNA uptake and recombination (Silva et al., 2003).

Genome-based phylogenetic approaches may be classified into three groups named gene content methods, sequence methods and gene order methods. Genome trees based on gene content were the first types of genome trees of complete genomes that were developed and published, showing reasonable correspondence to the know species tree (Fitz-Gibbon and House, 1999; Snel et al., 1999; Tekaia et al., 1999). Their reconstruction requires two steps. First, they must establish orthology relationships between genes in the compared genomes and second, they must convert the shared or unshared gene content into a tree structure by phylogenetic reconstruction methods that, in the absence of a more sophisticated method based on an explicit model of genome evolution, has been traditionally used distance based methods like Neighbor joining, although posterior developments based on more sophisticated phylogenetic methods like maximum likelihood or Dollo parsimony has been also proposed (Wolf et al., 2002; Gu and Zhang, 2004; Huson and Steel, 2004). These methods may be strongly affected when the number of genes in the compared genomes is markedly different because, in absolute terms, larger genomes of intermediate evolutionary distance shares more genes in common between them than large genomes do with closer relatives with smaller genomes (Snel et al., 1999). Several methods have been used to try to solve this problem, like to divide the

number of shared genes by the number of genes of the smaller genome, that represents the maximum number of genes the two genomes can share, but it is still usual to produce an incorrect position of *B. aphidicola* (Wolf et al., 2002; Dutilh et al., 2004). Genome-scale sequence-based methods help to resolve incongruent phylogenies inferred from single genes (Rokas et al., 2003).Genome-scale phylogenetic methods have been recently used to reconstruct the phylogenetic relationships among the three lineages of primary bacterial endosymbionts included in this study and the rest of the γ-proteobacterial species, and in all cases the monophyly of the bacterial endosymbionts, as well as their intermediate position in the γ-proteobacterial phylogenetic tree between the *E. coli-Y. pestis* and the *Pasteurellaceae* clades were inferred (Gil et al., 2003; Lerat et al., 2003; Canback et al., 2004; Brown and Volker, 2004). Finally, gene order data have been recently introduced in phylogenetics, and are considered to be especially suitable for resolving the phylogeny of closely related species, achieving a higher resolution at close evolutionary distances (Suyama and Bork, 2001; Snel et al., 2005). Problems associated to phylogenetic reconstructions based on gene order are the diversity in the shape and number of chromosomes, the variable number of genes, HGT events affecting different genomes, the rapid loss of gene order conservation in comparison to gene content and sequence evolution, and the heterogeneity in the rates of gene order evolution between genomes (Huynen and Bork, 1998; Blanchette et al., 1999; Moret et al., 2001; Bourque and Pevzner, 2002; Tang and Moret, 2003). Gene content and gene order methods for the reconstruction of genome phylogenies have been reported in SHOT, a web server for gene content and gene order phylogenies (Korbel et al., 2002). However, only one genome of *B. aphidicola* has been included in the database, and their position in gene order phylogeny using the default parameters was at the base of the γ-proteobacteria (Korbel et al., 2002).

For the resolution of the phylogeny of the 31 γ-proteobacterial species included in this study, a set of 244 orthologous genes presents in all genomes as either gene or pseudogene has been used. This number is slightly higher than the 205 orthologous gene set used by Lerat and collaborators to reconstruct the phylogeny of γ-proteobacteria (Lerat et al., 2003). The gene order reconstructions obtained by our analysis using two different genome rearrangement distances (breakpoints and inversions) and two different distance-based methods for phylogenetic reconstruction (Neighbor-joining and Fitch-Margoliash) rendered very similar results to other γ-proteobacterial phylogenies, with the basal position of the family Pasteurellaceae consequence of high rearrangement rates due to the inclusion in the analysis of a large set of non-easily detectable HGT genes. The most surprisingly result when compared the gene order and sequence-based phylogenies is the breakage of the monophyly of the three lineages of bacterial endosymbionts in gene order phylogenetic trees. Most of the sequence-based phylogenies tends to group these endosymbiotic species as sisterhood lineages, which is what is expected if the phenomenon of long branch attraction had taken place in fast evolving lineages,

grouping them together in a basal cluster near the outgroup based on their dissimilarity from the rest of species (Nei, 1996). This can be also applied to the phylogenetic tree obtained from the concatenated alignment of 10 proteins presented in Figure 3.2 because not only do the long branches of the 5 endosymbiotic genomes joint together, but also the two longest branches stick together (*Bl. floridanus* and *Wgl. glossinidia*). In addition, the branch leading to *B. aphidicola BBp*, which is the fastest evolving among the three *B. aphidicola* lineages, tends in some phylogenies to separate from the other *B. aphidicola* strains and to join *Bl. floridanus* and *W. glossinidia*. In order to avoid this artifact is necessary to select the slow evolving positions of genes and proteins in sequences alignments prior to phylogenetic reconstruction. In comparison, the lineage of *S. glossinidius*, despite having an extremely high rates of genome rearrangements that explains the long branches in gene order trees (Figures 3.5 and 3.6), appears clustered with enteric bacteria forming monophyletic group, closer to the lineage of *Yersinia*, in sequence-based phylogeny, although with low node support (Figure 3.2). The lineages of primary endosymbionts share characteristics such as a low GC content or a bias for the increase in AT-encoded amino acids (Moran, 1996; Clark et al., 1999), that is not shared by *S. glossinidius*, that with a GC content of 54 % have no sign of compositional bias. These biases in nucleotide and amino acid composition in primary endosymbionts can lead to a wrong clustering in phylogenetic reconstruction.

The phylogenies based on gene order distances presented in this study do not show a cluster of the three endosymbiotic species. In addition, the effect of long branch attraction is expected to be smaller because the acceleration of the branches leading to the three lineages of primary endosymbionts relative to free-living enteric bacteria are much smaller in the gene order than in sequence based phylogeny (Figures 3.2, 3.5, 3.6). However, several incorrect estimations may affect our gene order phylogenies. First, both rearrangement distances may underestimate the real number of genome rearrangements that are taking place during the evolution of γ-proteobacteria because both BP and INV distances are observed distances, and different evolutionary events such as multiple rearrangements occurring in a single rearrangement point or reciprocal inversions can lead to the underestimation of the real number of rearrangements. However, simulation studies indicated that this underestimation is very low (Bourque and Pevzner, 2002), specially for the intermediate values corresponding to the normalized distances among bacterial endosymbionts or between them and free-living enteric bacteria. Second, distance-based phylogenetic reconstruction methods like Neighbor joining or the least square method (Fitch-Margoliash) are not the most up-to-date methods in phylogenetic reconstruction, and have been outperformed by probabilistic methods like ML or Bayesian inference, although they are still valuable tools.

In light of the results presented in this study, the monophyly of bacterial endosymbionts is still an open question, and the retrieved monophyly observed in most genome-scale phylogenies could be an artifact of phylogenetic reconstruction due to the accelerated rates of sequence evolution associated to ancient bacterial endosymbionts. Even the use of ML methods for phylogenetic reconstruction from sequence data does not gives complete certainty that the correct phylogeny will be reconstructed. In fact, recent simulation studies shows that either distance or ML methods becomes inconsistent, with long branch attraction as one of the common forms of phylogenetic inconsistency(Susko et al., 2004), producing artifacts such as the joining of the fast evolving Microsporidian parasitic fungi lineage to the Archaea (Inagaki et al., 2004). It has been also demonstrated that even probabilistic phylogenetic reconstruction methods such as ML and Bayesian Markov chain Monte Carlo can become strongly biased and statistically inconsistent when the rates at which sequence sites evolve change non-identically over evolutionary time (Kolaczkowski and Thornton, 2004). Recently, γ-proteobacterial phylogenies based on two genes have shown that under non-homogeneous evolutionary rate models, *B. aphidicola* is separated from the other AT-rich endosymbiont bacterial species (Herbeck et al., 2005).

Finally, the estimation of genome rearrangement distances and the reconstruction of gene order phylogenies may be applied to the study of other bacterial groups. For example, an analysis of Firmicutes group, including the small genomes of Mycoplasma spp., may be done with around 138 genes shared for all the Firmicutes sequenced genomes. This number would increase to 159 removing the symbiont *Phytoplasma asteris*. In a recent study, INV distances and BP distances were also used to study phylogenetic relationships between 12 genomes of the photosynthetic marine bacterium *Prochlorococcus* spp., showing also high correlation between both gene order distances and similar topologies with sequence-based phylogenetic trees, helping to resolve the evolutionary relationship between low-light adapted ecotypes that were not clear with sequence based phylogenies (Luo et al., 2008).

3.5 CONCLUSSIONS

The analysis of genome rearrangements in γ-proteobacterial genomes has revealed that gene order change is a valid character for phylogenetic reconstruction, in light of the correlation observed with sequence divergence at amino acidic level. However, it is also clear that heterogeneity exist between different bacterial lineages, detecting lineages that are evolving with complete stability in their genome structures like the genomes of bacterial endosymbionts, whereas other lineages shows extremely high rates of genome rearrangements like the *Pasteurellaceae*. This heterogeneity is also present at sequence level, but the availability of large numbers

of models of sequence evolution makes that this variability could be taken into account into phylogenetic reconstruction in order to produce better phylogenetic trees (Posada and Buckley, 2004). In contrast, the models of genome evolution are at a very initial phase of development, and only very simple models are available that probably does not reflect the large complexity in the process of genome evolution in terms of rates of gene gain, loss, and the differential role of HGT in different bacterial lineages (Snel et al., 2002; Kunin and Ouzounis, 2003b). However, despite the lack of a model of genome evolution, phylogenetic reconstructions based on INV and BP distances are strongly similar to sequence-based phylogenies, with differences that can be explained by the observed heterogeneity in the rates of genome evolution, indicating their suitability for phylogenetic reconstruction.

The analysis of the average rates of genome rearrangement in bacterial endosymbionts has revealed a common trend shared by different endosymbiotic lineages in which initial transition from free-living to host dependent lifestyle appears associated to an acceleration in the rates of genome rearrangements followed by a period of genome stability that coincides with the divergence of the different endosymbiotic strains of each lineage. This is evident for long-term associated endosymbionts like *B. aphidicola* and *Blochmannia spp*., in light of the near absence of genome rearrangements between different strains of the same lineage and the average relative rates of genome rearrangement compared with free-living enterobacteria, and is confirmed by the situation of the secondary endosymbiont of tsetse flies *S. glossinidius*, in which a recent association with their host is accompanied by an acceleration in the rates of genome rearrangement approximately twice the rearrangement rate of free-living enteric bacteria. In addition, the comparison of gene-order distances and sequence based distances reveals that this increase is not accompanied by an increase in the rates of sequence evolution, a characteristic feature of most long-term bacterial endosymbionts like *B. aphidicola*, *Blochmannia spp.* or *Wgl. glossinidia*. This indicates that both processes are independent, and that genome rearrangements occurs prior in evolution, whereas acceleration of sequence evolution rates are characteristic of advanced stages of the genome reduction processes, probably due to the loss of genes involved in DNA recombination and repair (Silva et al., 2003).

4. Mobile genetic elements proliferation and gene inactivation impact over the genome structure and metabolic capabilities of Sodalis glossinidius, the secondary endosymbiont of tsetse flies

4.1 INTRODUCTION

The term symbiosis was coined by Anton de Bary in 1879 for the living together of two differently named organisms. Symbiosis can be defined as the association between two or more distinct organisms during at least one part of their lifecycles. Under this general definition is it possible to distinguish different types of symbiotic associations according to the ecological benefits or disadvantages that the relationship has on the symbiotic partners. If both partners of the symbiotic associations derive benefits from living together, the symbiotic association is defined as mutualistic. Parasitism is defined as a symbiotic association in which one of the partners benefits while the other is harmed. Finally, commensalism is an association in which one of the two members benefits from the association whereas the other is neither harmed nor obtains any advantage (Moran, 2006; Silva et al., 2007; Moya et al., 2008). In most cases, the symbiotic association is established between a multicellular eukaryote and a microorganism such as a bacterium or a unicellular fungus. These microbial symbionts can be also classified into two main groups depending of their degree of dependence on the symbiotic relationship for their survival. Facultative symbionts are those that retain the ability to return to a free-living condition, mainly due to their recent association with their corresponding eukaryotic hosts, whereas obligate endosymbionts have long evolutionary associations with their eukaryotic host and can only survive in association with them. In addition, symbiotic associations can be also classified based on the physical location of the microbial symbiont in their eukaryotic host. In this case, ectosymbiosis and ectosymbionts are applied to those symbionts that lives on the surface of their eukaryotic hosts, including those microorganisms that lives in the surface of digestive tract, whereas the terms endosymbiosis and endosymbionts refers to those symbionts that lives inside their corresponding eukaryotic hosts. This association may also be narrower when the microbial endosymbiont lives inside the host´s cells, in which case the term endocytobiosis is applied to these intracellular symbionts (Nardon and Nardon, 1998).

The origin of organelles of the eukaryotic cell such as mitochondria and plastids are considered as the oldest symbiotic event in evolution, dating back more than 1.5 billion years ago as postulated by Lynn Margulis in their generally accepted Serial Endosymbiosis Theory (Margulis, 1970; Margulis, 1981). According to this theory, eukaryotic organelles are the products of an ancient phagocytotic event of an ancestral α-proteobacteria, which has evolved into actual mitochondria, and an ancestral cyanobacteria, which has evolved into actual chloroplasts. The acquisition of metabolic traits such as oxygen respiration through mitochondria and photosynthesis through chloroplasts is considered as the key events in the evolutionary success of modern eukaryotes. These evolutionary events are not unique, and similar symbiotic events are known in all major vertebrate phyla with different impact on their evolution in terms of speciation and the ability to colonize

different ecological habitats. In *Protozoa*, different microbial endosymbionts inhabits in different locations in the host cells that are rich in metabolites, including the cytoplasm, the nuclei, and the periplasmic space (Gortz and Brigge, 1998), whereas *Porifera* (sponges) are frequently associated with large amounts of microorganisms that are thought to contribute to the chemical defense against the sponge´s predators, contributing from 40% to 60% of the sponges biomass (Friedrich et al., 1999). Another example is found in the genera *Cnidaria*, where symbiotic associations are established between corals and single celled algae, where the algae fix CO_2 and provide nutrients to the coral hosts, allowing them build thriving reef communities in the nutrient poor tropical waters (Steinert et al., 2000), whereas in the phylum *Mollusca*, the chemoautotrophic endosymbiont *Ruthia magnifica* from deep sea clam *Calyptogena magnifica* allows energy supply to the host by sulfur oxidation and CO_2 fixation in the ecological context of the hydrothermal vents (Newton et al., 2007).

Within the different examples of symbiotic associations between bacterial and eukaryotic lineages, symbiosis of insects and other arthropods with different bacterial lineages have been extensively studied in recent years associated to the massive process of genome reduction experienced by these bacterial endosymbionts during their transition from a free-living ancestor to a host-dependent lifestyle. The availability of complete genome sequences from several of these bacterial endosymbionts has revealed some of the most extreme examples of prokaryotic diversity in terms of genome organization and structure dynamics. These include the bacterial species with the smallest genomes known to date for any cellular organism, completely devoid of mobile genetic elements or bacteriophages, with the most biased nucleotide composition towards increases in their AT content associated with extremely high rates of sequence evolution, and with highly conserved genome organization in comparison with the highest genome plasticity observed between different strains of their closest free-living relatives (Shigenobu et al., 2000; Tamas et al., 2002; Gil et al., 2002; Akman et al., 2002; Gil et al., 2003; Degnan et al., 2005; Perez-Brocal et al., 2006; Nakabachi et al., 2006), but also bacterial genomes with the most extreme cases of mobile genetic elements proliferation that corresponds to bacterial endosymbionts in the initial phases of the association with their corresponding insect hosts (Wu et al., 2004; Foster et al., 2005; Plague et al., 2008; Gil et al., 2008; Klasson et al., 2009). In all cases, heritable symbionts of insects can be considered as obligatory symbiotic bacteria, in the sense that lacks of a replicative phase or dormant phase outside the insect hosts, although vary in the degree of dependence of the insect host on the symbiotic association for their successful development and reproduction that in most cases are correlated with the age of the symbiotic association. In this ecological context, three different types of heritable bacterial symbionts named primary and secondary endosymbionts (P- and S- type) and reproductive parasitic bacteria have been recognized. Primary endosymbionts or P-type endosymbionts correspond to obligatory mutualistic

bacteria that are required for the normal host development (Wernegreen, 2005; Moran et al., 2008; Moya et al., 2008). These primary endosymbionts are adapted to live intracellularly inside specialized host cells named bacteriocytes that form a specialized organ called bacteriome where the primary endosymbionts are restricted. Depending on the insect host lineage, the bacteriome can be derived from fat body cells, gut wall cells, or specialized insect cells that becomes developmentally determined in the embryos (Silva et al., 2007; Moran et al., 2008). Bacteriocyte-associated endosymbionts have been described in many different insect orders such as *Hemiptera* (aphids, psyllids or whiteflies), *Hymenoptera* (carpenter ants), *Diptera* (tsetse flies), *Dictyoptera* (termites or cockroaches) or *Coleoptera* (rice weevils) (Table 4.1). These bacteriocyte-associated endosymbionts have a strict vertical transmission from their mother hosts to their descendents, although there is different types of transmission routes depending on the insect lineage, like oocyte infection by bacterial endosymbionts that leaves the bacteriome and penetrates through the ovary in ants and parthenogenetic aphids, strict symbiont location in ovaries and female germ cells in some beetle species like those of the genus *Sitophilus*, or transmission through milk gland secretions used by females to feed developing larvae like in tsetse flies of the genera *Glossina* (Nardon and Nardon, 1998; Wernegreen, 2002; Baumann, 2005). The strict vertical transmission of bacterial endosymbiont from mothers to descendents combined with the absence of gene trasfer between endosymbiotic lineages originates a characteristic phenomenon of co-evolution between primary endosymbionts and their insect host lineages that has been used to date the age of the symbiotic association based on the fossil record of the insect hosts (Moran et al., 1993; Ochman et al., 1999; Chen et al., 1999; Funk et al., 2000). The nature of most of these ancient symbiotic associations are nutritional, in which the primary endosymbiont provides the insect hosts with different metabolites that are devoid in their corresponding diets, like the supply of essential amino acids by *Buchnera aphidicola* to their aphid hosts or the supply of different cofactors and vitamins by *Wigglesworthia glossinidia* and *Sitophilus oryzae* Primary Endosymbiont (SOPE) to the tsetse and grain weevil hosts, respectively (Zientz et al., 2001; Zientz et al., 2004; Dale and Moran, 2006).

In contrast to bacteriome-associated primary endosymbionts, facultative secondary symbionts or S-type symbionts have a wider tissue tropism, being able to occupy different tissues of the insect host, and are generally not essential for the host reproduction. These facultative symbionts resemble invasive bacterial pathogens in the sense that they are able to invade different cell types, including reproductive organs, are able to reside extracellularly in the body cavity (hemolymph) of their insect hosts (Fukatsu et al., 2000; Tsuchida et al., 2002; Haynes et al., 2003; Moran et al., 2005b) and, although are normally vertically transmitted, their distribution and phylogenetic patterns indicates that sporadic horizontal transmission between host individuals and host species must have occurred (Russell et al., 2003).

Table 4.1: Genomic data for mutualistic symbionts of animals

Organism	Host	Strain	Size (kb)	%GC	Protein coding genes	Pseudogenes	Reference
Buchnera aphidicola	*Acyrthosiphon pisum (aphid)*	APs	652	26,24	574	12	(Shigenobu et al 2000)
	Acyrthosiphon pisum (aphid)	5A	642,12	26	555	7	(Moran et al 2009)
	Acyrthosiphon pisum (aphid)	Tuc7	641,9	26	553	8	(Moran, McLaughlin, and Sorek 2009)
	Schizaphis graminum (aphid)	BSg	653	26,3	556	33	(Tamas et al 2002)
	Baizongia pistaciae (aphid)	BBp	618	25,3	507	9	(van Ham et al 2003)
	Cinara cedri (aphid)	BCc	422	20,2	362	3	(Perez-Brocal et al 2006)
Hamiltonella defensa (secondary)	*Acyrthosiphon pisum (aphid)*	5AT	2110,3	40	2094	187	(Degnan et al 2009)
Blochmannia floridanus	*Camponotus floridanus (carpenter ant)*		706	27,4	583	4	(Gil et al 2003)
Blochmannia pennsylvanicus	*Camponotus pennsylvanicus (carpenter ant)*	BPEN	792	29,6	610	4	(Degnan et al 2005)
Wigglesworthia glossinidia	*Glossina brevipalpis (tsetse fly)*		698	22,5	617	14	(Akman et al 2002)
Sodalis glossinidius (secondary)	*Glossina morsitans (tsetse fly)*		4171,	54,7	2431	972	(Toh et al 2006)
Sulcia muelleri	*Homalodisca coagulata (Sharpshooter)*	GWSS	245	22,4	227	_	(McCutcheon and Moran 2007)
Baumannia cicadellinicola	*Homalodisca coagulata (Sharpshooter)*		686	33,2	595	9	(Wu et al 2006)

Table 4.1 (Continuation)

Organism	Host	Strain	Size (kb)	%GC	Protein coding genes	Pseudogenes	Reference
Sulcia muelleri	*Diceroprocta semicincta (cicada)*	SMDSEM	277	22	242	4	(McCutcheon et al 2009a)
Hodgkinia cicadicola	*Diceroprocta semicincta (cicada)*		143,8	58	169	_	(McCutcheon et al 2009b)
Carsonella rudii	*Pachypsyllavenusta (psyllid)*		160	16,6	182	_	(Nakabachi et al 2006)
Blattabacterim str. Bge	*Blattella germanica (cockroach)*	Bge	636,8	27	586	1	(Lopez-Sanchez et al 2009)
Blattabacterium str. BPLAN	*Periplaneta americana (cockroach)*	BPLAN	637	28	576	4	(Sabree et al 2009)
Wolbachia str. wBm	*Brugia malayi (nematode)*	wBm	1080	34	805	98	(Foster et al 2005)
Wolbachia str. wMel	*Drosophila melanogaster*	wMel	1267,8	35	1195	74	(Wu et al 2004)
Wolbachia str. wPip	*Culex pipiens (mosquitoes)*	wPip	1482,4	34	1275	110	(Klasson et al 2008)
Ruthia magnifica	*Calyptogena magnifica (deep-sea clam)*		1200	34	976	_	(Newton et al 2007)
Vesicomyosocius okutanii	*Calyptogena okutanii (deep-sea clam)*		1000	31,6	937	_	(Kuwahara et al 2007)

In addition, these facultative endosymbionts have been experimentally introduced in uninfected insect hosts, where they are able to establish stable infections that are transmitted maternally to the descendents, revealing that their maintenance in insects hosts is achieved by symbiont-specific capabilities, like the presence of type III secretion systems specifically adapted for host cell invasion, rather than by host adaptations for maintaining symbiosis (Dale et al., 2001; Dale et al., 2002; Dale et al., 2005). In fact, in insects that harbor bacteriome-associated primary endosymbionts these facultative symbionts are able to invade bacteriocytes, where they co-reside with obligatory symbionts, even excluding them. However, in addition to harbor molecular mechanisms for the invasion of new hosts, cell entry, and the evasion of host immune responses, these facultative endosymbionts have to confer some beneficial selective feature to their insect hosts in order to ensure their spread and persistence across the evolution of the insect host lineage. In some cases, some of these secondary endosymbionts can rescue their insect host from heat damage, like *Serratia symbiotica* and *Hamiltonella defensa* in aphids (Russell and Moran, 2006), whereas *Hamiltonella defensa* has been also documented to protect aphid host from parasitoid wasp invasions (Oliver et al., 2003; Scarborough et al., 2005). In this concern, bacterial cells posses a large diversity of metabolic and biosynthetic capabilities that are devoid in insect hosts, so a wide variety of benefits for hosts are possible.

The third category of insect symbiosis corresponds to reproductive manipulators and parasites that does not provide any beneficial effect to their insect host but that ensures their transmission to the descendent lineages manipulating host reproductive biology by different strategies like cytoplasmic incompatibility between infected and uninfected strains, where males infected with the reproductive parasite sterilize uninfected females increasing the population ratio of infected matrilines, son killing, feminization of genetic males and parthenogenesis (Moran et al., 2008). Among these reproductive parasites, the best studied example corresponds to *Wolbachia pipientis*, an α-proteobacteria that is widely distributed in arthropods and in some other invertebrates. They show all above mentioned phenotypes (Dobson et al., 1999; Stouthamer et al., 1999; Brownlie and O'Neill, 2005), although reproductive parasitism has evolved repeatedly in different bacterial lineages, like *Rickettsia* and *Spiroplasma* species from the subdivision α-proteobacteria (Jiggins et al., 2000; Ogata et al., 2001; Fuxelius et al., 2007), or different bacteroidetes species like *Cardinium hertigii* (Hunter et al., 2003; Perlman et al., 2008).

These different types of bacterial endosymbionts are not exclusive, and is frequent to observe the presence of different types of bacterial symbionts that coexist within the same insect host, being also possible to found complex and stable symbiotic associations where different bacterial endosymbionts contributes to increase biological fitness of their insect hosts, like in the endosymbiotic consortium that is established between the aphid *Cinara cedri* and their endosymbionts

Buchnera aphidicola and *Serratia symbiotica*, where both bacterial endosymbionts contributes to different steps of the metabolic pathway of tryptophan biosynthesis (Gosalbes et al., 2008). Another nice example of metabolic complementation is found in the xylem-feeding sharpshooter *Homalodisca coagulata*, where their two primary endosymbionts *Baumannia cicadellinicola* and *Sulcia muelleri* have complementary biosynthetic capabilities, with *B. cicadellinicola* having genes needed for the biosynthesis of cofactors and vitamins but devoid of amino acid biosynthesis genes that are in turn presents in *S. muelleri* (Wu et al., 2006; McCutcheon and Moran, 2007). In addition, reproductive parasites like bacteria of the genera *Wolbachia* are frequently associated to insect lineages that harbor both primary and secondary endosymbionts (Heddi et al., 1999; Gomez-Valero et al., 2004b). One of these cases is the symbiotic association that is established between insect species of the genus *Glossina,* also known as tsetse flies (*Diptera: Glossinidae*), with almost three bacterial endosymbionts that covers the three different types of ecological associations between insects and bacteria described above. Tsetse flies comprises 31 different species and subspecies under the genus *Glossina* (order *Diptera*: family *Glossinidae*) that are largely classified into three groups (Morsitans, Palpalis, and Fusca groups) based on morphological differences in the structure of the genitalia (Gooding and Krafsur, 2005) (Figure 4.1).

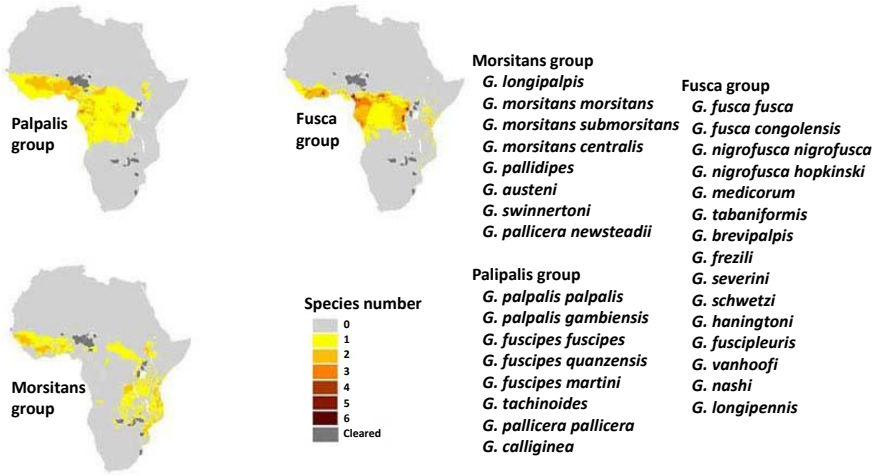


Figure 4.1: *Distribution of Glossina species of the three major subgenera (Morsitans, Palpalis and Fusca) across African continent.*

Tsetse flies are also the major vectors of several species of pathogenic trypanosomes in tropical Africa, which are the causative agents of Human African Trypanosomiases (HAT) or "sleeping sickness" in humans and Nagana in livestock.. Regarding their reproductive biology, tsetse females develop a single oocyte at the time of reproduction, and this oocyte is ovulated, fertilized and undergoes embryonic development in the mother´s uterus. The resulting larva hatches and is carried and nourished in the intrauterine environment for the complete duration of their development until an hour to parturition, where the larva burrows into the earth and pupates. This kind of viviparous cycle is termed pseudo-placental unilarviparity, and it has been also observed in three other families of haematophagous flies (blood feeding) evolutionary closer to tsetse flies (family *Glossinidae*), *Hippoboscidae*, *Nycterbiidae* and *Streblidae* (Attardo et al., 2008). To be able to carry out this viviparous reproductive physiology, tsetse have a uterus that is a modified vaginal canal covered with a highly treached muscle tissue, and that has the capacity to harbor a mature third instar larva equivalent in weight to the mother. The nutrition of the developing larvae is carried out by a modified accessory gland termed the milk gland that empties into the uterus. This milk gland is connected to the dorsal side of the uterus and expands throughout the abdominal cavity of the fly intertwining with fat body tissue. The lumen of the milk gland is surrounded by secretory and epithelial cells, and this epithelial cells secrete and maintain the chitinous lining of the lumen (Attardo et al., 2008).Tsetse flies harbor both primary and secondary endosymbionts primarily in their gut tissue, both belonging to the γ-subdivision of the proteobacteria (See Figure 4.2). The primary endosymbiont *Wigglesworthia glossinidia* (from now *W. glossinidia*) is an obligatory mutualistic bacteria that resides free in the cytoplasm of specialized bacteriocyte cells that forms the bacteriome structure located in anterior midgut of all tsetse flies lineages (Aksoy, 1995). Their genome was sequenced in 2002 (Akman et al., 2002), and reveals all the characteristic features of an ancient mutualistic symbiotic association similar to that of *Buchnera aphidicola* in aphids or *Blochmannia floridanus* and *Blochmannia pennsylvanicus* in carpenter ants, with a small genome size (697 kb) and an extremely biased nucleotide composition towards adenine and thymine (78% A+T). Their genome sequence contains only 621 protein coding sequences (CDSs) due to a massive process of genome reduction from a free-living enterobacteria ancestor, but retaining a large fraction of genes involved in the biosynthesis of cofactors (Akman et al., 2002). In this context, tsetse flies feed exclusively on vertebrate blood, which is vitamin deficient, and it was postulated that the reason for the origin of its symbiotic association with *W. glossinidia* was to cover this deficiency. This has been confirmated by experimental data with antibiotic-fed symbiotic-free flies that are able to survive with a nutritional supply of B-complex vitamins in their blood meal (Nogge, 1976; Nogge, 1981). These experiments also revealed the essential character of this symbiotic association for the tsetse fly, given that aposymbiotic flies shows significant retarded growth and decreased egg production, which results

in a loss of the reproductive capacity (Aksoy, 1995; Dale and Welburn, 2001). The ancestral character of this endosymbiotic association is revealed by the extensive concordance in phylogenetic reconstructions between *W. glossinidia* strains from different *Glossina* species and their corresponding hosts that allows to estimate the age of the symbiotic association some 50-100 million years ago, with an ancestral infection of the tsetse ancestor followed by evolutionary radiation together with the tsetse lineages without horizontal transfer of genetic material between *Glossina* species (Chen et al., 1999; Akman et al., 2002).

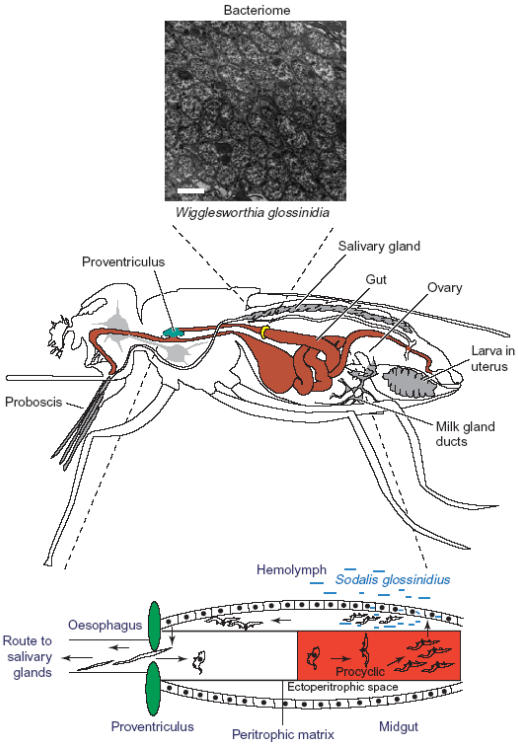


Figure 4.2: *Representation of female Glossina internal physiology, including one of the salivary glands, proventriculus organ in the foregut-midgut junction and the bacteriome organ in the gut are shown. The ovary and the larva in the uterous together with milk-gland ducts that provides them with nutrients are also depicted. W. glossinidia resides intracellularly in the bacteriome structure (micrograph showing W. glossinidia cells lying free within bacteriocyte cytoplasm), whereas S. glossinidius resides both intra- and extracellularly predominantly associated to hemolymph and gut cells. The diagram below represents the trypanosome lifecycle in the tsetse midgut and foregut. Adapted from (Rio et al., 2004)*

In addition to harboring *W. glossinidia*, most tsetse flies analyzed from the field also harbor a secondary endosymbiont from the family *Enterobacteriaceae* named *Sodalis glossinidius* (from now *S. glossinidius*) (Dale and Maudlin, 1999). This secondary endosymbiont exhibits a facultative association with tsetse flies, and have a more wider tissue tropism than *W. glossinidia*, being able to reside both intra- and extracellularly primarily in the midgut tissue but also in the muscles, fat bodies, hemolymph, milk glands, and salivary glands, showing high variation in their localization among different tsetse lineages (Cheng and Aksoy, 1999). Phylogenetic analysis with ribosomal sequences of *S. glossinidius* from different *Glossina* species reveals an evolutionary pattern corresponding with a very recent association with the tsetse host, with extremely low divergence at 16S rDNA sequence level that originate a complete absence of concordant evolution between *S. glossinidius* and their corresponding tsetse hosts, pointing out to an evolutionary scenario of independent acquisition of *S.glossinidius* symbiont by different tsetse lineages or alternatively with an initial acquisition by one of the *Glossina* species followed by horizontal transfer events among different tsetse species (Chen et al., 1999). However, despite the lack of phylogenetic congruence between *S. glossinidius* and tsetse phylogenies based on ribosomal sequence data, it has been detected more subtle differences in *S. glossinidius* populations from different *Glossina* species that constitutes genetically distinct populations in terms of their patterns of amplified fragment length polymorphism (AFLP) that are indicatives of a certain degree of population differentiation (Geiger et al., 2005; Geiger et al., 2007). In addition, there is extensive variation in relative density of *S. glossinidius* between different tsetse species and sexes, increasing the density of infections with the age of the fly although three orders of magnitude lower than *W. glossinidia* (Shaw and Moloo, 1991; Cheng and Aksoy, 1999; Rio et al., 2006). Another feature that postulates for a recent transition from a free living to a endosymbiotic lifestyle is that *S. glossinidius* is one of the few bacterial endosymbionts of insects that is able to be cultured in vitro, both in insect cell lines and on solid media (Welburn et al., 1987; Beard et al., 1993). In addition, is it possible to transform *S. glossinidius*-free tsetse flies of a given *Glossina* species with *S. glossinidius* originally isolated from a different *Glossina* specie without affecting the fitness of the recipient host and without differences in the densities of *S. glossinidius* populations between the transformed and control *Glossina* species (Weiss et al., 2006). The availability of pure cultures of *S. glossinidius* has enabled the development of genetic transformation systems that allow the introduction and expression of foreign genes in *S. glossinidius* and, in turn, in their corresponding *Glossina* species, an approach that has been called paratransgenesis and that has been used with the objective to express in the tsetse flies antitrypanocidal gene products through recombinant *S. glossinidius* to control trypanosome infections (Aksoy et al., 2003; Rio et al., 2004; Aksoy and Rio, 2005).. Finally, the specific elimination of *S. glossinidius* from tsetse flies by antibiotic treatment with streptozoin have very little effect upon tsetse fecundity, although F1 descendents of females treated with streptozoin lacks of *S.*

glossinidius and their longevity was significatively reduced (Dale and Welburn, 2001). This contrast with the more severe effects of *W. glossinidia* elimination, which produces a significative loss of fecundity, revealing the more integrated and essential role of the primary endosymbiont *W. glossinidia* in contrast to the more recently established secondary endosymbiont *S. glossinidius*.

In contrast to *W. glossinidia*, *S. glossinidius* depends on type III secretion systems to invade host cells and to proliferate inside them. This was initially detected by Dale and collaborators in 2001, where a *S. glossinidius* mutant in a type III secretion system invasion gene *invC* generated by random mutagenesis are unable to invade host cells in aposymbiotic tsetse flies (Dale et al., 2001). Phylogenetic analysis of *invC* and their neighboring genes reveals close proximity to genes from SPI-1 pathogenicity island from pathogens like *Salmonella enterica* and *Shigella flexneri* that constitutes a cluster of functionally related genes encoding the complete machinery for a type III secretion system together with genes encoding effector proteins that are secreted by this system, revealing an ancestral acquisition of these genes by the free-living ancestor of *S. glossinidius*, and the possible pathogenic character of *S. glossinidius* free-living ancestor that has been changed in the transition to a endosymbiotic lifestyle, using virulence determinants of enteric pathogens like *inv/spa* genes for the invasion of tsetse host cells (Dale et al., 2001). Posterior screening of *S. glossinidius* genome for *invA/spaP* genes and sequencing of the identified BAC clones allowed the complete characterization of two different symbiotic islands (SSR-1 and SSR-2) encoding type III secretion systems with different evolutionary origins and functional profile; The previously characterized SSR-2 island, closer to SPI-1 from *Salmonella enterica*, which lacks of genes encoding effector proteins and some genes of the needle substructure, and a novel SSR-1 island closer to *ysa* pathogenicity island from *Yersinia enterocolitica* that maintains a full complement of genes for the type III secretion system (Dale et al., 2005). Analysis of gene expression reveals differential expression patterns of genes from both symbiotic islands, with genes from SSR-1 being expressed when *S. glossinidius* contacts with the host cells and genes from SSR-2 being expressed once *S. glossinidius* have entered inside the tsetse cells (Dale et al., 2005). The presence of homologous *inv/spa* genes in the closer relative genome of the primary endosymbiont of grain weevils *Sitophilus zeamais* (SZPE) and their clustering in phylogenetic reconstruction reveals the ancestral acquisition of these symbiotic islands by the ancestor of both SZPE and *S. glossinidius* (Dale et al., 2002).

S. glossinidius biological role in the context of the symbiotic association with the tsetse fly is less clear than that of primary endosymbiont *W. glossinidia*. It has been shown that it plays a role in potentiating the susceptibility to trypanosome infection in tsetse flies, affecting the efficacy of the tsetse immune system (Welburn et al., 1989). Trypanosome infections in tsetse can only be established in the fly midgut if this parasites evades the action of specific trypanocidal lectins secreted by the tsetse

flies during feeding (Maudlin and Welburn, 1987; Mihok et al., 1992). *S. glossinidius* is thought to influence lectin activity through the production of N-Acetyl Glucosamine (GlcNAc), an amino sugar that inhibits tsetse lectins, which is accumulated during pupal development as result of *S. glossinidius* chitinolytic activity by chitinases and N-acetylglucamidases that degrades chitin from the tsetse midgut peritrophic membrane to produce GlcNAc and that has been demonstrated as essential enzymatic activities for the establishment of permanent *S. glossinidius* infections in tsetse flies in vivo (Welburn et al., 1993; Dale and Welburn, 2001). However, this is strictly a teneral phenomenon that affects teneral flies, that are tsetse flies in the developmental stage before their first blood meal, and this susceptibility is lost at subsequent blood meals, decreasing the rate of trypanosomal infection with the age of tsetse flies (Welburn and Maudlin, 1992).

S. glossinidius genomics started with the characterization of their four extracellular plasmids by Darby and collaborators in 2005, characterizing three different extrachromosomal plasmids (pSG1,2,4) and one bacteriophage-like extrachromosomal element (pSG3) (Darby et al., 2005).. *S. glossinidius* complete genome was sequenced in 2006 by Toh and collaborators, and all their genome features confirmed the recent symbiotic association of this bacterium with their tsetse host (Toh et al., 2006). *S. glossinidius* has a genome size of 4.3 megabases, strongly similar to closer enterobacterial free living relatives like different *Escherichia coli* strains, and far from the smallest bacterial genomes of obligatory mutualistic bacteria with long-term evolutionary associations with their corresponding insect hosts (0.1-1.5 megabases). In addition, *S. glossinidius* have no bias in their nucleotide composition, with an average GC content of 54%, also far from the highly biased genomes of obligatory mutualistic bacteria like tsetse primary endosymbiont *W. glossinidia* (22% GC). However, their coding density is very low, around 50%, consequence of a massive process of gene inactivation, with a total number of 972 pseudogenes described in the original genome paper but that remains unnanotated in the different genome-specific files available in public databases.

In contrast to the well studied genomes of obligatory mutualistic bacteria, for which several complete genomes are available from different insect species and even from different strains of the same species, *S. glossinidius* represents one of the few examples of a genome at the initial steps of the transition from a free living to a endosymbiotic lifestyle, and in the context of this thesis, one of the objectives is to characterize this evolutionary transition starting with the annotation and functional characterization of the complete set of pseudogenes presents in this genome that constitutes a hallmark of the functional character of *S. glossinidius* ancestor. In addition, another objective is to evaluate the impact of mobile genetic elements proliferation in the evolution of this genome, including the characterization of different types of insertion sequence elements (IS) and prophages and their impact over the functional capabilities of *S. glossinidius*, in light of the results obtained in

the previous chapter of these thesis, where a significative acceleration in the rates of genome rearrangement has been determined in the lineage of *S. glossinidius*. Finally, a complete reconstruction of *S. glossinidius* metabolism including the information from newly characterized pseudogenes and their comparison with the metabolic profile of the primary endosymbiont *W. glossinidia* has been carried out in order to define the functional role of both bacterial endosymbionts in the context of the tsetse symbiotic association and to identify possible cases of metabolic complementation like those observed in other insect-symbiont associations.

4.2 MATERIAL AND METHODS

4.2.1 Pseudogene annotation

The genome sequence of *S. glossinidius str. 'morsitans'* with their originally annotated 2431 protein coding sequences (CDSs) were recovered from GenBank (accession number NC_007712). In order to characterize pseudogene limits at nucleotide level, a two-step approach was followed. In a first step, intergenic regions of *S. glossinidius* genome comprised between non-overlapping genes with a minimum length of 50 base pairs (2215 intergenic regions) were used as query sequences against the proteome of *E.coli strain* K12 *substrain* MG1655, their closest free-living relative (accession number NC_000913) using BLASTX (Altschul et al., 1997) with a maximum e-value cutoff of 10^{-5}, extracting the beginning and the end of each BLASTX hit in *S. glossinidius* genome. Synteny with the genome of *E. coli* K12 and BLASTX alignment coverage were analyzed in order to fix the limits of the potential pseudogenes at nucleotide level. If a genome region presented homology with a single *E. coli* K12 protein, only those segments in which the BLASTX homology covered at least 25% of the E. coli K12 protein were retained. If a given genome region presented homology with several *E. coli* K12 proteins, synteny were analyzed, and pseudogene limits were fixed according with the beginning and the end of the BLASTX homology with the corresponding syntenic *E. coli* K12 CDS. If no synteny is observed between *S. glossinidius* and *E. coli* K12, the limits of the segment corresponding with the BLASTX hit with the lowest e-value were retained. This analysis produced a preliminary set of 845 potential pseudogenes homologous to *E. coli* K12 CDS. In a second step, a new extraction of intergenic regions of *S. glossinidius* was carried out this time considering the whole set of genes and pseudogenes characterized in the first step, and those with a minimum length of 50 base pairs (2378 intergenic regions) were used as query sequences against the proteomes of all completely sequenced bacterial genomes available in KEGG database (http://www.genome.jp/kegg/) at December of 2007 using BLASTX with a minimum e-value cutoff of 10^{-5}. The beginning and the end of the BLASTX hit with the lowest e-value were retained as nucleotidic

limits of the potential pseudogene. This analysis produced a second preliminary set of 879 potential pseudogenes, having a total number of 1724 potential pseudogenes delimited at nucleotide level by BLASTX. Finally, to characterize each pseudogene at amino acid level, Genewise program (Birney et al., 2004) was used to predict the open reading frames of each potential pseudogene based on the protein sequence of their corresponding best BLASTX hit. The information from the beginning and end of each pseudogene predicted by BLASTX and their corresponding open reading frames predicted by Genewise were integrated in the genome sequence of *S. glossinidius* with Artemis software release 10 (Carver et al., 2008), and open reading frames corresponding to the same pseudogene consequence of frameshift mutations were joined in a single CDS in order to reflect the complete amino acid sequence of the ancestral protein regardless of its lack of functionality. With this protocol, the 1724 predicted pseudogenes were characterized at both nucleotide and amino acid level.

4.2.2 Insertion sequence elements characterization

In the original annotation of *S. glossinidius*, 29 CDSs were annotated as putative transposases with no more specification about their corresponding IS type and family. In order to characterize the major types of mobile genetic elements presents in *S. glossinidius*, their genome sequence were self-compared by using the NUCMER program from the MUMMER package (Kurtz et al., 2004) with default conditions, allowing the identification of four main blocks of highly conserved repeated sequences with a minimum length of 500 base pairs that contained 28 of the 29 originally annotated transposase genes. In order to cover the complete IS length, repeated sequences of each type (named ISSgl1, ISSgl2, ISSgl3, and ISSgl4) plus 100 bp flanking regions were extracted (16, 6, 2, and 4 transposase genes respectively). The sequences were aligned with the program ClustalW (Larkin et al., 2007), and inverted repeats flanking the originally annotated transposase gene that defines IS limits were identified with the programs *Palindrome* and *Etandem* included in the EMBOSS package (Rice et al., 2000). Direct repeats generated after the insertion of each IS element were identified by visual inspection of the nucleotide sequence flanking complete IS elements. Consensus sequence for each alignment was extracted with the program *Consense* from the EMBOSS package and used as query sequence against the complete genome of *S. glossinidius* in BLASTN searches (Altschul et al., 1997) in order to identify more divergent or partial copies of each IS element not characterized in the initial approach. Finally, in order to characterize minor IS elements not detected in the previous approach, the sequence of *S. glossinidius* genome out of the previously characterized IS elements were used as query sequence against IS finder database (http://www-is.biotoul.fr/is.html) by using BLASTX with a minimum e-value cutoff of 10^{-5}. This allows the identification of a fifth family of IS elements (named ISSgl5), which was characterized by the same procedure described above. Genes disrupted by IS

element insertion were identified by BLASTX using the two sequences flanking complete and partial IS elements as queries against non-redundant protein database subdivision of GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) with an e-value cutoff of 10^{-5}. The results of these BLASTX searches were manually curated in order to detect coincident BLASTX profiles for flanking sequences of each IS element indicatives of gene inactivation by IS transposition.

For each type of IS element characterized, complete and partial copies were aligned with ClustalW. The observed p-distance between copies of the same type of IS element (number of differences / alignment length) were estimated with the pairwise deletion option of the MEGA4 software package (Tamura et al., 2007), and the segment of the multiple alignment corresponding to each transposase gene were extracted and translated. Elements with stop codons or frameshifts in the region corresponding to the transposase gene were considered as defective, whereas, in absence of any other information, non-synonymous substitutions that originate amino acid change as well as small insertions multiple of three nucleotides were considered to yield functional products.

4.2.3 Whole genome functional re-annotation

Manual re-annotation of all originally annotated *S. glossinidius* protein coding genes (2431 CDSs) and the 1724 potential pseudogenes characterized was carried out integrating different sources of information. Hierarchical functional classification scheme adapted from the Sanger institute, that is derived from the more general MultiFun classification scheme of gene products for bacterial genome annotation (Serres and Riley, 2000), were used to assign a "class" qualifier to each *S. glossidinius* CDS, both genes and pseudogenes, reflecting their cellular function (See Supplementary Table 4.1). In addition, the annotation of *E. coli* K12 genome were adopted as reference annotation given their close evolutionary relationship with *S. glossinidius* and their high accuracy of their annotation, that was recently updated by an international consortium specially focused on bringing up-to-date all functional information available for each gene product (Riley et al., 2006). Putative orthologous genes between *S. glossinidius* and *E. coli* K12 were identified by reciprocal best match on FASTA searches with amino acid sequences (Pearson, 1990) with cutoffs of 80% of the protein alignment length and 30% of amino acid identity, and annotation from the *E. coli* K12 was transferred to the putative *S. glossinidius* ortholog, including qualifiers "product", "EC_number", "class", "gene", "function", and "note". Reciprocal-best-match FASTA searches were also carried out with *S. glossinidius* genome against proteins of HAMAP database of manually annotated microbial proteomes (http://www.expasy.ch/sprot/hamap/) with the same cutoffs described above, transferring all annotation from HAMAP hits to their corresponding *S. glossinidius* CDSs, including "primary_name", "product" and "EC_number when they were available. The same approach was followed in

FASTA searches against an internal database of 24 bacterial genomes that are annotated by the same procedure, transferring as much annotation as possible from FASTA hits. Finally, individual BLASTP and FASTA searches were carried out with all *S. glossinidius* CDSs against the bacterial subdivision of UniProt database (http://www.expasy.ch/sprot/hamap/), and additional functional data were provided by individual searches of each *S. glossinidius* CDS against protein domain databases such as PFAM (http://pfam.sanger.ac.uk/) and Prosite (http://www.expasy.ch/prosite/). To help the functional assignment during the reannotation process, TMHMM (http://www.cbs.dtu.dk/services/TMHMM/)and SIGNALP (http://www.cbs.dtu.dk/services/SignalP/) programs were used to predict transmembrane domains and signal peptide sequences respectively based on amino acid sequences, and PSORT program (http://psort.ims.u-tokyo.ac.jp/) were use to predict cellular localization of gene products. The results of all these analyses were integrated together in the *S. glossinidius* genome sequence using Artemis software release 10, and manual re-annotation of all CDSs, both genes and pseudogenes, was carried out based on all this different sources of functional information at both functional and physical level (CDSs limits).

4.2.4 Metabolic reconstructions

In order to reconstruct the complete metabolic map of *S. glossinidius* to determine the impact of the pseudogenization process over the metabolic capabilities of the bacteria, KEGG pathway maps of *S. glossinidius* were generated with KEGG Automatic Annotation Server (Moriya et al., 2007) from amino acid sequences of genes and pseudogenes, and predicted EC numbers by KAAS were compared to those transferred during functional re-annotation process in order to ensure a correct assignment of enzymatic functions. In addition, BLAST2GO program (Gotz et al., 2008) were used to complement KAAS pathway reconstruction in order to predict multifunctional enzymes associated with more than one EC number that were not detected by the KAAS server. To avoid over assignment of EC numbers that would produce misleading inferences about metabolic capabilities of the bacteria, all additional EC numbers predicted by BLAST2GO that are not predicted by other methods (KAAS and functional re-annotation EC assignment) were individually analyzed by carrying out individual BLASTP searches of the corresponding aminoacidic sequence against non-redundant protein database subdivision of GenBank but restricting the search only to those sequences annotated with the additionally predicted EC numbers. BLASTP result hits with a minimum identity of 60% were considered significant. Extensive literature search and specialized metabolic databases like ECOCYC (Keseler et al., 2009) and METACYC (Caspi et al., 2009) were also used during the analysis of the reconstructed metabolic pathways.

4.2.5 Whole genome comparisons

TBLASTX genome comparisons were generated between *S. glossinidius* and *E. coli k12* genome to help the reannotation process and metabolic reconstructions. TBLASTX comparisons were also generated between the genome of *S. glossinidius* and the complete phage genomes subdivision of GenBank in order to characterize complete or partial prophage insertions. Whole genome comparisons were analyzed with Artemis Comparison Tool software version 6 (Carver et al., 2005).

4.3 RESULTS

4.3.1 Pseudogene number adjustment

It was previously described that the genome of *S. glossinidius* harboured 972 pseudogenes (Toh et al., 2006). However, their coordinates and the function of the active genes from which they derived were not annotated neeither in the manuscript nor in the GenBank file. We performed a complete pseudogene re-annotation identifying and characterizing 1501 pseudogenes that supposes an increase of 529 pseudogenes compared with the 972 pseudogenes described but not annotated in the original genome paper. First, we performed a BLASTX search over *S. glossinidius* intergenic regions, which allowed us to identify 1724 potential pseudogenes. Second, individual inspection of FASTA and BLASTP results for each CDS allowed us the detection of adjacent pseudogenes that were different frames of a same original functional gene. We merged these adjacent pseudogenes in single CDSs. Third, IS element characterization allowed the detection of ancestral genes that were inactivated by their insertions, with the two parts of the ancestral gene flanked by the annotated IS element. The two segments of each disrupted gene were merged in a single CDS. Fourth, during the re-annotation process we detected 142 situations in which a putative functional gene had been split in an ORF included in the primary annotation as a functional gene (Toh et al., 2006) and a pseudogene detected in our FASTA and BLASTP searches. These pseudogenes were eliminated from the final re-annotation and an additional "misc_feature" qualifier was added to the corresponding original gene specifying the proportion of the ancestral gene represented by the originally annotated gene and the re-annotated pseudogene. These 142 situations are described in Supplementary Table 4.2. Finally, potential pseudogenes with no significant homology on BLASTP and FASTA searches and any possible functional assignment by domain analysis (PFAM or Prosite analysis) or cellular location (PSORT or TMMM analysis) were eliminated from the final re-annotation. Despite the reduction over the initial estimations, the final set of 1501 pseudogenes overcame the 972 pseudogenes described in the original annotation.

4.3.2 Whole genome functional re-annotation

A systematic re-annotation of all CDSs was performed based on information derived from similarity searches (BLASTP, FASTA), protein motif searches (PFAM, PROSITE, TMHMM and SIGNALP), and protein localization prediction (PSORT). Supplementary Table 4.1 contains the outline of the functional classification scheme used in this re-annotation. This functional classification scheme was adopted from the Pathogen Sequencing Unit (PSU) of the Sanger Institute. Figure 4.3 summarizes graphically the results of the functional re-annotation process over genes and pseudogenes based on the functional classification scheme depicted in Supplementary Table 4.1.

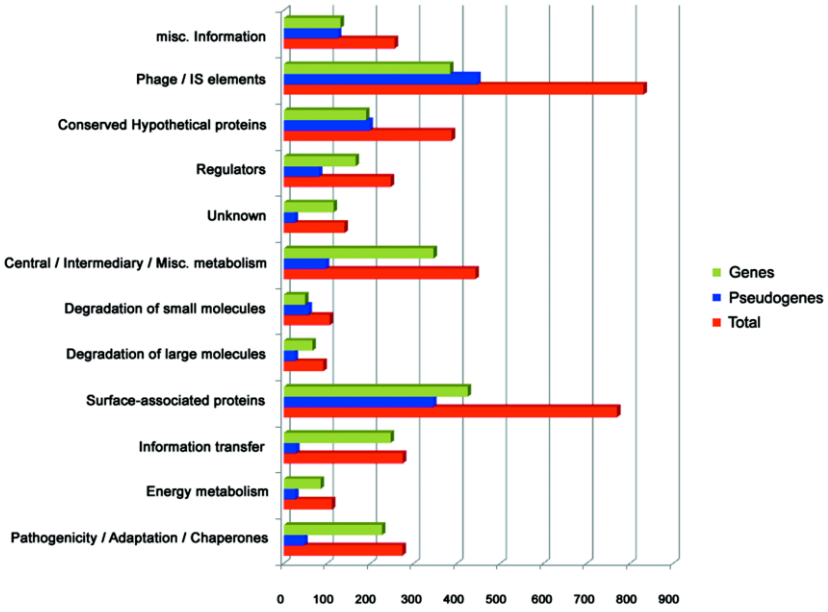


Figure 4.3: *Results of the re-annotation process. The functional categories correspond to the "colour" qualifier in Supplementary Table 4.1. Red bars show the number of CDSs (genes and pseudogenes) in each functional category. Green bars show the number of genes in each functional category. Blue bars show the number of pseudogenes in each functional category*

The highest gene proportion was assigned to the functional class of mobile genetic elements (831 CDSs) followed by surface proteins functional class (770 CDSs). These numbers correspond to 21.1% and 19.6% of the total number of CDSs respectively. Surface proteins functional class includes all integral membrane proteins (inner and outer membrane proteins), secreted proteins, and different types of membrane transporters, including ABC and MFS type transporters and components of the different types of PTS systems presents in *S. glossinidius* genome, but also includes all enzymatic activities involved in the biosynthesis of different components of cell envelope like peptidoglycan or bacterial lipopolysaccharide. Mobile genetic elements functional class includes all CDSs corresponding to transposases and different components of prophage elements, including integrases, reverse transcriptases, phage tail proteins, and hypothetical phage proteins. In addition, these two functional classes harbour the highest number of pseudogenes, with 447 and 345 pseudogenes respectively, being also the functional classes most affected by pseudogenization, with pseudogenes representing 53.8% and 44.8% of the CDSs assigned to mobile genetic elements and surface proteins functional class respectively. On the opposite side, the functional classes with the smallest number of CDSs correspond to degradation of large and small molecules, with 91 and 106 CDSs that represents 2.3% and 2.7% of the total number of CDSs of the genome respectively. It agrees with the previous observation that *S. glossinidius* has mainly retained biosynthetic rather than degradative pathways (Toh et al., 2006).

The original annotation of *S. glossinidius* genome includes 787 genes encoding hypothetical proteins. After the re-annotation process, 190 genes remained annotated as conserved hypothetical proteins, that corresponds to genes encoding proteins conserved in related bacterial genomes but that have an unknown function, and 115 genes as unknown proteins, representing orphan genes with the corresponding protein having no significant homology with any entry of public databases. This latter result contrasts with the 221 genes described in the original *S. glossinidius* genome paper with no homology with any entry from public databases (Toh et al., 2006). This is a consequence of the exponential growth in the number of complete bacterial genomes available in public databases due to the new generation of sequencing technologies (454, Solexa, and SOLID sequencing techniques), with more than 1,000 complete bacterial and archaeal genomes available since the publication of *S. glossinidius* genome paper in 2006 (http://genomesonline.org). In addition, the use of different sources of functional information out of BLASTP and FASTA searches allow making functional assignments for CDSs with no clear functional assignment based only on sequence similarity searches. A detailed survey of the functional reassignment of the 787 protein coding genes originally annotated as hypothetical proteins indicates that most functional reassignments were done to mobile DNA (245 genes) and surface protein (117 genes) functional classes (Figure 4.4).

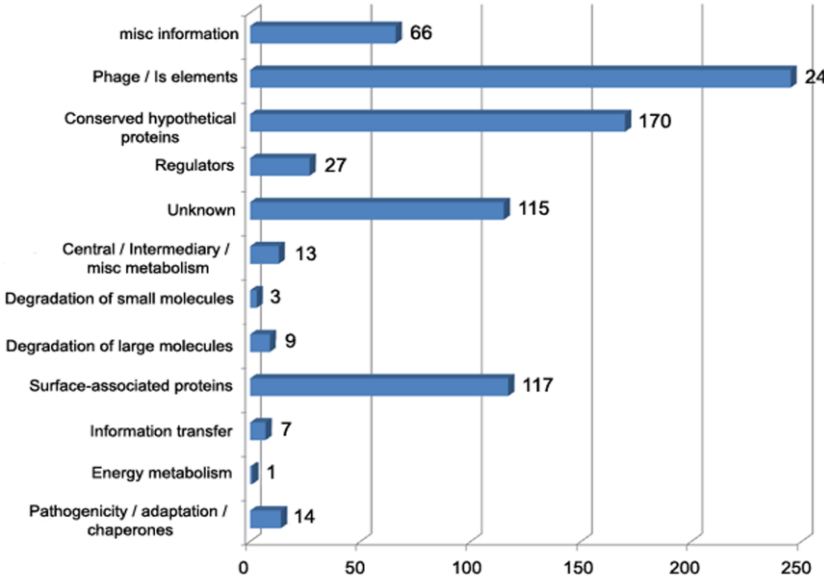


Figure 4.3: *Functional re-assignments of genes originally annotated as "hypothetical proteins"*

4.3.3 Prophage elements characterization

Due to the massive presence of phage-related CDSs and their significative clustering in different regions of the genome, a more detailed analysis of these CDSs was done in order to characterize possible prophage elements in the genome of *S. glossinidius*. Functional class 5.1.2 corresponding to Phage-related functions and prophages contains 698 genes and pseudogenes that represent 17.8% of the total number of CDSs of *S. glossinidius* genome, being also the functional class most affected by pseudogenization, with 353 pseudogenes. In order to confirm the phage-related role of these CDSs, additional BLASTP searches were carried out with these phage genes and pseudogenes against ACLAME database of mobile genetic elements (http://aclame.ulb.ac.be/), detecting only 41 genes and 62 pseudogenes that have no significant homology with any phage-related gene from ACLAME database (e-value cutoff of 10^{-6}), although have significant homology with hypothetical phage proteins from non-redundant databases on BLASTX searches. In order to characterize complete and partial prophage regions, TBLASTX analysis was carried out with *S. glossinidius* complete genome against all prophage genomes available in GenBank, and the results were analyzed with Artemis Comparison Tool (ACT) in order to identify regions of global synteny indicatives of a prophage element inserted in *S. glossinidius* genome. This analysis allowed the detection of two regions with homology with two complete prophage elements belonging to the Mu-family of double-stranded DNA bacteriophages at whole genome level. The first region, named SGLp1, is located between pseudogenes ps_SGL0195 and

ps_SGL0213, and shows strong collinearity with enterobacterial phage Mu (NC_000929.1) at whole genome level (Figure 4.4A). Pseudogenization has affected to 19 of the 28 CDSs included in the prophage region, including the genes *c* and *ner* responsible of the regulation of lytic and lysogenic development, *A* and *B* genes involved in phage integration and transposition, *I* gene encoding a protease and scaffolding protein, and the majority of genes encoding phage tail assembly proteins, indicating the inactivity of this prophage element (Morgan et al., 2002). In addition to prophage genes, this region includes 5 orphan genes that have no significant homology with any protein from ACLAME database or non-redundant database, that can be explained by the modularity associated with prophage evolution, with regions of significative homology between phage elements interdispersed with unrelated segments (Casjens, 2003; Canchaya et al., 2003). The second region, named SGLp2, includes 47 CDSs comprised between gene SG0816 and pseudogene ps_SGL0453, and shows strong colinearity at whole genome level with *Burkholderia phage BcepMu* (NC_005882) (Figure 4.4B), a Mu-like bacteriophage isolated from *Burkholderia cenocepacia strain J2315* that has closely related homologs in prophage elements from several bacterial genomes including *Salmonella typhi CT18* (NC_003198), *Salmonella typhi Ty2* (NC_004631), *Photorhabdus luminescens* (NC_005126), and *Chromobacterium violaceum* (NC_005085). All these prophage regions share a high degree of collinearity, with much less of the mosaicism detected in other bacteriophages, and share a common inversion of the entire left end region of their genomes in comparison with other Mu-like prophages, also present in SGLp2 (Summer et al., 2004) (Figure 4.4B). In contrast with SGLp1, where pseudogenization has affected the majority of CDS of the prophage region, in SGLp2 only 5 of their 47 CDSs are pseudogenes, retaining functional genes for most of the essential prophage activities, like *rve* gene that encodes a phage transposase (SG0823), the lysis gene cassette comprised of genes *slt* (SG0830), that encodes a lytic transglycolase with two transmembrane domains similar to SAR sequences (Signal Arrest and Release) that directs their integration in the bacterial inner membrane, SG08289 that encodes a holin-protein that activates the lytic transglycolase Slt allowing their access to the periplasmic space in order to start the lytic cycle (Xu et al., 2004), and SG0831, homologous to BcepMu nested genes 23 and 24, that encodes proteins involved in the degradation of the outer membrane and the links between the outer membrane and the cell wall during the lytic cycle. In addition, SGLp2 also contains functional genes for the phage capside formation and DNA packaging (SG0834-SG0838, SG0842) and genes for the biosynthesis of the phage contractile tail (SG0845, SG0846, SGL0849-SG0857). However, there is no homology between SGLp2 and BcepMu at the control region comprised between BcepMu11 and BcepMu21. This control region is also deleted in the prophage regions of *Salmonella typhi CT18* and *Salmonella typhi Ty2*, indicating that these prophage regions are uninducible cryptic prophages (Summer et al., 2004).

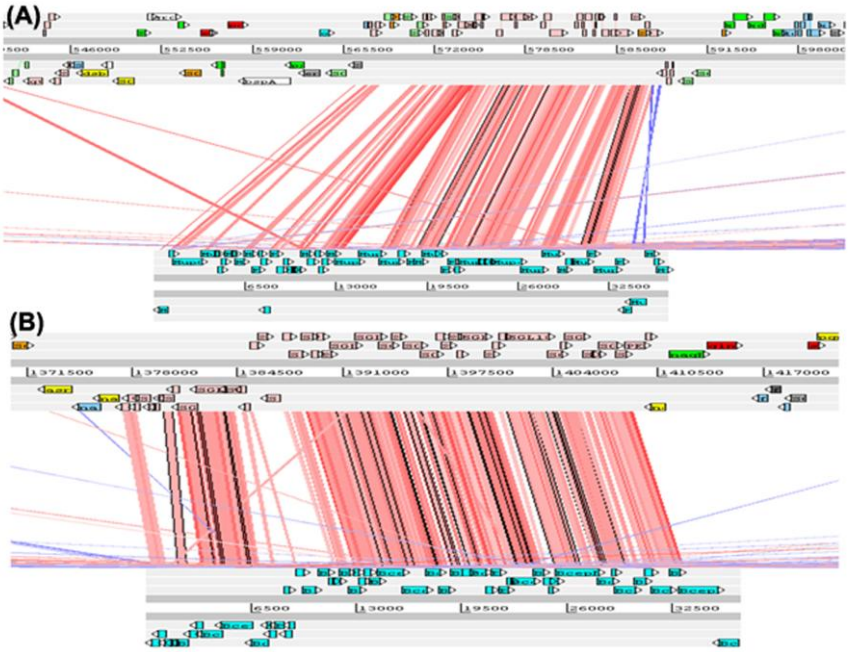


Figure 4.4: *Complete prophages characterized during the re-annotation process. Comparisons were generated with ACT based on TBLASTX comparisons of whole genome sequences: (A) S. glossinidius complete prophage region SGLp1 (top) vs. enterobacteria phage Mu (NC_000929.1) (bottom) (B) S. glossinidius complete prophage region SGLp2 (top) vs. Burkholderia phage BcepMu (NC_005882) (bottom).*

In addition to these two complete prophage elements, ACT comparison with complete prophage genomes allowed to identify 11 genomic regions that showed significant homology with different domains of completely sequenced phage genomes (Table 4.2). These include two different genome regions (SG0710-SG0725 and SG2350-ps_SGL14650c) that shows strong collinearity with structural and morphogenetic genes of *Enterobacteria phage SfV* (NC_003444) from *Shigella flexneri*, including genes involved in DNA packaging, capside morphogenesis, and tail structure and assembly but with no homology at specific phage genes for O-antigen modification, immunity and regulation responsible of lysogenic conversion between different serotypes of *S. flexneri* (Allison et al., 2002), together with two prophage domains that are homologous to the sequenced phage epsilon 15 (NC_004775.1), that is one of the precursors of *S. glossinidius* extrachromosomal bacteriophage-like element pSOG3 from *Glossina morsitans morsitans*(Clark et al., 2007).

Phage region	Start	Stop	Length (bp)	%GC	Total CDS	G	PS	Best homologous phages
1	168823	177292	8469	57,8	8	3	5	Klebsiella phage phiKO2
2	448798	463188	14390	57,3	13	6	7	Vibrio phage kappa
3	566249	587834	21585	57,9	28	9	19	Enterobacteria phage Mu
4	1192981	1211518	18537	49	21	16	5	Enterobacteria phage SfV
5	1377364	1410000	32636	56,2	47	42	5	Burkholderia phage BcepMu
6	1661326	1677343	16017	54,8	19	7	12	Yersinia phage L-413C
7	2011312	2027367	16055	56,6	16	8	8	Yersinia phage L-413C
8	2667966	2682203	14237	58,2	16	6	10	Yersinia phage L-413C
9	1930800	1963872	33072	51,3	28	12	16	Acyrthosiphon pisum bacteriophage APSE-1
10	1989376	2003483	14107	57,8	10	10	0	Enterobacteria phage epsilon15
11	1315020	1334014	18994	58,3	11	2	9	Pseudomonas phage D3112
12	2878251	2893450	15199	57,1	12	6	6	Enterobacteria phage epsilon15
13	4032451	4048913	16462	48,7	18	8	10	Enterobacteria phage SfV

Table 4.2: *S. glossinidius genome regions corresponding to domains of completely sequenced phage genomes. The coordinates are extracted from whole genome TBLASTX comparisons between S. glossinidius and completely sequenced phage genomes presents in GenBank at May 2008. Length, GC content, total numer of CDS, genes (G) and pseudogenes (PS), and the best homologous phage in TBLASTX searches are represented.*

4.3.4 Insertion Sequences characterization

Insertion Sequence (IS) element proliferation has been proposed as one of the main evolutionary events in the first stages of the genome reduction process associated with the lack of selective pressure over large genomic regions that become non-essential in the more stable host environment (Moran and Plague, 2004; Silva et al., 2007; Touchon and Rocha, 2007). In addition, IS proliferation has been extremely high in the genomes of rice weevils primary endosymbionts like *Sitophilus oryzae* (SOPE) and *Sitophilus zeamais* (SZPE), that appear as the closest relatives of *S. glossinidius* in phylogenetic reconstructions and that are also at initial stages of the genome reduction process, although their type of association with the rice weevil hosts are strictly mutualistic in contrast with the facultative association of *S. glossinidius* with the tsetse flies (Plague et al., 2008; Dougherty and Plague, 2008; Gil et al., 2008). In order to evaluate the effect of IS element proliferation in the genome of *S. glossinidius*, a complete characterization of the different types of IS elements presents in this genome has been carried out. Five different types of IS elements were identified in the genome of *S. glossinidius* representing 2.52% of the overall genome sequence, all of which showed clear homology with known γ-proteobacterial IS families, with their main structural characteristics described in Table 4.3 and Figure 4.5.

There are clear differences between IS types in terms of functionality and sequence divergence, which are indicative of different stages of IS degeneration. Among the five different types of IS elements characterized, only IS_Sgl1 and IS_Sgl2 can be considered as functional IS elements by the presence of a functional transposase gene clearly conserved at whole sequence level between IS copies of the same element and with homology with complete PFAM domains corresponding to transposases. The most abundant IS element is IS_Sgl1, that belongs to the IS5 family of IS elements (Mahillon and Chandler, 1998), and that contains a functional transposase gene encoding a protein of 307 amino acids that harbours a functional DDE domain common to most transposase proteins that are responsible for co-ordinate divalent metal ions, mainly Mg+2, that are needed during the course of the transposition reaction (Siguier et al., 2006a). Among the 16 genes originally annotated as transposase genes belonging to IS_Sgl1 type, only 6 genes contain a complete DDE domain, and can be considered with high confidence as potentially functional IS elements. The rest of originally annotated transposase genes of IS_Sgl1 have no homology with DDE transposase domain (complete or truncated), and can be considered as inactive IS elements.

		ISSgl	ISSgl	ISSgl	ISSgl	ISSgl
IS family		IS5	IS110	IS256	IS110	ISNCY
Structural features						
	Consensus sequence length(bp)	1052	1175	1247	1210	939
	GC content (%)	49.8	49.7	51	48.7	52.7
	Number of open reading frames	1	1	1	1	1
	Inverted repeats (bp)	17	11	34	10	n.d. [(c)]
	Direct repeats derived from transposition (bp)	9	n.d.[(d)]	9	n.d.[(d)]	n.d.[(d)]
Complete copies [(a)]		47	9	7	7	4
Partial copies [(b)]		16	9	12	2	9
IS with functional transposase genes		6	4	0	0	0
Pseudogenes generated by IS insertion		10	2	2	3	0
p-distance (number of nucleotide differences / total length						
	Minimum	0.0	0.0024	0.0062	0.0048	0.1097
	Maximum	0.066	0.0268	0.0756	0.0240	0.3935
	Mean	0.02	0.0162	0.0287	0.0131	0.2388
Total length (nt) [(e)]		52130	15328	16915	9375	11411

Table 4.3: *Structural and evolutionary characteristics of the 5 main groups of IS elements characterized in the genome of S. glossinidius.*

[a] *Complete copies are those that present inverted repeats at both ends of the transposase gene.*

[b] *Partial copies correspond to fragments of the putative IS element.*

[c] *No inverted repeats were detected in ISSgl5*

[d] *No direct repeats were detected in ISSgl2, ISSgl4, and ISSgl5 (n.d.).*

[e] *Sum of the length of all corresponding complete and partial IS elements.*

The other IS type that contains functional transposase genes is IS_Sgl2, that belongs to the IS110 family of IS elements, a family characterized by the absence of inverted repeats flanking transposase gene in most of their members, as well as by the absence of direct repeats flanking IS element after their transposition (Mahillon and Chandler, 1998). This is also observed in copies of IS_Sgl2, with no inverse nor direct repeats flanking the 9 complete IS_Sgl2 elements characterized, although inverted repeats of 11 base pairs are detected internally in the element, with right inverted repeat that are located inside the transposase gene (Figure 4.5)

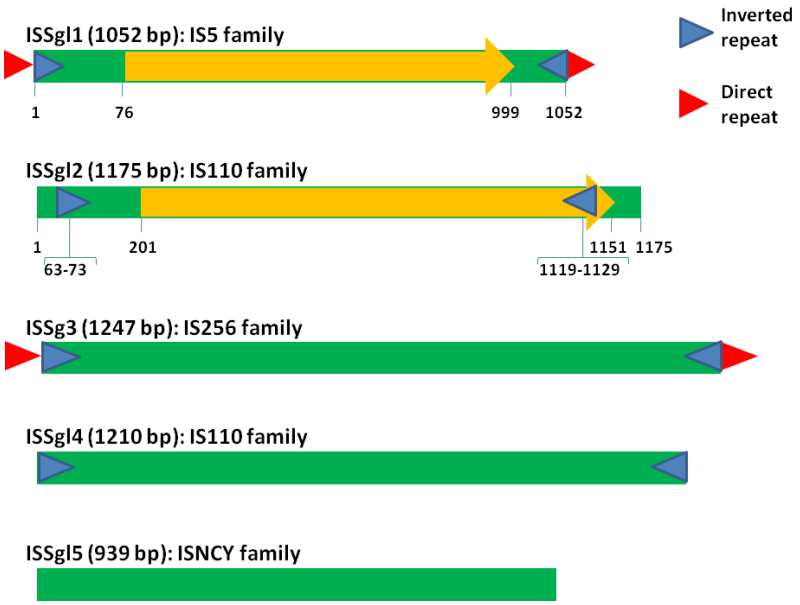


Figure 4.5: *Schematic representation of the 5 main types of IS elements of S. glossinidius. The position of the putative transposase gene is indicated in those IS types for which a functional transposase gene has been confirmed (ISSgl1 and ISSgl2). For ISSgl2 type, the relative position of internal inverted repeats is also indicated.*

The 6 transposase genes detected in IS_Sgl2 elements present the same profile on PFAM searches, with a complete transposase domain 20 (PF02371) associated to IS elements of IS110 family together with a partial transposase domain 9 (PF01548), being all potentially functional transposases except in one case where transposase gene SG0296, included in a partial IS_Sgl2 element, shows a premature stop codon compared with the other transposase genes probably reflecting an ongoing process of IS degradation. For the rest of IS types (IS_Sgl3, IS_Sgl4, IS_Sgl5), it was no possible the identification of a functional transposase gene common to all components of the IS type, despite each IS type included originally annotated

transposase genes. IS_Sgl3 elements contain only two originally annotated transposase genes (SG1677 and SG2070). BLASTP and FASTA searches of their encoded proteins against ISfinder database showed truncated homology with transposases from IS256 family with an average length of 400 amino acids approximately, with SG1677 (222 amino acids) being homologous with the first half of the transposases (position 1-200 in BLASTP searches) and SG2070 being homologous with the second half of the transposases (position 160-end in BLASTP searches), and with an overlap of 247 nucleotides between the end of SG1677 and the beginning of SG2070. In both cases, the absent part of the transposase was identified as a neighbour pseudogene upstream (in SG2070) and downstream (in SG1677) of the originally annotated gene, indicating that these originally annotated transposase genes (SG1677 and SG2070) are actually pseudogenes. IS_Sgl4 elements include 5 originally annotated transposase genes that share a common start codon at position 224 of the consensus IS_Sgl4, but that show variability at their stop codons. Three transposase genes (SG0243, SG1885, and SG2216) share the same stop codon at position 1187 of the consensus IS_Sgl4 sequence but SG0243 encodes a protein of 279 amino acids whereas SG1885 and SG2216 encodes proteins of 203 and 200 amino acids respectively. The other 2 transposase genes (SG1164 and SG1833) have premature stop codons at positions 690 and 820 of the consensus IS_Sgl4 sequence, rendering proteins of 154 and 152 amino acids, respectively. In all cases, the encoded transposases have truncated homology with transposase domains from PFAM database, suggesting that IS_Sgl4 elements are inactive. IS_Sgl5 type is the most divergent of all IS families characterized, with a rank of 0.110-0.394 nucleotide differences per site, which is an order of magnitude higher than differences observed in the other characterized IS families (see Table 4.3). IS_Sgl5 family includes 8 protein coding genes originally annotated as conserved hypothetical proteins that have homology with transposase 31 domain from PFAM database (PF04754), although neither flanking inverted repeats nor direct repeats were found. It was also not possible to get a well defined consensus sequence because there is no sequence similarity out of transposase genes. BLASTX searches against ISfinder database gave only significant homology with ISplu15 from *Photorhabdus luminescens* included in ISNCY family of poorly characterized IS elements, so members of IS_Sgl5 family probably reflect an ongoing process of IS inactivation.

In addition, 3 complete and 5 partial IS elements present in single copy were characterized during the re-annotation process. This raises the fraction of *S. glossinidius* genome represented by IS elements to 2.72%.

In order to evaluate the impact of IS proliferation over the functional capabilities of *S. glossinidius*, ancestral genes disrupted by IS element insertion were characterized by BLASTX searches of IS flanking sequences against non-redundant protein databases followed by manual inspection of the results in order to detect

coincident BLASTX profiles at both flanking regions of the same IS element (See Table 4.4). Of the 1501 characterized pseudogenes, only 18 were originated by the insertion of an IS element with no major effect in metabolic capabilities of the bacteria, affecting mainly to surface proteins (7 pseudogenes), mobile genetic elements (3 pseudogenes), degradation of small molecules (3 pseudogenes) and central and intermediary metabolism (2 pseudogenes) functional classes, indicating that IS transposition has not been a major force in the process of pseudogenization in *S. glossinidius*.

Pseudogene	Product	Functional class	Disrupted by
ps_SGL0005	Galactonate dehydratase	Degradation of small molecules	ISSgl1
ps_SGL0078	Putative outer membrane usher protein (LpfC)	Surface proteins	ISSgl2
ps_SGL0133	Aldehyde dehydrogenase A	Degradation of small molecules	ISSgl4
ps_SGL0325c	Putative oligogalacturonide transporter	Surface proteins	ISSgl1
ps_SGL0346	putative elongation factor G	Information transfer	(*1)
ps_SGL0567c	Conserved hypothetical protein	Conserved hypothetical protein	ISSgl1
ps_SGL0601	Glucose dehydrogenase	Metabolism	ISSgl1
ps_SGL0703c	Putative integral membrane protein	Surface proteins	ISSgl1
ps_SGL0820c	Cyclopropane-fatty-acyl-phospholipid synthase	Metabolism	ISSgl1
ps_SGL1041c	Putative uracil permease	Surface proteins	ISSgl3
ps_SGL1059c	Putative short chain dehydrogenase	Miscellaneous information	ISSgl4
ps_SGL1138	Putative propanediol utilization protein	Degradation of small molecules	ISSgl1
ps_SGL1164c	Putative membrane protein	Surface proteins	ISSgl4
ps_SGL1269	Transposase ISSgl3	Phage/IS elements	ISSgl3
ps_SGL1336	Putative citrate-proton symporter	Surface proteins	ISSgl2
SG2104 (*3)	Putative phage integrase	Phage/IS elements	ISSgl1
ps_SGL1420c	L-fucose permease	Surface proteins	ISSgl1
ps_SGL1461c	Putative phage protein	Phage/IS elements	(*2)

Table 4.4: *Pseudogenization events generated by IS insertion.*
*(*1): Partial IS element in single copy detected during re-annotation process*
*(*2): Complete IS element in single copy detected during re-annotation process*
*(*3): Situation of a longer ancestral gene, with ISSgl1 insertion in the corresponding adjacent pseudogene*

4.3.5 Metabolic reconstruction

A detailed reconstruction of the metabolic capabilities of *S. glossinidius* has been carried out based on the combined results of KEGG Automated Annotation Server (KAAS) and Blast2GO with the complete set of genes and pseudogenes (3932 CDSs) together with extensive literature searches and specialized metabolic databases like ECOCYC or METACYC that allowed the identification of several

features do not described in the original annotation. In the next sections I will describe the main results of this analysis, specially focused in the effect of gene inactivation over metabolic capabilities of *S. glossinidius* and the comparative analysis with the metabolic profile of the primary endosymbiont *W. glossinidia* at cofactors biosynthesis level in order to identify possible cases of metabolic complementation between both tsetse endosymbionts.

4.3.5.1 Carbohydrates metabolism

4.3.5.1.1 Glycolysis

Glycolysis is perhaps the most universal metabolic pathway, present with variations in nearly all living organisms, both aerobic or anaerobic. Through glycolysis, glucose is oxidized in 10 reactions to pyruvate with the concomitant energy release in the form of ATP and reducing power as NADH. *S. glossinidius* retains completely functional glycolytic pathway, although pseudogenization has affected to enzymatic steps catalyzed by different isozymes, retaining as functional gene the most processive isozyme and with pseudogenization affecting the less processive one.This is the case of phosphoglycerate mutase, where functional gene is retained for the major phosphoglycerate mutase I (*gpmA*) whereas pseudogenization has affected the minor phosphoglycerate mutase II (*gpmB*) (Figure 4.6). Pseudogenes corresponding to fructose 1,6 biphosphatase II (*glpX*) and 6-phosphofructokinase II (*pfkB*) are detected by means of sequence similarity searches and gene order comparison with *E. coli* K12. *S. glossinidius* also retains a functional gene *fbp* encoding fructose-1,6-biphosphatase I that catalyzes the dephosphorylation of fructose-1,6-biphosphate to fructose-6-phosphate, the key reaction of gluconeogenesis. Gluconeogenesis can start from TCA cycle intermediate malate in two enzymatic steps through functional malate oxidoreductase (*maeA*), that catalyzes the decarboxylation of malate to pyruvate, and functional phosphoenolpyruvate synthase (*pps*), that produces phosphoenolpyruvate from pyruvate, or from TCA cycle intermediate oxalacetate in one enzymatic step through functional phosphoenolpyruvate carboxykinase (*pck*), that catalyzes the phosphorylative decarboxylation of oxalacetate to pyruvate.

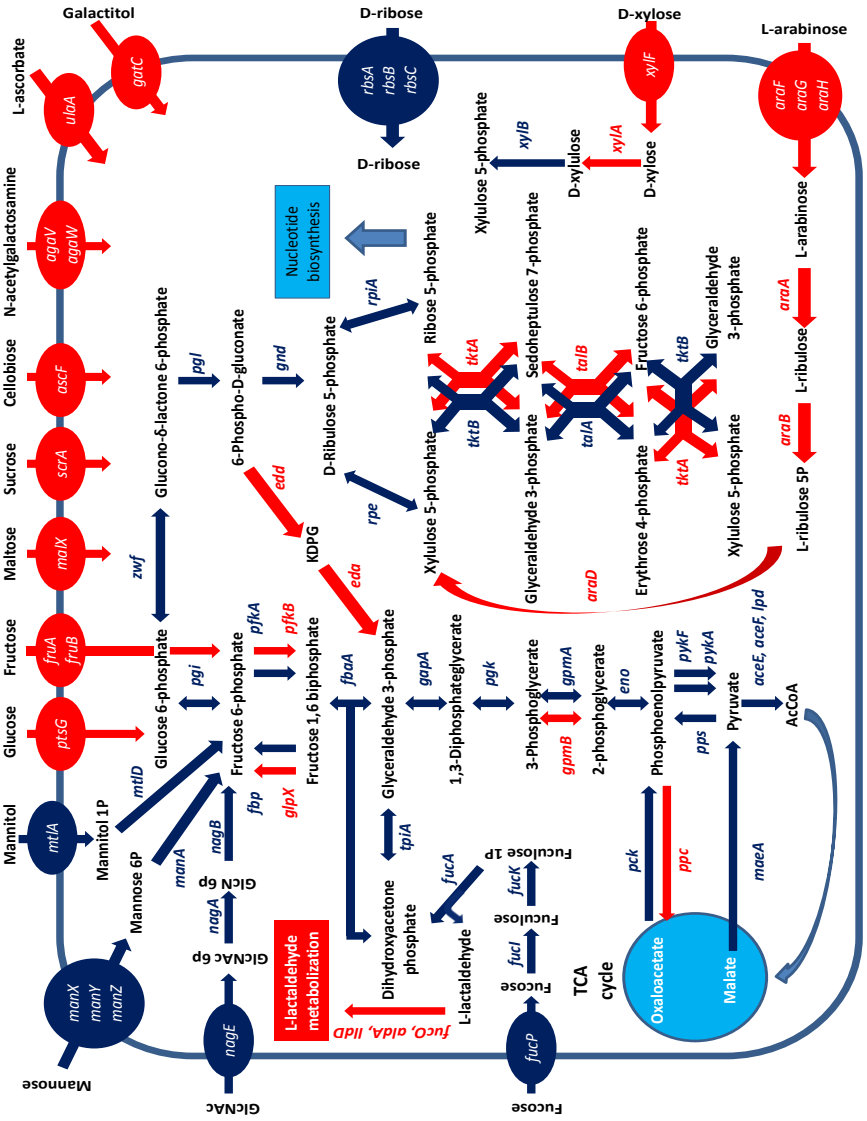


Figure 4.6: *Metabolic profile of S. glossinidius in terms of carbohydrate metabolization capabilities. In blue are represented functional reactions. In red are represented pseudogenized reactions.Double arrows represents different isozymes catalyzing the same reaction*

As carbon sources, *S. glossinidius* retains functional PTS systems for transport and phosphorylation of N-acetyl-D-glucosamine (GlcNAc), Mannitol and Mannose. Cytoplasmic GlcNAc-6-phosphate can be deacetylated to D-glucosamine-6-phosphate (GlcN-6P) by GlcNAc-6P deacetylase (*nagA*), and GlcN-6P can be converted to glycolytic intermediate Fructose-6-phosphate by GlcN-6P deaminase (*nagB*). GlcNAc polymer are known as chitin, and is considered as the second major biopolymer in nature after cellulose. Chitin is an essential component of the tsetse peritrophic matrix, a highly organized glycosamineglycan-rich layer that covers the tsetse midgut epithelia acting as a physical barrier that protects midgut epithelium from abrasive food particles, digestive enzymes, and infectious pathogens, being also a biochemical barrier that blocks and inactivates ingested toxins, and allowing the compartimentalization of the digestive process (Lehane et al., 1996; Welburn and Maudlin, 1999; Hegedus et al., 2009). The degradation of chitin polymers from peritrophic matrix by *S. glossindius* secreted chitinase has been postulated as a major factor increasing susceptibility to trypanosome infections in teneral flies due to the inhibitory effect of GlncNAc over trypanocidal midgut lectins produced by the tsetse fly (Kubi et al., 2006; Peacock et al., 2006; Roditi and Lehane, 2008). The genome of *S. glossinidius* retains two originally annotated genes encoding chitinases, although one of them encodes a truncated chitinase detecting the absent part of the protein as adjacent pseudogene. The products of chitin degradation are GlcNAc monomers and dimers known as chitobiose that can be used as carbon source by bacterial cells. In the case of *S. glossinidius*, GlcNAc can be transported to the cytoplasm and metabolized to glycolytic intermediates, although there is no signal in *S. glossindius* genome of the *chb* operon encoding chitobiose PTS system and specific cytoplasmic glycosyl hydrolase for chitobiose degradation to GlcNAc and GlcNAc-6-phosphate.

In addition, cytoplasmic mannitol-1-phosphate can be oxidized to glycolytic intermediate fructose-6-phosphate by functional mannitol-1-phosphate 5-dehydrogenase (*mltD*), whereas cytoplasmic manose-6-phosphate can be also converted to fructose-6-phosphate by functional mannose-6-phosphate isomerase (*manA*) (Figure 4.6). *S. glossinidius* retains also the complete pathway for transport and degradation of fucose, a methylpentose that can be used as carbon source by different bacteria including *E. coli* K12. Fucose can be transported by specific MFS transporter encoded by functional *fucP* gene, and once in the cytoplasm, can be converted to fuculose by functional L-fucose isomerase (*fucI*); L-fuculose can be phosphorylated to L-fuculose-1-phosphate by functional L-fuculokinase (*fucK*) that can be hydrolyzed to the glycolytic intermediate dihydroxyacetone phosphate and L-lactaldehyde by functional L-fuculose-phosphate aldolase (*fucA*). In addition, *S. glossinidius* retains a functional gene *fucR* encoding a transcriptional activator for the expression of the operon for fucose degradation and a functional *fucU* gene encoding a L-fucose mutarotase that catalyze the conversion between L- and D-fucose. However, the genes responsible of L-lactaldehide metabolization are all

pseudogenized in *S. glossinidius*, which has inactivated both *fucO* gene encoding L-1,2-propanediol oxidoreductase for the anaerobic conversion of L-lactaldehyde to L-1,2-propanediol, and the genes *aldA* and *lldD* that encode L-lactaldehyde dehydrogenase and L-lactate dehydrogenase for the conversion of L-lactaldehyde to pyruvate under aerobic conditions, with *aldA* being inactivated by IS_Sgl4 insertion.

Pseudogenization has also affected specific components of the PTS transporters of Glucose, Fructose, Maltose, Sucrose, Cellobioses, N-acetylgalactosamine, Galactitol, and N-ascorbate. In these cases, pseudogenization has affected to some or all of the components of each PTS system. However, mannose PTS systems have broad substrate specificity in enteric bacteria, and can also transport glucose, fructose, GlcN and GlcNAc, and 2-deoxyglucose that are transported and phosphorylated to their 6-phosphate derivatives (Curtis and Epstein, 1975; Stock et al., 1982; Postma et al., 1993; Kornberg, 2001).

S. glossinidius shows also pseudogenized the genes encoding 6-phosphogluconate dehydratase (*edd*) and 2-keto-3-deoxy-6-phosphogluconate aldolase (KDPG aldolase; *eda*), which are the key enzymatic activities of the Entner-Doudoroff pathway for the degradation of sugar acids like glucuronate, galacturonate, idonate, or fructuronate to glycolytic intermediates glyceraldehydes-3-phosphate and pyruvate, and that are essential for bacterial cells to growth under this hexuronides as carbon sources (Peekhaus and Conway, 1998). In a same way, genes of the *dgo* operon analogous to Entner-Doudoroff pathway but specific for D-galactonate transport and degradation to glyceraldehydes-3-phosphate and pyruvate are also pseudogenized in *S. glossinidius*.

4.3.5.1.2 Pentose phosphate pathway

This pathway allows the generation of the essential cofactor NADH, which acts as reductive agent in redox reactions of multiple biosynthetic pathways, and the essential monomer ribose-5-phosphate that constitutes the basic building block for nucleotide biosynthesis. This pathway can be divided in two main branches. The oxidative branch allows the direct oxidation of glucose-6-phosphate to ribulose-5-phosphate and NADPH, whereas the non-oxidative branch allows the regeneration of the glycolytic intermediates fructose-6-phosphate and glyceraldehyde-3-phosphate from ribulose-5-phosphate. *S. glossinidius* retains a completely functional pentose phosphate pathway (Figure 4.6), although as happens with glycolysis, pseudogenization has affected enzymatic steps catalyzed by different isozymes, like the pseudogenization of *tktA* and *talB* genes encoding transketolase A and translaldolase B respectively, retaining functional *tktB* and *talA* genes encoding transketolase B and transaldolase A respectively. *S. glossinidius* also shows a pseudogenized *xfp* gene that encodes phosphoketolase, an enzymatic activity associated with carbohydrate degradation in heterofermentative bacteria where

pentose phosphate pathway intermediate xylulose-5-phosphate is hydrolyzed to the glycolytic intermediate glyceraldehydes-3-phosphate and acetylphosphate. In addition, *S. glossinidius* has also inactivated different pathways for the transport and metabolization of pentose sugars to pentose phosphate pathway intermediate xylulose-5-phosphate, like pseudogenization of genes encoding L-arabinose high-affinity transport system (*araF, araG, araH*), L-arabinose isomerase (*araA*), and L-ribulokinase (*araB*), for the transport and metabolization of L-arabinose, and the pseudogenization of *xylF* gene of xylose ABC transport system together with pseudogenization of *xylA* gene encoding xylose isomerase for the conversion of xylose to xylulose, although retains a functional *xylB* gene encoding a xylulokinase for the phosphorylation of xylulose to xylulose-5-phosphate. For D-ribose transport and metabolization, *S. glossinidius* retains a functional high-affinity D-ribose ABC transport system encoded by the genes *rbsA, rbsB,* and *rbsC,* although there is no signal for the gene *rbsK,* which encodes a ribokinase for the phosphorylation of D-ribose to the pentose phosphate pathway intermediate D-ribose-5-phosphate.

4.3.5.1.3 Pyruvate metabolism and tricarboxylic-acid cycle

Tricarboxilic acid cycle (TCA cycle) is the key catabolic pathway of aerobic respiration, which generates energy in the form of ATP and reducing power as NADH from the complete oxidation of pyruvate to carbon dioxide and water. In addition, TCA cycle intermediates are also essential precursors for amino acid biosynthesis. *S. glossinidius* retains a completely functional pyruvate dehydrogenase complex (*aceE, aceA,* and *lpdA* genes) for the production of Acetyl Coenzyme A (AcCoA) from pyruvate and Coenzyme A in aerobic conditions, and a functional pyruvate-formate lyase (*pflB*) for the anaerobic conversion of pyruvate to AcCoA and formate, together with functional *pflA* gene encoding a *PflB*-activating enzyme and functional *adhE* gene encoding alcohol dehydrogenase-aldehyde dehydrogenase that acts also as PflB-deactivase. However, pseudogenization has affected to *poxB* gene encoding anaerobic pyruvate oxidase, which catalyzes the decarboxylation of pyruvate to acetate and carbon dioxide, essential for anaerobic cell growth with acetate as carbon source, although retains functional genes encoding phosphate acetyltransferase (*pta*), acetate kinase (*ackA*), and acetyl-CoA synthase (*acs*) for the interconversion between acetate and AcCoA (Figure 4.7).

A complete pathway of tricarboxylic acid cycle (TCA cycle) is also functional in *S. glossinidius*, although as happens with the glycolysis and pentose phosphate pathway, pseudogenization has affected enzymatic steps catalyzed by different isozymes (Figure 4.7). The most remarkable feature is the pseudogenization of the gene *mdh* encoding malate dehydrogenase that catalyzes the last step of the cycle, the reversible oxidation of malate to oxalacetate using NAD+ as electron acceptor; however, *S. glossinidius* retains a functional gene *mqo* that encodes a malate:quinone oxidoreductase that catalyzes the same reaction but irreversible in

the direction of malate to oxalacetate, a common feature shared with other bacterial endosymbionts like *W. glossinidia* and carpenter ants primary endosymbiont *Blochmannia spp* (Zientz et al., 2004).

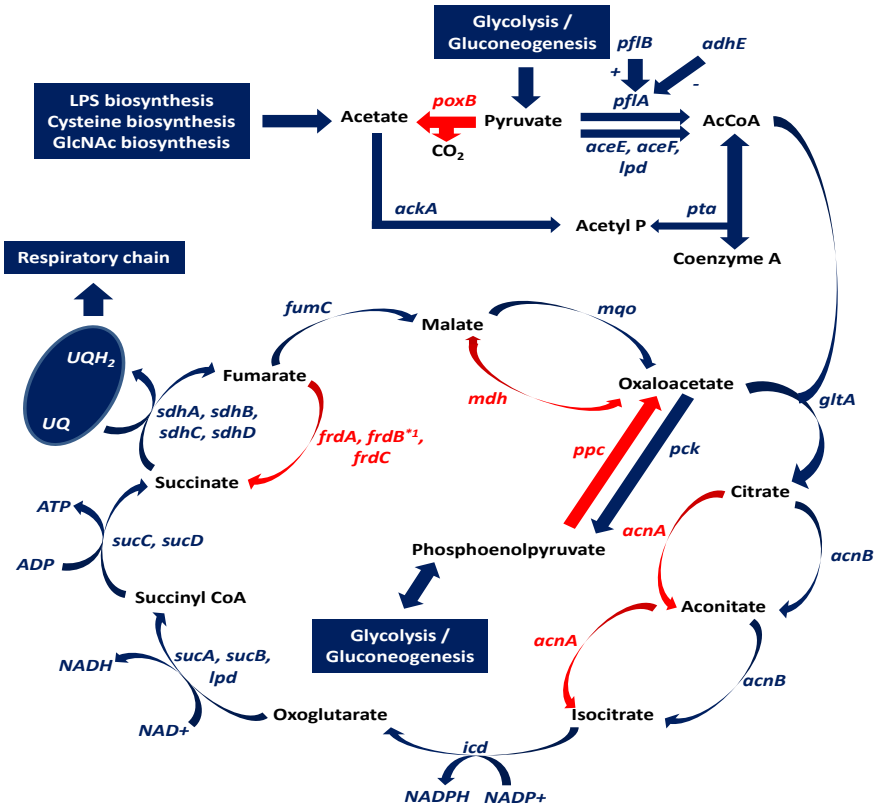


Figure 4.7:*Pyruvate metabolism and TCA cycle in S. glossinidius. In blue are represented functional reactions, whereas in red are represented pseudogenized reactions. Double arrows represent isozymes catalyzing the same reaction, whereas multiple genes asociated with a single arrow represents enzymatic complexes (p.ex. succinate dehydrogenase complex)*

However, pseudogenization have affected to the gene *ppc*, that encodes a key anaplerotic enzyme phosphoenolpyruvate carboxylase, responsible of oxalacetate replenishment to TCA cycle by carboxylation of glycolytic intermediate phosphoenolpyruvate, whereas there is no signal of genes encoding isocitrate dehydrogenase (*aceA*) and malate synthase (*aceB*), the enzymes of the glyoxylate bypass that catalyzes direct conversion of D-isocitrate to malate without NADH and ATP generation, a second anaplerotic pathway that is activated in *ppc* knockouts

(Peng and Shimizu, 2004). Glyoxylate bypass is also essential for bacterial survival on carbon sources like acetate or fatty acids. The absence of these enzymatic activities together with the absence of *aceK* gene that encodes the regulatory enzyme isocitrate dehydrogenase kinase/phosphatase that controls the flux between TCA cycle and glyoxylate bypass by posphorylation and dephosphorylation of isocitrate dehydrogenase indicates that *S. glossinidius* is not able to growth on acetate and fatty acids as carbon source, a conclusion supported by the pseudogenization of all *fad* genes involved in fatty acid degradation and the previously described pseudogenization of *poxB* gene encoding pyruvate oxidase. *S. glossinidius* also shows a pseudogenized *acnA* gene that encodes the isozyme aconitase A, retaining a functional *acnB* gene that encodes the major isozyme aconitase B. In addition, *S. glossinidius* retains a functional *fumC* gene that encodes the isozyme fumarase C, active under aerobic conditions, with no signal of the genes *fumA* and *fumB* that encodes the other two isozymes actives under microaerophilic and anaerobic conditions respectively. For the oxidation of succinate to fumarate, *S. glossinidius* retains a completely functional succinate dehydrogenase enzyme complex encoded by the genes *sdhA,sdhB, sdhC* and *sdhD* that catalyzes the oxidation of succinate to fumarate with the concomitant electron transference to the ubiquinone pool of the respiratory chain under aerobic conditions. However, *S.glossinidius* has pseudogenized *frdA* and *frdC* genes that encodes catalytic subunit and membrane-anchor subunit of fumarate reductase enzyme complex respectively, which catalyzes the reverse reaction, the reduction of fumarate to succinate under anaerobic conditions.

4.3.5.2 Energetic metabolism

S. glossinidius retains a completely functional electron transfer chain for oxidative phosphorylation, which couples the energy released by the oxidation of different nutrients and cellular compounds with ATP biosynthesis by ATP synthase enzyme complex. During this pathway, electrons are transferred from different electron donors (NADH, glucose, pyruvate, succinate, proline, etc.) through several redox reactions to different final acceptors like molecular oxygen (O_2) in aerobic respiration or other organic compounds like formate or nitrite in anaerobic respiration. These redox reactions generate free energy in the form of proton gradient by proton translocation across cytoplasmic membrane to the periplasmic space, which is coupled to ATP biosynthesis by ATP synthase enzyme complex (Figure 4.8).

S.glossinidius retains a completely functional *nuo* operon (*nuoA, B, C, D, E, F, G, H, I, J, K, L, M, N*) encoding primary NADH-dehydrogenase I, which constitutes the main proton pump of aerobic respiration coupled with the transfer of electrons from NADH to the intermediate ubiquinone pool. In addition, *S. glossinidius* retains a functional *ndh* gene encoding secondary NADH-dehydrogenase II, a monomerical

peripheral enzyme that catalyzes the same reaction but without proton translocation. In addition to NADH, other possible electron donors to the ubiquinone pool in aerobic respiration are succinate through their oxidation to fumarate in the TCA cycle by functional succinate dehydrogenase enzyme complex and the phospholipids precursor glyceraldehyde-3-phosphate by aerobic glyceraldehyde-3-phosphate dehydrogenase (*glpD*). *S. glossinidius* also retain a functional gene encoding proline oxidase (*putA* gene), a bifunctional enzyme that, when is associated to cytoplasmic membrane, catalyzes the two steps degradation of L-proline to L-glutamate, that includes the oxidation of L-proline to 1-pyrroline-5-carboxylate with the concomitant electron transfer to the ubiquinone pool (Wood, 1987; Zhang et al., 2004). However, pseudogenization have affected genes encoding D-glucose dehydrogenase (*gcd*), inactivated by ISSgl1 insertion, L- and D-Lactate dehydrogenases (*lldD, dld*), and pyruvate oxidase (*poxB*).

S. glossinidius also retains the complete pathway for ubiquinone biosynthesis from glycolytic intermediates glyceraldehyde-3-phosphate and pyruvate (Figure 4.8). These glycolytic intermediates are converted to isopentenyl pyrophosphate (IPP) and dimethylallyl diphosphate (DMAPP) through mevalonate pathway, in a set of reactions catalyzed by the genes *dxs, dxr, ispD, ispE, ispF, ispG* and *ispH*. DMAPP is the starting point for isoprenoid biosynthesis, which is synthesized by polymerization of IPP monomers generating polyisoprenoids of different length. This pathway starts with the biosynthesis of farnesyl diphosphate by the sequential addition of two molecules of IPP to DMAPP catalyzed by bifunctional geranyl/farnesyl diphosphate synthase encoded by the gene i*psA*. From farnesyl diphosphate, an octaprenyl diphosphate synthase encoded by the gene *ispB* catalyzes the production of the ubiquinone precursor octaprenyl diphosphate by sequential addition of IPP molecules to farnesyl diphosphate. Ubiquinone is synthesized by condensation of octaprenyl diphosphate with aromatic compounds derived from chorismate in a set of reactions catalyzed by the genes *ubi* (*ubiC, A, X, D, B, G, H, E, F,* and *G*) (Figure 4.8).

As final electron acceptor of the respiratory chain, *S. glossinidius* retains completely functional *cyd* operon (*cydA, B*) and *cyo* operon (*cyoA, B, C*) encoding cytochrome *bd* and cytochrome *bo* terminal oxidases enzyme complexes respectively, which catalyzes the electron transfer from the ubiquinone pool to oxygen in aerobic respiration, with the concomitant reduction of molecular oxygen to water. Both enzyme complexes catalyze the same reaction, differing in their activity as proton pump, with cytochrome *bd* complex that is not able to carry out proton pumping coupled to redox reaction, being active under oxygen-limited conditions, whereas cytochrome *bo* complex is able to act as proton pump coupled to electron transfer from ubiquinone to oxygen.

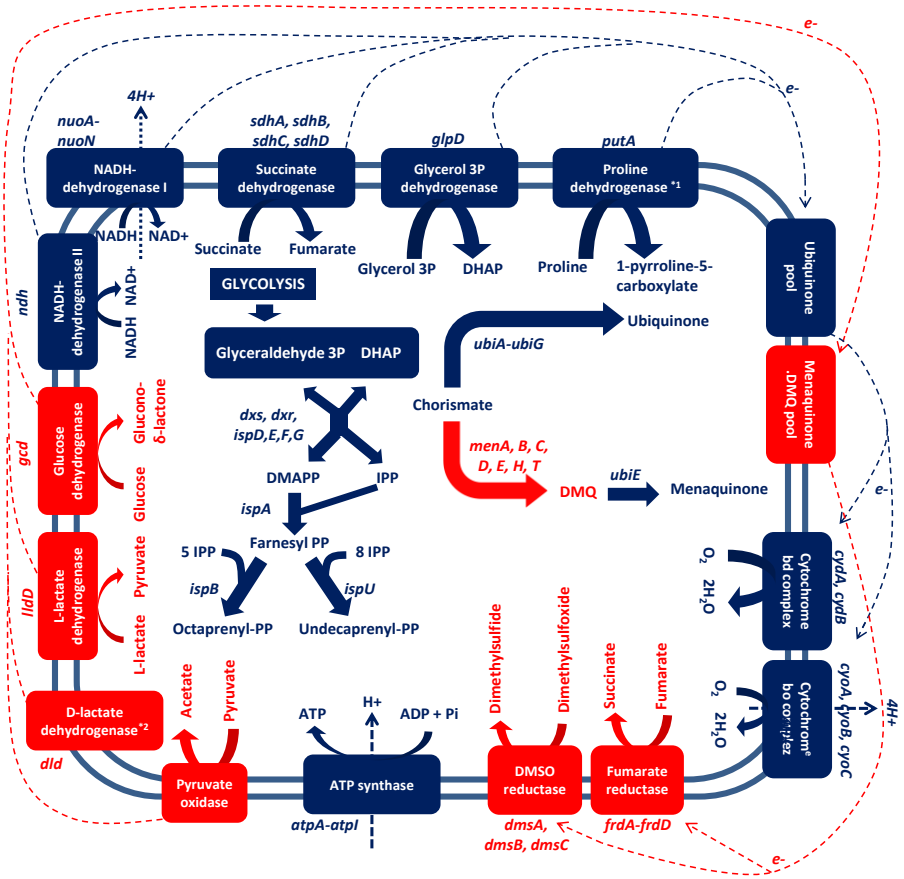


Figure 4.8: *Energetic metabolism profile of S. glossinidius. In blue are represented functional reactions, whereas in red are represented pseudogenized reactions.*
*[*1] putA encodes a bifunctional enzyme with proline dehydrogenase/1-pyrroline-5-carboxylate dehydrogenase responsible of the degradation of proline to glutamate. Proline dehydrogenase activity couples oxidation of proline to pyrroline-5-transfering electrons to the respiratory chain (Wood, 1987).*
*[*2] D-lactate dehydrogenase catalyzes the same reaction as L-lactate dehydrogenase but with D-lactate as substrate*

S. glossinidius also retains a complete *atp* operon (*atpA, B, C, D, E, F, G, I*) encoding the ATP synthase complex that carries out ATP biosynthesis using energy from the proton gradient generated by the respiratory chain. In addition, functional *ppk* and *ppi* genes encoding polyphosphate kinase and pyrophosphatase generates orthophosphates necessary for ATP biosynthesis from polyphosphate molecules (Figure 4.8).

In contrast, pseudogenization has affected most enzymatic complexes for anaerobic respiration, like the inactivation of the *dms* operon (*dmsA, B, C*) encoding dimethylsulfoxide reductase enzyme complex for the utilization of dimethylsulfoxide as terminal electron acceptor in anaerobic respiration, or the previously described inactivation of the genes *frdA, frdB,* and *frdC* encoding the enzymatic complex fumarate reductase, and with no signal for genes encoding nitrite or nitrate reductases, trimethylamine N-oxide reductase, or thiosulfate and tetrathiosulfate reductases. In accordance with the absence of functional enzyme complexes for anaerobic respiration, pseudogenization has also affected to *men* operon (*menE, C, B, ybfB, mend, menF*) that encodes enzymes for the biosynthesis of menaquinone and dimethylmenaquinone, the intermediate electron carriers of anaerobic respiration.

4.3.5.3 Amino acid metabolism

S. glossinidius retains functional biosynthetic pathways for all amino acids except for L-arginine biosynthesis, which is inactivated in four enzymatic steps. In contrast, pathway of L-alanine biosynthesis is completely functional, a feature that contradicts the conclusions of Toh and collaborators in *S. glossinidius* genome paper where postulated that the unique amino acid biosynthetic pathway inactivated in *S. glossinidius* was the pathway for L-alanine biosynthesis (Toh et al., 2006). In the next sections I will describe briefly the most relevant features of the different amino acid biosynthetic pathways in *S. glossinidius*.

4.3.5.3.1 Glutamate and glutamine metabolism

S. glossinidius retains *glnA* gene encoding glutamine synthase, the unique enzyme responsible of L-glutamine biosynthesis from L-glutamate, ammonia and ATP. For L-glutamate biosynthesis, *S. glossinidius* retains *gdhA* gene encoding glutamate dehydrogenase for the biosynthesis of L-glutamate from the TCA cycle intermediate α-ketoglutarate, ammonia and NADPH, and functional *aspC* gene encoding aspartate transaminase that utilizes α-ketoglutarate and L-aspartate to produce L-glutamate and oxalacetate, catalyzing also the reverse reaction for L-aspartate biosynthesis. In contrast, pseudogenization has affected *gltB* and *gltD* genes that encode the glutamate synthase enzyme complex that catalyzes the same reaction as glutamate dehydrogenase but being able to utilize both ammonia and L-glutamine as amino group donor. *S. glossinidius* also retains functional *gltP* gene encoding glutamate and aspartate DAACS transporter, although shows pseudogenized the genes *gltJ* and *gltK* encoding internal membrane components of glutamate/aspartate ABC transport system, retaining functional *gltL* and *gltI* genes encoding the ATP-binding component and the periplasmic-binding component of the ABC transport system respectively. (Figure 4.9).

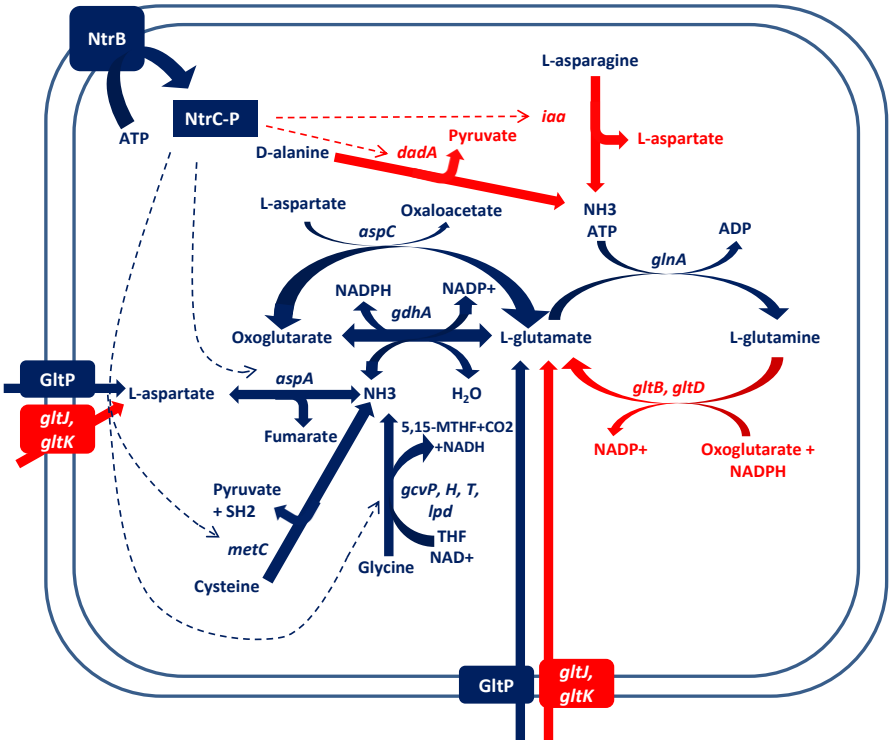


Figure 4.9: *Metabolic profile of nitrogen metabolism in S. glossinidius. In blue are represented functional reactions, whereas in red are represented pseudogenized reactions. ntrB encodes a transmembrane protein with sensor histidine kinase activity that phosphorylates nitrogen regulatory protein NtrC in response to ammonia insuficiency. Phosphorylated NtrC (NtrC-P) activates enzymes involved in the degradation of nitrogen compounds to ammonia (dashed arrows).*

L-glutamine and L-glutamate are also the main intracellular nitrogen donors for different biosynthetic reactions. The main source of nitrogen in bacterial cells is ammonia, which is assimilated as glutamate and glutamine by glutamine synthase, glutamate dehydrogenase, and glutamate synthetases enzyme complex described above. The activity of these enzymes depends on the availability of ammonia in the cell environment, with glutamine synthase and glutamate dehydrogenase being active under ammonia-rich environments, whereas glutamate synthase being mostly functional under ammonia-limited conditions. The inactivation of the genes encoding glutamate synthase enzyme complex impedes cell growth under nitrogen-limited environments in *E. coli*, *S. thypimurium*, and different *Klebsiella* species (Reitzer and Magasanik, 1987), so the inactivation of this enzymatic activity in *S.*

glossinidius indicates that the tsetse environment must be rich in nitrogen compounds in order to allow *S. glossinidius* survival.

S. glossinidius also retains *ntrB* and *ntrC* genes encoding a two component regulatory system responsible of cell response to ammonia-limited conditions. *ntrB* encodes a transmembrane protein that acts as sensor-histidine kinase, phosphorylating Nitrogen regulatory protein (NtrC) encoded by *ntrC* gene in response to increases of the ratio α-ketoglutarate/ammonia, which indicates a relative insufficiency of ammonia and an excess of carbon. Phosphorylated NtrC activates the transcription of different genes involved in transport and degradation of different nitrogen compounds, mainly amino acids, to ammonia that will be assimilated by glutamine synthase and glutamate dehydrogenase to glutamine and glutamate (Zimmer et al., 2000). Within the genes activated by this regulatory mechanism, *S. glossinidius* retains functional genes *gcvP*, *gcvH*, *gcvT* and *lpd* that encode a glycine cleavage enzyme complex responsible of glycine decarboxylation to the cofactor 5,10-methylene tetrahydrofolate, carbon dioxide and ammonia, functional *aspA* gene encoding an aspartate-ammonia lyase responsible of L-aspartate deamination to ammonia and fumarate, and a functional *metC* gene encoding a L-cysteine desulfhydrase responsible of the deamination of L-cysteine to ammonia, hydrogen sulfide and pyruvate. *aspA* gene corresponds to a situation of a longer ancestral gene where the absent part of the ancestral gene is encoded by adjacent pseudogene, although protein domain analysis indicates that the originally annotated gene retains a complete Lyase domain (Gly-Ser-XX-Met-XX-Lys-X-Asn) that has been postulated as responsible of the catalytic activity of the enzyme (Woods et al., 1988), so is reasonably to assume the functionality of these gene despite the inactivating mutations happened in the ancestral gene. In contrast, pseudogenization has affected to *dadA* gene, which encodes D-amino acid dehydrogenase responsible of the deamination of D-alanine to ammonia and pyruvate, and *iaaA* gene that encodes an asparaginase responsible of the deamination of L-asparagine to ammonia and L-aspartate. In addition, there is no functional urease enzyme complex for urea degradation to ammonia and carbon dioxide, with a single pseudogene corresponding to UreC subunit of urease, and with a single pseudogene homologous to allophanate hydrolase but no signal for urea carboxylase, responsibles of the two-steps degradation of urea to ammonia and carbon dioxide (See Figure 4.9).

4.3.5.3.2 Aspartate biosynthesis

S. glossinidius retains a functional *aspC* gene encoding aspartate transaminase, which catalyzes the biosynthesis of L-aspartate from oxalacetate and L-glutamate as amino group donor. In addition, *aspA* gene described above encoding aspartate – ammonia lyase catalyzes the reversible transamination of fumarate with ammonia to

produce L-aspartate. In addition, glutamate/aspartate DAACS transporter described above allows also the transport of L-aspartate from the environment.

4.3.5.3.3 Arginine biosynthesis

L-arginine is synthesized from L-glutamate in eight enzymatic steps, with five steps involved in L-ornithine production from L-glutamate and three enzymatic steps for the production of L-arginine from L-ornithine and carbamoyl phosphate, an essential metabolic intermediate of pyrimidine biosynthesis (Cunin et al., 1986). Two main patterns for L-arginine biosynthesis differing in the strategy followed to remove acetyl group from the intermediate N-acetyl ornithine have been described. One is a linear pathway that is found in members of the *Enterobacteriaceae* and *Bacilleae* in which N-acetyl ornithine is deacetylated by the hydrolytic enzyme acetylornithine deacetylase encoded by the gene *argE* producing L-ornithine and acetate. The other pattern is a cyclic pathway, energetically more efficient and found in most prokaryotic and eukaryotic microbes, which involves a transacylation of N-acetyl ornithine using L-glutamate as acetate acceptor, yielding L-ornithine and N-acetyl-glutamate that is the first intermediate of arginine biosynthesis pathway, through a reaction catalyzed by an ornithine acetyltransferase encoded by the gene *argJ* (Cunin et al., 1986). More recently, a third pattern for arginine biosynthesis has been characterized in *Xanthomonadales* in which N-acetyl ornithine is directly transcarbamylated to N-acetyl citrulline by specific N-acetyl ornithine transcarbamylase and then deacetylated by the common acetylornithine deacetylase to the L-arginine precursor citrulline (Morizono et al., 2006). The ancestor of *S. glossinidius*, as other members of the *Enterobacteriaceae*, had a complete pathway for L-arginine biosynthesis from L-glutamate. However, in *S. glossinidius*, pseudogenization has affected the first (*argA*), third (*argC*), fourth (*argD*) and seventh (*argG*) steps of the pathway, indicating a complete inability for L-arginine biosynthesis by none of the above described pathways given the inactivation of three of the four enzymatic activities involved in the biosynthesis of the common intermediate N-acetyl ornithine (Figure 4.10). The inactivation of the arginine biosynthetic pathway indicates that *S. glossinidius* needs L-arginine supply from the tse-tse host, which could be accomplished by a functional arginine ABC transport system encoded by the genes *artM*, *artQ*, *artI* and *artP* (Wissenbach et al., 1995), despite the inactivation of the major Lysine/Arginine/Ornithine (LAO) ABC transport system by pseudogenization of *hisM* and *hisQ* genes encoding their integral membrane components of the ABC transporter. The inactivation of the major LAO ABC transport system has also consequences in other biosynthetic pathways like putrescine biosynthesis, a polyamine involved in cell development and ribosomal function together with their derivative spermidine, which can be synthesized by two alternate pathways from L-arginine or L-ornithine. In concordance with the inactivation of L-ornithine biosynthesis from L-glutamate and L-ornithine transport through LAO ABC transport system, *S. glossinidius* has

retained functional *speA* (encoding arginine decarboxylase) and *speB* (encoding agmatinase) genes for putrescine biosynthesis from L-arginine, but shows an inactivated *speE* gene (encoding ornithine decarboxylase) required for putrescine biosynthesis from L-ornithine.

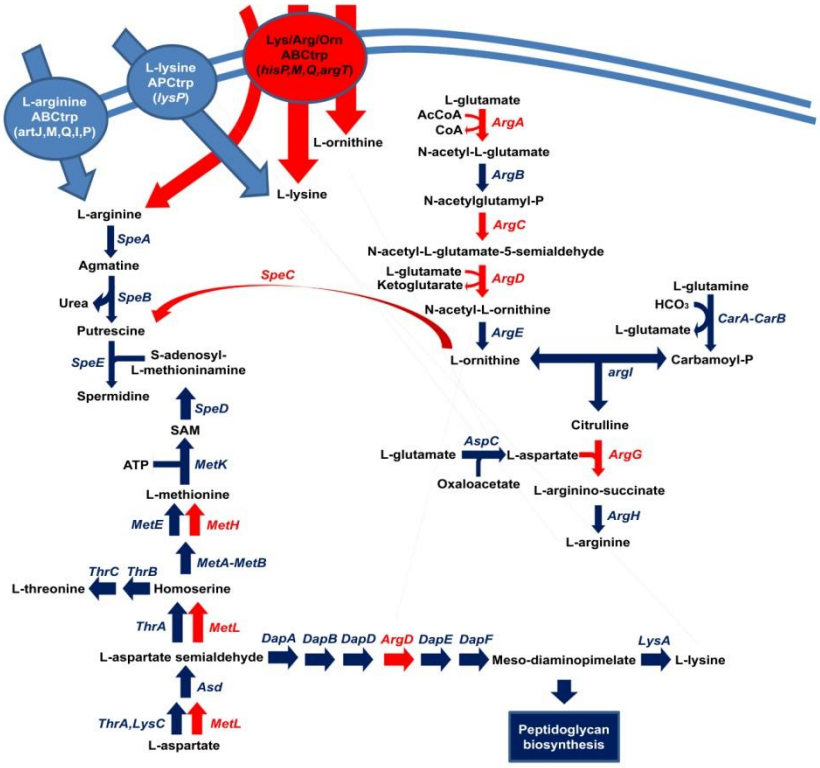


Figure 4.10: *Metabolic profile of L-arginine biosynthesis from L-glutamate in S. glossinidius and its implication in spermidine, L-lysine and peptidoglycan biosynthesis. Red coloured enzymatic reactions represent pseudogenization events, whereas blue coloured ones are those encoded by functional genes*

argD gene has been also postulated to have a succinyldiaminopimelate aminotransferase activity, required for the sixth step of L-Lysine biosynthetic pathway that produces L-Lysine and their precursor Mesodiaminopimelate, which constitutes an essential cross-linking moiety of the peptidoglycan component of bacterial cell walls (Ledwidge and Blanchard, 1999). However, this evidence comes from in-vitro assays of purified ArgD protein with different substrates, and shows that the affinity for acetylornithine is higher than for succinyldiaminopimelate. In

addition, other studies have determined that there is a specific succinyldiaminopimelate activity in *Escherichia coli* additional to *argD* encoded aminotransferase, in concordance with the presence of PLP-dependent aminotransferases of unknown function in the genome of Escherichia coli like *b2290* that encodes a predicted PLP-dependent aminotransferase, and that have a functional ortholog in the genome of *Sodalis glossinidius* (SG1602) not assigned to any pathway, so probably this uncharacterized PLP-dependent aminotransferase is responsible of succinyldiaminopimelate aminotransferase activity in L-lysine biosynthetic pathway (Cox and Wang, 2001). Alternatively, it has been also described the replacement of mesodiaminopimelate by the methionine biosynthesis intermediate cistathionine in *E.coli* (Mengin-Lecreulx et al., 1994), so another possible explanation is that cystathionine replaces mesodiaminopimelate in peptidoglycan biosynthesis and the bacteria depends also on extracellular L-lysine through a functional APC transporter encoded by the gene *lysP*.

4.3.5.3.4 Asparagine biosynthesis

L-asparagine is synthesized in a single transamination step from L-aspartate, and this reaction can be catalyzed by two different asparagine synthetases that differ in the amino group donor used for transamination reaction. Asparragine synthetase A encoded by the gene *asnA* utilizes ammonia as amino group donor, whereas asparragine synthetase B encoded by the gene *asnB* utilizes ammonia or L-glutamine as amino group donor. *S. glossinidius* retains a functional *asnB* gene, whereas there is no signal for *asnA* gene. There is also two different pseudogenes with low homology with *ansP* gene encoding L-asparagine APC transporter in *E. coli* (ps_SGL0174 and ps_SGL0881).

4.3.5.3.5 Alanine biosynthesis

L-alanine can be synthesized by three different reactions. A pyruvate-alanine aminotransferase encoded by the gene *avtA* catalyzes the transamination of pyruvate with L-valine as amino group donor to produce L-alanine and 2-ketoisovalerate, which is the precursor of L-valine biosynthesis. L-alanine is also produced by cysteine desulfurase encoded by the gene *iscS* in a reaction critical for the generation of iron-sulphur clusters, thiouridine groups in tRNA, and different sulphur- and selenium-containing proteins, although their major role in L-alanine biosynthesis has not been demonstrated (Reitzer and Magasanik, 1987; Kambampati and Lauhon, 1999; Lauhon and Kambampati, 2000). Finally, L-alanine can be also synthesized from pyruvate with L-glutamate as amino group donor by glutamate-pyruvate aminotransferase, although this enzymatic activity comes from cell extract studies with the isolation of an unsequenced *alaB* gene during a screen for genes able to complement an alanine auxotroph (Wang et al., 1987). *S. glossinidius* retains functional *avtA* and *iscS* genes, indicating that *S. glossinidius* is able to produce L-

alanine from L-valine or L-cysteine, a conclusion that contradicts results of *S. glossinidius* genome paper where postulates that the unique amino acid biosynthetic pathway inactive in *S. glossinidius* is L-alanine biosynthesis (Toh et al., 2006). In addition, *S. glossinidius* also retains a functional *yaaJ* gene encoding an alanine AGSS transporter that allows the assimilation of exogenous L-alanine.

S. glossinidius also retains a functional *dadX* gene encoding the major alanine racemase responsible of the interconversion between L- and D-alanine that is an essential component of bacterial peptidoglycan, although shows a pseudogenized *dadA* gene, which encodes a D-amino acid dehydrogenase that catalyzes the deamination of D-alanine to pyruvate and ammonia, essential for cell growth with L-alanine as sole energy and nitrogen source.

4.3.5.3.6 Serine and glycine biosynthesis

S. glossinidius retains functional *serA*, *serB* and *serC* genes for L-serine biosynthesis from the glycolytic intermediate 3-phosphoglycerate, together with a functional *glyA* gene encoding a serine hydroxymethyltransferase that catalyzes L-glycine biosynthesis from L-serine and tetrahydrofolate, producing also 5,10-methylene-tetrahydrofolate, the major source of one-carbon units in bacterial cells. In addition, *S. glossinidius* retains a completely functional glycine cleavage system (*gcvP, H, T, lpd*) for the decarboxylation of glycine to ammonia, carbon dioxide and 5,10-methylene-tetrahydrofolate. This enzymatic complex, in addition to provide ammonia in conditions of nitrogen scarcity through Ntr response, maintains appropriate levels of glycine and 5,10-methylene-tetrahydrofolate by conversion of glycine excess to 5,10-methylene-tetrahydrofolate. For L-serine degradation, *S. glossinidius* retains a functional *sdaA* gene encoding serine deaminase that catalyzed the deamination of L-serine to pyruvate and ammonia. *S. glossinidius* also retains a functional *sdaC* gene that encodes a serine/proton symporter.

4.3.5.3.7 Threonine, methionine and lysine biosynthesis

These three amino acids are synthesized from the common precursor L-aspartate semialdehyde, which is produced from L-aspartate through their phosphorylation to L-asparthyl-4-phosphate by aspartate kinase followed by their reduction to L-aspartate semialdehyde by aspartate semialdehyde dehydrogenase. *S. glossinidius* retains functional *thrA* and *lysC* genes encoding a bifunctional aspartate kinase/homoserine dehydrogenase and an aspartate kinase respectively, together with functional *asd* gene encoding aspartate semialdehyde dehydrogenase. Pseudogenization has affected *metL* gene that encodes a second bifunctional aspartate kinase/homoserine dehydrogenase, by a single frameshift mutation in the middle of the ancestral gene, indicating a recent inactivation event (Figure 4.11). In fact, protein domains analysis with the two open reading frames of *metL* pseudogene shows that N-terminal ORF contains the complete aspartate kinase domain

(PF:00696) whereas the C-terminal ORF contains the complete homoserine dehydrogenase domain (PF:00742, 03447), so it is possible that *S. glossinidius* could still express aspartate kinase activity of *metL*, given their high identity with functional homologs at start codon level, with stop codon 88 base pairs downstream PF:00696 domain.

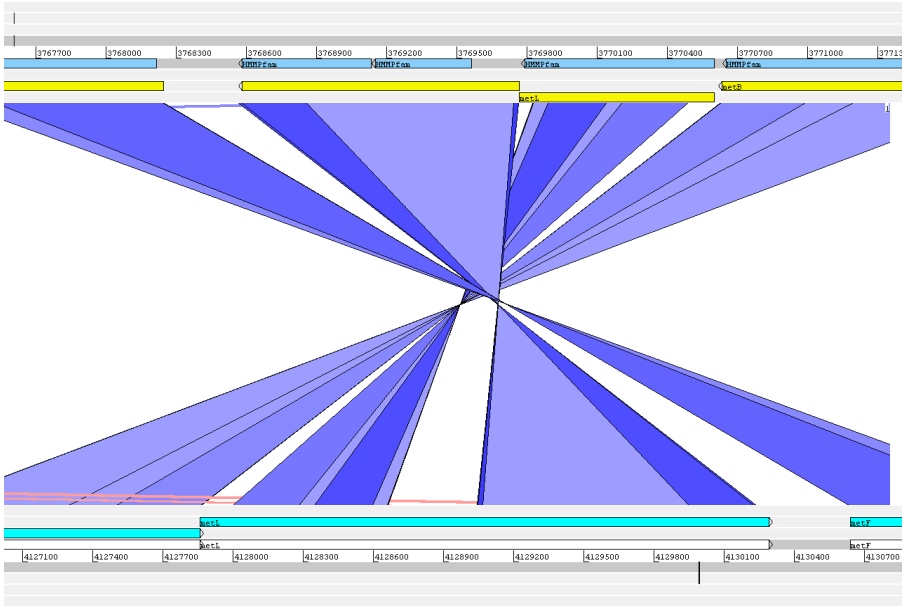


Figure 4.11: *ACT comparison between S. glossinidius (up) and E. coli K12 (down) at metL gene level. Blue features in S. glossinidius metL pseudogene corresponds to PFAM domains aspartate kinase (upstream frame) and homoserine dehydrogenase (downstream frame).*

From L-aspartate semialdehyde, L-lysine is synthesized in seven enzymatic steps catalyzed by enzymes encoded by the genes *dapA*, *dapB*, *dapD*, *argD*, *dapE*, *dapF* and *lysA*, all of which are functional in *S. glossinidius* except *argD*, which encodes bifunctional N-succinyldiaminopimelate aminotransferase/acetylornithine transaminase also involved in arginine biosynthesis (Figure 4.12).

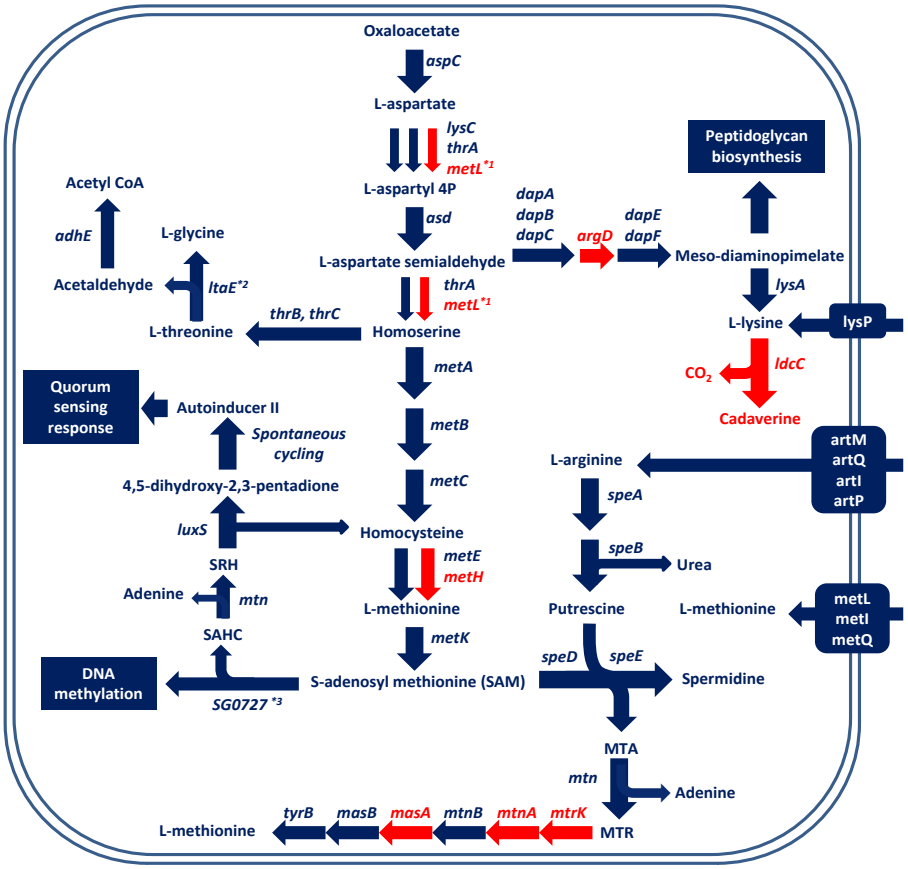


Figure 4.12: *Metabolic profile of methionine, lysine and threonine biosynthesis in S. glossinidius. Red coloured enzymatic reactions represent pseudogenization events, whereas blue coloured ones are those encoded by functional genes.*

**1 Pseudogenization ofmetL by single frameshift mutation in the middle of ancestral gene, retaining functional aspartate kinase domain in the N-terminal frame of the pseudogene (start and stop codons flanking aspartate kinase domain), pointing out to the possible functionality of aspartate kinase activity.*

**2 Situation of longer ancestral gene where the absent part of the protein is present as adjacent pseudogene.*

**3 Phage methyltransferase*

As is described in the arginine biosynthesis section, alternative PLP-dependent aminotransferase encoded by the gene SG1602 can be responsible of the N-succinylaminopimelate activity. L-methionine and L-threonine are synthesized from the common precursor homoserine, which is produced by oxidation of L-aspartate semialdehyde by one of the two bifunctional enzymes aspartate kinase/homoserine dehydrogenase described above (*metL* or *thrA*).

For threonine catabolism, *S. glossinidius* retains *ltaE* gene encoding a low-specificity threonine aldolase, which catalyzes the degradation of L-threonine to acetaldehyde and glycine, with acetaldehyde that can be converted to AcCoa functional alcohol dehydrogenase (*adhE*). However, *lteE* corresponds to a longer ancestral gene where the absent part of the protein is present as an adjacent pseudogene that have disrupt the domain Beta-eliminating lyase (PF:01212) responsible of the catalytic activity of the enzyme. Pseudogenization has also affected to *ldcC* gene encoding a constitutive lysine decarboxylase responsible of the decarboxylation of L-lysine to carbon dioxide and cadaverine, a polyamine that is part of the lysine-dependent resistance system 4 that confers resistance to weak organic acids produced during carbohydrate fermentation under conditions of anaerobiosis and phosphate starvation, with no signal of the *cadA* gene encoding the inducible isoform of lysine decarboxylase (Lemonnier and Lane, 1998).

S. glossinidius also retains functional *metK* gene encoding a methionine adenosyltransferase responsible of S-adenosylmethionine (SAM) biosynthesis from L-methionine and ATP. SAM is a methyl-group donor involved in DNA methylation and spermidine biosynthesis (Figure 4.12). In spermidine biosynthesis, spermidine is synthesized from SAM and putrescine by the sequential action of adenosylmethionine decarboxylase (*speD*) and spermidine synthase (*speE*), generating as by-product S-methyl-5'-thioadenosine (MTA) (Figure 4.12). *S. glossinidius* retains functional *speE* and *speD* genes for spermidine biosynthesis together with functional genes for putrescine biosynthesis from L-arginine (see L-arginine metabolism section). MTA is the starting point for the salvage pathway for L-methionine (*mtn,mtrK, mtnA, mtnB, masA, masB*) that is present in *S. glossinidius* although with pseudogenization that have affected to *mtnK*, *mtnA* and *masA*, retaining functional *mtnN, mtnB,* and *masB* genes, reflecting an ongoing process of pathway inactivation. In *E. coli*, no methionine salvage pathway is found but functional 5-methilthioadenosine/S-adenosylhomocysteine nucleosidase (*mtnN* gene, b0159) hydrolyzes MTA to S-methyl-5-thio-D-ribose (MTR) that is excreted into the environment (Hughes, 2006), so similar behaviour can be expected in *S. glossinidius* despite the ancestral functionality of MTA recycling pathway to L-methionine.

In DNA methylation, SAM acts as methyl group donor for cytosine methylation by cytosine methyltransferase, which in *E. coli* appears encoded by the gene *dcm*, producing 5-methylcysteine and S-adenosyl-L-homocysteine that is hydrolyzed to adenine and S-D-ribosyl-L-homocysteine (SRH) by the bifunctional MTA/SAHC nucleosidase encoded by the gene *mtnN* also involved in the first step of methionine salvage pathway (Figure 4.12). SRH is finally degraded to the methionine precursor L-homocysteine and 4,5-dihydroxy-2,3-pentadione by SRH lyase encoded by the gene *luxS*. *S. glossinidius* retains functional *mtnN* and *luxS* genes, although there is no ortholog to *dcm* gene encoding cysteine methyltransferase, with only a phage

methyltransferase gene with similar function (SG0727). The spontaneous cycling of the final metabolite 4,5-dihydroxy-2,3-pentadione produces a molecule named autoinducer II, an important molecule for quorum-sensing response in bacteria, controlling gene expression in response to cell population density through their secretion and detection by bacterial cells (Schauder et al., 2001). Quorum sensing response mediated by autoinducer II is responsible of different cellular responses in pathogenic bacteria like activation of virulence genes and flagellar regulon genes through cell-to-cell signalling, being characteristic of bacteria that lives in close proximity to plants and animals (de Kievit and Iglewski, 2000). The presence of this mechanism in *S. glossinidius* that have a broad tissue specificity both intra- and extra-cellularly, may be involved in bacterial migration across tsetse body in response to nutrient availability.

In addition to endogenous biosynthesis, *S. glossinidius* also retains functional ABC transport systems for L- and D-methionine (*metN, I, Q*) and a functional lysine APC transporter (*lysP*), although pseudogenization has affected to all genes of the high-affinity lysine/arginine/ ornithine ABC transport system (*hisP, M, Q, J*).

4.3.5.3.8 Cysteine biosynthesis

S. glossinidius retains all genes for cysteine biosynthesis from inorganic sulphate (SO_4^{-2}) that includes *cysD, N, C, H, I, J* genes for hydrogen sulphide (SH_2) biosynthesis from inorganic sulphate and *cysK* gene encoding cysteine synthase A for L-cysteine biosynthesis from hydrogen sulphide and O-acetyl-L-serine. In addition, *S. glossinidius* also retains functional *cysE* gene encoding a serine acetyltransferase for O-acetyl-L-serine biosynthesis by acetylation of L-serine with acetate group from AcCoA.

In addition to their role in protein biosynthesis, the pathway for L-cysteine biosynthesis is the main route for inorganic sulfur assimilation in bacterial cells. Inorganic sulfur in the form of sulphate enters inside bacterial cells through specific ABC transport system encoded by the genes *cysA, T, W, P* and *sbp*, which is also completely functional in *S. glossinidius*. This ABC transport system is also capable of transport thiosulfate, which is converted to S-sulfocysteine by a second cysteine synthase B encoded by the gene *cysM* that is also functional in *S. glossinidius*.

For cysteine catabolism, *S. glossinidius* retains functional *metC* gene encoding a bifunctional enzyme with cystathionine β-lyase/cysteine desulfhydrase activity. Cystathionine β-lyase catalyzes the degradation of cystathionine to L-homocysteine, pyruvate, and ammonia during L-methionine biosynthesis (Figure 4.12), whereas cysteine desulfhydrase activity degrades L-cysteine to pyruvate, ammonia, hydrogen sulphide (Awano et al., 2003).

4.3.5.3.9 Valine, leucine and isoleucine biosynthesis

S. glossinidius retains all functional genes for branched-chain aminoacid biosynthesis, including *ilvG, M, C, D, E, A* genes for L-valine and L-isoleucine biosynthesis from pyruvate, and *leuA, C, D, B* for L-leucine biosynthesis from 2-ketoisovalerate. *S. glossinidius* also retains functional *tyrB* gene encoding a broad-specificity aromatic-acid aminotransferase that is able to catalyze the final transamination reaction in tyrosine, phenylalanine, and leucine biosynthesis (Onuffer et al., 1995). Pseudogenization affects to *ilvI* and *ilvH* genes, which encodes the acetolactate synthase enzyme complex III, although retains functional *ilvG* and *ilvM* genes encoding the acetolactate synthase enzyme complex II. In addition to endogenous biosynthesis, *S. glossinidius* retains functional *brnA* gene encoding a low-affinity leucine/isoleucine/valine-proton symporter, although pseudogenization has affected the complete *liv* operon (*livF, G, M, H, K, J*) that encodes a high-affinity ABC transport system for branched-chain amino acids.

4.3.5.3.10 Proline biosynthesis

S. glossinidius retains functional *proB*, *proA*, and *proC* genes encoding a glutamyl kinase, glutamate-5 –semialdehyde dehydrogenase, and pyrroline-5-carboxylate reductase respectively, responsible of L-proline biosynthesis from L-glutamate. In addition, L-proline can be also synthesized by enzymes of the pathway for L-arginine biosynthesis, due to the fact that both pathways are rather similar. L-arginine is synthesized from L-glutamate that is first converted to N-acetyl-L-glutamate, but the next steps of phosphorylation by acetylglutamate kinase (*argB*) and NADPH-dependent reduction to N-acetyl-L-glutamate-5-semialdehyde by N-acetylglutamylphosphate reductase (*argC*) are equivalent to the reactions catalyzed by glutamate kinase (*proB*) and glutamate-5-semialdehyde dehydrogenase (*proA*) in L-proline biosynthesis. In situations of inactivation of *argD*, bacteria accumulates N-acetyl-L-glutamate-semialdehyde, which is directly converted to the L-proline precursor L-glutamate semialdehyde by the acetylornithine deacetylase encoded by the gene *argE*, increasing the production of L-proline (Leisinger, 1987). In fact, it has been demonstrated that double mutants *proA-argD* or *proB-argD* are capable of L-proline biosynthesis through this alternative pathway with enzymes of L-arginine biosynthesis (Itikawa et al., 1968; Kuo and Stocker, 1969; Berg and Rossi, 1974). However, in *S. glossinidius*, the pseudogenization of *argA* and *argC* genes involved in N-acetyl-L-glutamate semialdehyde biosynthesis from L-glutamate in the context of L-arginine biosynthesis described above blocks this alternative pathway for L-proline biosynthesis.

Pseudogenization has affected to different genes involved in L-proline transport, like the inactivation of *proV* and *proX* that encodes the ATP-binding component and

the periplasmic-binding protein respectively of a L-proline ABC transport system, and the inactivation of *proP* gene encoding a L-proline MFS transporter.

For L-proline degradation, *S. glossinidius* retains functional *purA* gene encoding a bifunctional proline dehydrogenase that catalyzes the degradation of L-proline to L-glutamate during conditions of nitrogen absence or starvation.

4.3.5.3.11 Histidine biosynthesis

S. glossinidius retains a completely functional *his* operon (*hisG, D, C, B, A, F, I*) for L-histidine biosynthesis from the pentose ribose-5-phosphate, and intermediate of the pentose phosphate pathway. In addition, during histidine biosynthesis is produced aminoimidazole carboxamide ribonucleotide (AICAR), an important intermediate for purine ribonucleotides biosynthesis. Pseudogenization has affected to the complete operon *hisP, M, Q, J* that encodes a high-affinity histidine ABC transport system.

4.3.5.3.12 Phenylalanine, tyrosine and tryptophan biosynthesis

Phenylalanine, tyrosine, and tryptophan, also known as aromatic amino acids due to the presence of an aromatic ring in their structure, are synthesized from the pentose phosphate pathway intermediate erythrose-4-phosphate through the shikimate pathway that finishes with the production of chorismate, a common metabolic precursor for the biosynthesis of aromatic amino acids, folate, ubiquinone, and menaquinone. *S. glossinidius* retains all *aro* genes for chorismate biosynthesis from erythrose-4-phosphate (*aroF, H, G, D, E, K, L, A, C*), including functional genes for the three different isozymes corresponding to 2-dehydro-3-deoxyphosphoheptonate aldolase (*aroF, G, H*) and the two isozymes for shikimate kinase (*aroK* and *aroL*).

S. glossinidius also retains functional *trp* genes (*trpD, E, C, A, B*) for L-tryptophan biosynthesis from chorismate, and functional *tyrA*, *pheA*, and *tyrB* genes for L-phenylalanine and L-tyrosine biosynthesis from chorismate through their common intermediate prephenate. In addition, *S. glossinidius* also retains functional *aroP* gene encoding an APC transporter for aromatic amino acid transport.

4.3.5.4 Fatty acids metabolism

S. glossinidius retains the complete pathway for "de-novo" biosynthesis of saturated and unsaturated fatty acids from AcCoA, with a functional AcCoA carboxylase enzyme complex (*accA, D*), biotin carboxylase (*accC*) and biotin carrier protein BCCP (*accB*) for the biotin-dependent carboxylation of AcCoA to MalonylCoA, which will be converted to the fatty acid precursor Malonyl-ACP by a functional MalonylCoA-ACP transacylase (*fabD*). *S. glossinidius* also retains

functional *fabB*, *fabF* and *fabH* genes encoding three different β-ketoacyl-ACP synthases for acetoacetyl-ACP biosynthesis from malonyl-ACP and functional *fabG*, *fabZ*, and *fabI* genes for the biosynthesis of saturated fatty acids by successive cycles of elongation of the acetoacetyl-ACP precursor with subunits of malonyl-ACP, increasing fatty acid chain length by two carbons at each elongation cycle (Figure 4.13). In addition, functional *fabA* gene encodes a 3-hydroxydecanoyl-[ACP] dehydrase responsible of the first step of unsaturated fatty acid biosynthesis by interconversion of saturated fatty acid precursor trans-decenoyl-ACP into cis-decenoyl-ACP, which will be elongated by the same enzymatic activities responsible of saturated fatty acid elongation (Figure 4.13).

S. glossinidius also retains functional *plsB* and *plsC* genes encoding two different acyltransferases that catalyze the esterification of acyl-chains of different fatty acids with positions 1 and 2 of glycerol-3-phosphate producing 1,2-diacyl-sn-glycerol-3-phosphate, and functional *cdsA* gene encoding CDP-diglyceride synthase for CDP-diacylglycerol biosynthesis from 1,2-diacyl-sn-glycerol-3-phosphate. CDP-diacylglycerol is the common precursor of the three major membrane phospholipids presents in bacterial cells, phosphatidylethanolamine (PE), phosphatidylglycerol (PG), and cardiolipin (CL). *S. glossinidius* retains functional pathways for the biosynthesis of all three major phospholipids, with functional *pssA* and *psd* genes for PE biosynthesis from CDP-diacylglycerol and serine, functional *pgsA*, *pgpA* and *pgpB* genes for PG biosynthesis from CDP-diacylglycerol and glycerol-3-phosphate, and functional *cls* gene encoding the major cardiolipin synthase for CL biosynthesis by condensation of two PG molecules (Figure 4.13).

With regard to glycerol-3-phosphate biosynthesis, the essential backbone of all bacterial phospholipids, it can be synthesized by transport and phosphorylation of extracellular glycerol, degradation of membrane phospholipids, and "de-novo" biosynthesis from the glycolytic intermediate dihydroxyacetone-3-phosphate (Figure 4.13). *S. glossinidius* retains a functional *glpK* gene encoding a glycerol kinase for glycerol phosphorylation to glycerol-3-phosphate, although pseudogenization has affected *glpF* gene, which encodes a facilitator channel protein for glycerol diffusion through cytoplasmic membrane, and with no signal of *glpT* gene that encodes a specific MFS transporter of glycerol. Of the two different bacterial phospholipases, *S. glossinidius* retains functional *pldA* gene encoding phospholipase A1, also known as OMPLA (outer membrane phospholipase A1), but have pseudogenized *pgpB* gene, which encodes phospholipase A2. In addition, pseudogenization has affected the complete *ugp* operon (*ugpB, A, E, C, Q*) that encodes a specific glycerophosphodiester ABC transport system for the transport of glycerophosphodiesters produced by phospholipase-mediated phospholipid degradation to the cytoplasm (*ugpB, A, E, C*) and a cytoplasmic phosphodiesterase (*ugpQ*) for glycerophosphodiester degradation to glycerol-3-phosphate and alcohol,

with no signal of *glpQ* gene that encodes a second periplasmic phosphodiesterase catalyzing the same reaction as *ugpQ*.

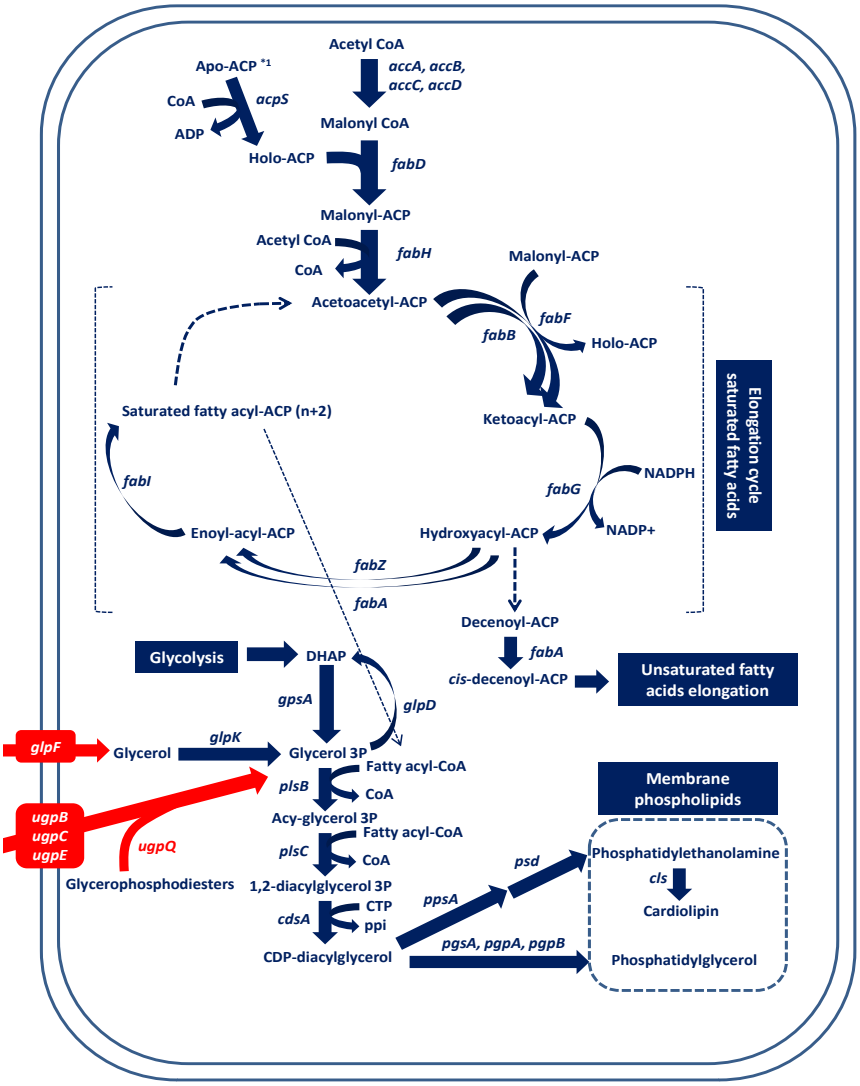


Figure 4.13: *Metabolic profile of fatty acid and membrane phospholipid biosynthesis in S. glossinidius. Red coloured enzymatic reactions represent pseudogenization events, whereas blue coloured ones are those encoded by functional genes.*
*[*1] ACP precursor (apo-ACP) is encoded by functional acpP gene*

Finally, *S. glossinidius* retains a functional *gpsA* gene that encodes a biosynthetic glycerol-3-phosphate dehydrogenase that catalyzes the NADPH-dependent reduction of the glycolytic intermediate dihydroxyacetone-3-phosphate to glycerol-3-phosphate. *S. glossinidius* retains also functional *glpD* gene encoding an aerobic glycerol-3-phosphate dehydrogenase that catalyzes the reverse reaction, the oxidation of glycerol-3-phosphate to dihydroxyacetone-3-phosphate with the concomitant transfer of electrons to the ubiquinone pool of the respiratory chain, essential for utilization of glycerol-3-phosphate or their precursors, glycerol and glycerophosphodiesters, as carbon sources.

In contrast, there is no signal of *fad* genes for fatty acid degradation, and pseudogenization has affected also to *tesA* and *yciA* genes encoding thioestherase I and Acyl-CoA thioestherase respectively that catalyze the hydrolysis of a broad range of acyl-CoA molecules to Coenzyme A and fatty acids.

4.3.5.5 Nucleotides biosynthesis

Purine and pyrimidine nucleotides for DNA and RNA biosynthesis can be synthesized in bacterial cell "de novo" from ribose-5-phosphate and L-glutamine respectively, or through recycling or salvage pathways that allows exogenous nucleotides or nucleotides generated during RNA turnover to be used for DNA biosynthesis through different interconversion reactions. In the next sections I will briefly describe the nucleotides metabolism in *S. glossinidius* and how pseudogenization has affected differentially both biosynthetic and salvage pathways for both purines and pyrimidines.

4.3.5.5.1 Purine nucleotides biosynthesis

S. glossinidius retains functional *prsA* gene encoding a ribose-phosphate diphosphokinase that catalyzes the initial phosphorylation of ribose-5-phosphate to 5-phosphoribosyl 1-pyrophosphate (PRPP), together with functional *pur* genes (*purF, D, N, T, L, M, K, E, C, B, H*) for the biosynthesis of inositol-5-phosphate (IMP) from PRPP. Pseudogenization have affected only to *purN* gene encoding a phosphoribosylglycinamide formyltransferase (GAR formyltransferase), one of the two isozymes responsible of the third step in purine nucleotide biosynthesis, the transfer of formyl group to 6-phosphoribosyl glycineamide (GAR) using 5,10-formyl tetrahydrofolate as formyl group donor. However, *S. glossinidius* retains functional *purT* gene encoding the second isozyme GAR transformylase that uses directly formate as formyl group donor. Formate can be obtained from degradation of pyruvate through functional *pflB* gene encoding anaerobic pyruvate-formate lyase or through functional *purU* gene encoding 10-formyl tetrahydrofolate hydrolase that catalyzes the hydrolysis of 10-formyl tetrahydrofolate to formate and tetrahydrofolate (Figure 4.14).

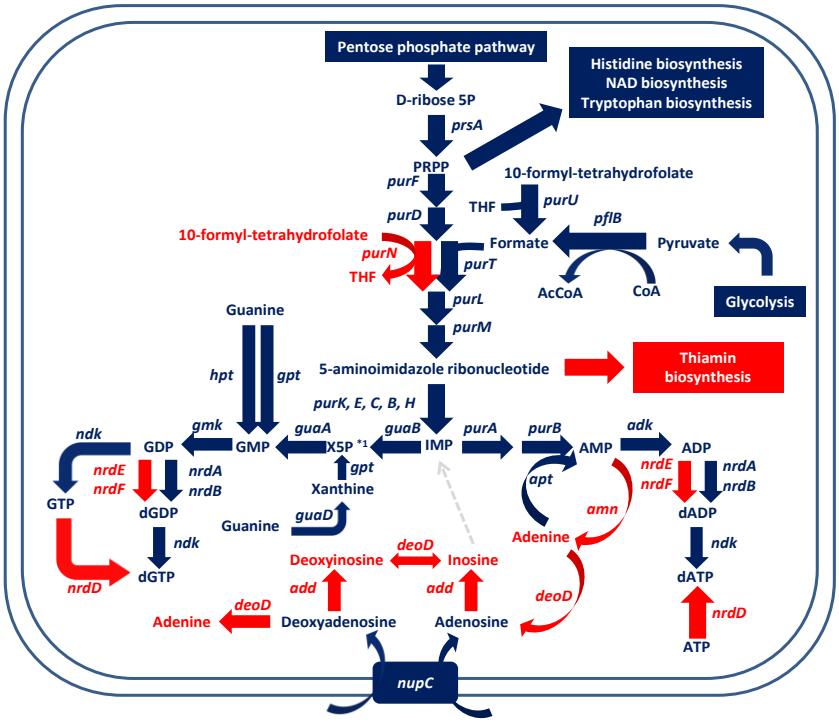


Figure 4.14: *Metabolic profile of purine nucleotides biosynthesis and recycling pathways in S. glossinidius. Red coloured enzymatic reactions represent pseudogenization events, whereas blue coloured ones are those encoded by functional genes. Dashed arrow represents reaction catalyzed by bifunctional inosine/guanosine kinase, absent in S. glossinidius.*
*[*1] Xanthosine 5-phosphate*

From IMP, *S. glossinidius* retains functional adenylosuccinate synthase (*purA*) and adenylosuccinate lyae (*purB*) for ADP biosynthesis and IMP dehydrogenase (*guaB*) and GMP synthetases (*guaA*) for GMP biosynthesis. Of the three different ribonucleotide reductase complexes for the conversion of ribonucleotides to deoxyribonucleotides, *S. glossinidius* retains functional *nrdA* and *nrdB* genes, which encodes the major ribonucleotide diphosphate reductase Ia enzyme complex under aerobic conditions, but pseudogenization has affected to *nrdE* and *nrdF* genes encoding the minor ribonucleotide diphosphate reductase Ib complex, involved in response to oxidative stress, and to *nrdD* gene encoding the anaerobic ribonucleotide triphosphate reductase III, essential for nucleotide biosynthesis under anaerobic conditions. *S. glossinidius* also retains functional *ndk* gene encoding nucleoside diphosphate kinase responsible of phosphorylation of ribonucleotides and deoxyribonucleotides diphosphate (CDP, GDP, UDP, dADP, dTDP, dCDP, dGDP) to their corresponding triphosphate derivatives (Figure 4.14). Intracellular levels of

ATP are considerably higher than other nucleotides diphosphate due to their prevalent role as the main energy source for different biosynthetic reactions, being produced by alternative pathways like through functional ATP synthase enzyme complex coupled with respiratory chain or through functional pyruvate kinase coupled to the dephosphorylation of glycolytic intermediate phosphoenolpyruvate to pyruvate during the glycolysis.

In contrast, pseudogenization has affected genes involved in purine salvage and interconversion pathways (Figure 4.14). Pseudogenization of *deoD* gene encoding reversible purine nucleoside phosphorylase impedes the recycling of purine nucleosides (adenosine and guanosine), whereas pseudogenization of *amn* and *add* genes impedes AMP-GMP interconversions through IMP. In concordance with the inability of purine nucleoside recycling, there is no signal of *nupG* gene that encodes a high-affinity transport system for purine and pyrimidine nucleosides, retaining functional *nupC* gene encoding a specific high-affinity transport system for pyrimidine nucleosides and adenosine that is in concordance with the retention of functional *deoA* gene encoding a specific reversible pyrimidine nucleoside phosphorylase equivalent to *deoD* but active only with pyrimidine nucleosides. However, *S. glossinidius* retains all genes encoding purine phosphoribosyltransfeases (*hpt*, *gpt*, *apt*), which allows the direct conversion of purine bases (hypoxanthine, xanthine, guanine, adenine) to their nucleoside monophosphate derivatives, as well as functional *guaD* gene encoding guanine deaminase for guanine deamination to xanthine (Figure 4.14).

S. glossinidius also retains functional *surE* and *ybfR* genes encoding periplasmic nucleotide phosphatases that catalyze the dephosphorylation of exogenous nucleotides to their corresponding nucleosides. This means that exogenous nucleotides can be dephosphorylated by periplasmic phosphatases SurE and YbfR to their corresponding nucleosides, but only pyrimidine nucleosides and adenosine can be assimilated by *S. glossinidius* through functional high affinity nucleoside transporter NupC, and only pyrimidine nucleosides can be assimilated through specific pyrimidine nucleoside phoshorylase DeoA. The pseudogenization of genes involved in purine salvage pathways may also explain the inactivation of *purR* gene that encodes a purine repressor of *pur* genes for "de-novo" purine biosynthesis. PurR binds guanine and hypoxanthine and represses the expression of *pur* genes involved in IMP biosynthesis from ribose-5-phosphate in conditions with purines presents in the medium.

Finally, *S. glossinidius* retains functional *spoT* and *relA* genes that encodes pyrophosphokinases I and II respectively, responsibles of the condensation of GTP and ATP to produce ppGpp, an important regulator of the stringent response in bacterial cells that coordinates a variety of cellular activities in response to changes in nutritional abundance (Murray and Bremer, 1996).

4.3.5.5.2 Pyrimidine nucleotides biosynthesis

The pathway for "de-novo" biosynthesis of pyrimidine nucleotides can be considered as an unbranched sequence of reactions that finishes with the production of dTTP and in which UTP, CTP, and dCTP are intermediates of the pathway (Figure 4.15).The unique branching point of the pathway corresponds to the first enzymatic reaction, the production of carbamoyl phosphate from L-glutamine, bicarbonate, and ATP, because carbamoyl phosphate is also an important intermediate of arginine biosynthesis. *S. glossinidius* retains the complete pathway for "de-novo" pyrimidine nucleotides biosynthesis, with functional *carA* and *carB* genes encoding carbamoyl phosphate synthase enzyme complex and functional *pyr* genes (*pyrC, D, E, F*) for UMP biosynthesis from carbamoyl phosphate. UMP is then phosphorylated to UDP by a functional uridylate kinase (*udk*), and UDP is substrate of the general nucleoside diphosphate kinase (*ndk*) producing UTP. CTP is synthesized by direct transamination of UTP by functional CTP synthetases (*pyrG*). The general ribonucleotide reductase Ia enzyme complex synthesizes pyrimidine deoxyribonucleotides (dUTP and dCTP) from their corresponding ribonucleotides (UTP and CTP). Thymine nucleotides are deoxy compounds that do not have ribonucleotide counterparts because are only used for DNA biosynthesis, and therefore cannot be produced by ribonucleotide reductases. Their synthesis starts with dUTP dephosphorylation by functional deoxyuridine triphosphatase (*dut*) to dUMP, which is methylated by functional thymidylate synthase (*thyA*) using 5,10-methylene tetrahydrofolate producing dTMP and the folate biosynthesis intermediate 7,8-dihydrofolate. dTMP is then phosphorylated by specific dTMP kinase (*tmk*) producing dTDP, which is finally phosphorylated by the general nucleoside diphosphate kinase Ndk to dTTP (Figure 4.15).

In addition, *S. glossinidius* also retains functional genes for most enzymatic activities of pyrimidine salvage pathway, with functional *nupC* gene encoding a high-affinity pyrimidine nucleoside transporter (both ribonucleosides and deoxyribonucleosides) and functional *deoA* gene encoding a thymidinephosphorylase that catalyzes the phosphorolysis of pyrimidine deoxyribonucleosides (deoxycytidine, deoxyuridine, and deoxytymidine) to their corresponding pyrimidine bases and deoxyribose 1 phosphate. Deoxyribose 1 phosphate can be metabolized to the glycolytic intermediate glyceraldehyde-3-phosphate and acetaldehyde by functional *deoB* and *deoC* genes encoding a phosphopentomutase and deoxyribose aldolase respectively (Figure 4.15). Acetaldehyde can be converted to AcCoA by functional alcohol dehydrogenase (*adhE*).

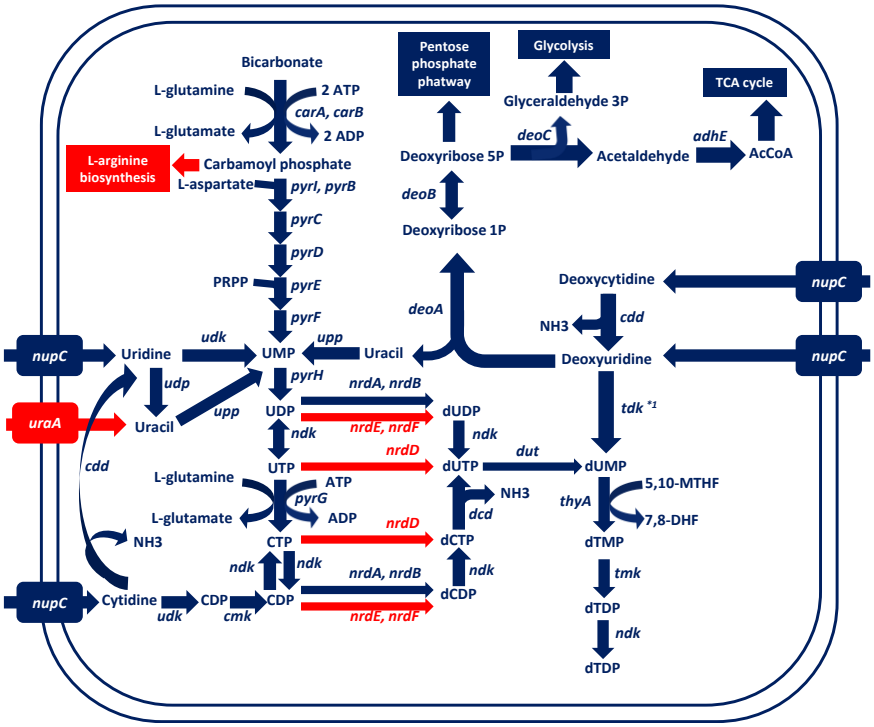


Figure 4.15: *Metabolic profile of pyrimidine nucleotides biosynthesis and recycling pathways in S. glossinidius. Red coloured enzymatic reactions represent pseudogenization events, whereas blue coloured ones are those encoded by functional genes.*

**1 Situation of a putatively longer ancestral gene where the absent part of the protein is present as adjacent pseudogene.*

For the salvage of pyrimidine ribonucleosides (cytidine and uridine), there is two possible salvage pathways both functional in *S. glossinidius*. Uridine and cytidine can be phosphorylated to UMP and CMP respectively by a bifunctional cytidine/uridine kinase (*udk*). UMP enters to the pathway of "de-novo" biosynthesis, whereas CMP can be phosphorylated by a functional cytidylate kinase (*cmk*) yielding CDP, which is phosphorylated by the general nucleotide diphosphate kinase Ndk to CTP (Figure 4.15). In a second pathway, cytidine can be converted to uridine by bifunctional cytidine/deoxycytidine deaminase (*cdd*), and uridine can be hydrolyzed to uracil and ribose-1-phosphate by specific uridine phosphorylase (*udp*). Ribose-1-P is converted to the pentose phosphate pathway intermediate ribose-5-phosphate by the phosphopentomutase DeoB described above, whereas uracil can be converted to UMP by functional uracil phosphoribosyltransferase (*upp*) (Figure 4.15).

Pseudogenization has affected to *codA* gene that encodes a cytosine deaminase for cytosine conversion to uracil, and *uraA* gene that encodes a specific uracil transporter. In addition, the originally annotated *tdk* gene that encodes a bifunctional thymidine/deoxyuridine kinase for the phosphorylation of thymidine and deoxyuridine to dTMP and dUMP respectively corresponds to a situation of a longer ancestral gene where the absent part of the protein appears as an adjacent pseudogene in a different frame, disrupting the functional domain thymidine kinase (PF:00265), so this enzymatic activity would be probably inactivated.

4.3.5.6 Cofactors biosynthesis

The supply of cofactors to the tsetse fly is the main objective of the symbiotic association between the primary endosymbiont *W. glossinidia* and the tsetse host, because cofactors are devoid in the vertebrate blood that constitutes the strict diet of tsetse flies. This is reflected by the preservation of cofactor biosynthesis genes in the highly streamlined genome of *W. glossinidia*. Comparison of the functional profile of both *W. glossinidia* and *S. glossinidius* (Akman et al., 2002; Toh et al., 2006) reveals that all metabolic capabilities of *W. glossinidia* are present completely functional in *S. glossinidius*, with the only exception of genes involved in thiamine biosynthesis. In the next sections, I will compare the cofactor biosynthetic capabilities of *S. glossinidius* and *W. glossinidia* to gain insight in the ecological association between the tsetse fly and their primary and secondary endosymbionts.

4.3.5.6.1 Riboflavin biosynthesis

Riboflavin is the precursor of the essential cofactors flavin adenine dinucleotide (FAD) and flavin mononucleotide (FMN), essential for a wide range of redox reactions. *S. glossinidius* and *W. glossinidia* retains functional *rib* genes (*ribA*, *D*, *E*, *C*) for riboflavin biosynthesis from GTP together with functional *ribC* gene encoding a bifunctional riboflavin kinase/FMN adenyltransferase for FMN biosynthesis by riboflavin phosphorylation and FAD biosynthesis by FMN adenylation.

4.3.5.6.2 Pyridoxal 5-phosphate biosynthesis

Pyridoxal-5-phosphate (PLP) is the biologically active form of Vitamin B6, a water-soluble vitamin that can be found in three different chemical conformations named pyridoxamine (PM), pyridoxal (PL), and pyridoxine (PN), all of which are precursors of the active form PLP, which is an essential cofactor of enzymes involved in amino acid metabolism and for glycogen phosphorylases (Hill and Spenser, 1996). Their "de-novo" biosynthesis proceeds through the convergence of two different pathways that lead to the synthesis of 4-phosphohydroxy-L-threonine

(4PHT) from the pentose phosphate pathway intermediate erythrose-4-phosphate (*epd*, *pdxB*, and *serC* genes) and 1-deoxy-D-xylulose-5-phosphate from glycolytic intermediates glyceraldehyde-3-phosphate and pyruvate (*dxs* gene) respectively. 4PHT is oxidized by 4PHT dehydrogenase encoded by the gene *pdxA* and the product is spontaneously decarboxylated before their condensation with 1-deoxy-D-xylulose-5-phosphate to produce pyridoxine-5-phosphate (PNP) in a reaction catalyzed by pyridoxine-5-phosphate synthase encoded by the gene *pdxJ*. PNP is finally oxidized to the biologically active form PLP by PNP oxidase encoded by the gene *pdhX*.

S. glossinidius and *W. glossinidia* retain all functional genes for all enzymatic activities of PLP biosynthesis except the gene *epd* that appears absent in both genomes. The *epd* gene encodes an erytrhose-4-phosphate dehydrogenase, the enzyme responsible for the first step of 4PHT biosynthesis, the oxidation of erythrose-4-phosphate to erythronate-4-phosphate. This enzyme is structurally and evolutionary related with the glycolytic enzyme glyceraldehyde-3-phosphate dehydrogenase (*gapA*), showing 41% of identity at amino acid level (Zhao et al., 1995). Glyceraldehyde-3-phosphate dehydrogenase catalyzes the oxidation of glyceraldehyde-3-phosphate to 1,3-diphosphoglycerate during glycolysis, an analogous reaction to the reaction catalyzed by erythrose-4-phosphate dehydrogenase, and it has been demonstrated that erythrose-4-phosphate dehydrogenase has residual activity glyceraldehyde-3-phosphate dehydrogenase , although not enough to allow cell growth in *gapA* mutant cells (Boschi-Muller et al., 1997). In contrast, deletion experiments with both *epd* and *gapA* genes have demonstrated that both are needed for PLP biosynthesis, and that *epd* gene is specially required for PLP biosynthesis under conditions of cell growth with non-glycolytic carbon sources like glycerol or succinate, but under cell growth with glycolytic carbon sources like glucose or fructose, *epd* deletion have no phenotypic effect, detecting an increase in GapA activity in comparison with control cells consequence of its double activity in both glycolysis and PLP biosynthesis (Yang et al., 1998b). Both *S. glossinidius* and *W. glossinidia* retain a functional *gapA* gene, which could be responsible of PLP biosynthesis, in concordance with their growth under glycolytic carbon sources in the context of their ecological niche inside the tsetse fly.

In contrast, *S. glossinidius* and *W. glossinidia* have no functional PLP salvage pathways. These pathways proceeds through the transport of PL, PN, and PM followed by their phosphorylation by two different multifunctional kinases encoded by the genes *pdxK* and *pdxY* (Yang et al., 1996; Yang et al., 1998a). In *S. glossinidius*, pseudogenization has affected to *pdxY* gene, whereas there is no signal of *pdxK* gene. *W. glossinidia* lacks both genes.

4.3.5.6.3 NAD and NADP biosynthesis

NAD and NADP are essential cofactors of bacterial cells, involved in hundreds of redox reactions. *S. glossinidius* retains functional *nad* genes (*nadB, C, D, E*) for NAD biosynthesis from L-aspartate, whereas in *W. glossinidia* there is no signal for *nadB* gene, which encodes an L-aspartate oxidase that catalyzes the first enzymatic step of the pathway, the oxidation of L-aspartate to iminoaspartate. Both *S. glossinidius* and *W. glossinidia* retain functional *nadK* gene encoding NAD kinase responsible for NAD phosphorylation to produce NADP.

NAD and NADP molecules, even if they are not consumed in oxidation reactions, have very short half-life, and are degraded to AMP and nicotinamide mononucleotide (NMN). NMN can be recycled to nicotinate mononucleotide (NAMN), an intermediate of "de-novo" pathway for NAD and NADP biosynthesis through three main recycling pathways that are also known as pyridine nucleotide cycles (Figure 4.16). However, none of these salvage pathways appears functional in *S. glossinidius* and *W. glossinidia*. In *W. glossinidia*, there is no signal for none of the genes involved in NAD salvage pathways, whereas *S. glossinidius* retains functional *pncB* gene encoding a nicotinate phosphoribosyltransferase for NMN biosynthesis from nicotinate and PRPP, but pseudogenization has affected to *nudC* gene, which encodes a NAD+ diphosphatase responsible of the degradation of NAD to AMP and NMN (Reed et al., 2003), *pncA* gene that encodes a nicotinamidase responsible for nicotinamide deamination to nicotinate, and *nadR* gene, which encodes a multifunctional protein that acts as transcriptional repressor of *nad* regulon (*nadA, nadB,* and *pnuB*) and has also ribosylnicotinamide kinase and NMN adenyltransferase activity for the transport and transformation of extracellular N-ribosylnicotinamide to NAD through NMN intermediate. NadR protein plays a pivotal role in the metabolism of NAD, acting as an allosteric protein that, in presence of high levels of NAD in the medium inhibits the genes involved in "de-novo" biosynthesis of NAD (*nadA* and *nadB*) and N-ribosylnicotinamide transport for NAD salvage pathway, whereas under low NAD concentrations binds to membrane transporter encoded by *pnuC* gene and induces both NAD "de-novo" biosynthesis by de-repression of *nadA* and *nadB* genes and NAD biosynthesis from salvage pathways through N-ribosylnicotinamide transport by PnuC followed by their phosphorylation to NMN by NadR-ribonucleotide kinase activity and NMN adenylation to NAD by NadR-adenyltransferase activity, allowing a rapid response to NAD depletion in bacterial cells (Foster et al., 1990; Zhu et al., 1991; Raffaelli et al., 1999; Kurnasov et al., 2002).

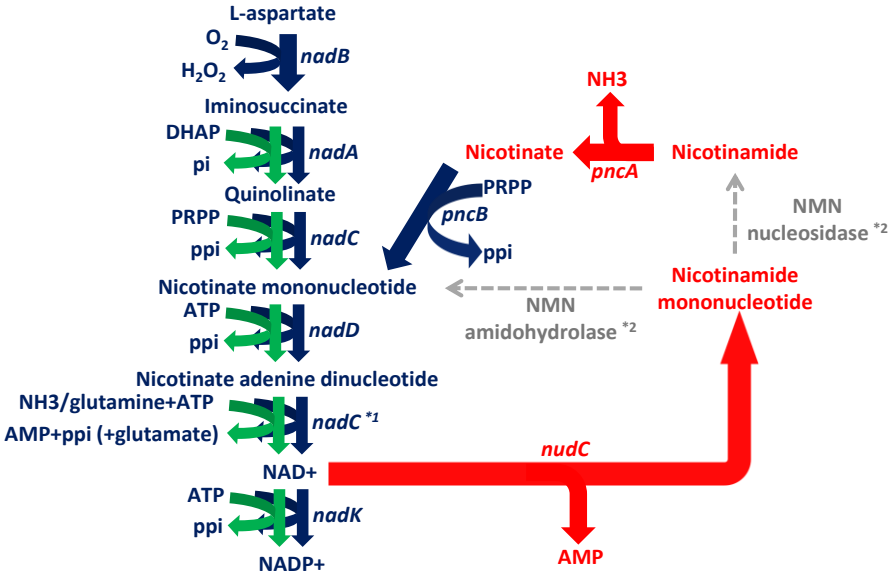


Figure 4.16: *Metabolic profile of NAD/NADP biosynthesis and recycling pathways in S. glossinidius and W. glossinidia. Blue coloured reactions represents S. glossinidius functional genes, whereas green coloured reactions represents W. glossinidia functional genes. Red coloured reactions represents pseudogenized enzymatic activities in S. glossinidius.*

*[*1] NAD synthetase encoded by the gene nadC catalyzes the transamination of nicotinate adenine dinucleotide to NAD+ by using NH3 or glutamine as amino group donor*

*[*2] Dashed arrows represents reactions without associated gene in database searches.*

The absence of salvage pathways in both bacterial endosymbionts, and specially the inactivation of *nadR* gene in *S. glossinidius*, may reflect a constitutive production of NAD and NADP through "de-novo" pathway that can be explained as a mechanism to provide this essential coenzyme for both the bacteria and the tsetse host.

4.3.5.6.4 Folate biosynthesis

Tetrahydrofolate, the biologically active form of folate, is the precursor of important cofactors involved in one-carbon unit transfer reactions, essential in pathways like methionine, purine, and pyrimidine biosynthesis, being also involved in the interconversion between serine and glycine and in histidine catabolism. Their "de-novo" biosynthesis starts with GTP that is converted to 6-hydroxymethyl-

dihydropterin diphosphate by the sequential action of enzymes encoded by the genes *folE, nudB, folB* and *folK*. In a second branch of the pathway, p-aminobenzoate is synthesized from chorismate by the sequential action of aminodeoxychorismate synthase enzyme complex (*pabA* and *pabB*) and aminodeoxychorismate synthase (*pabC*). P-aminobenzoate and 6-hydroxymethyl-dihydropterin diphosphate are condensed by dihydropteroate synthase (*folP*) to 7,8-dihydropteroate, which is converted to the final tetrahydrofolate by the sequential action of dihydrofolate synthetase (*folC*) and dihydrofolate reductase (*folA*). Dihydrofolate synthetase catalyzes also the addition of glutamate residues to tetrahydrofolate to produce folate polygutamates that prevents the efflux of folates outside the cells and increases the binding of folate cofactors to the enzymes of folate interconversions and biosynthesis (Figure 4.17).

Tetrahydrofolate is modified with different one-carbon units yielding different tetrahydrofolate derivatives that constitute the active cofactors. 5,10-methylene tetrahydrofolate, an essential cofactor involved in pantothenate and pyrimidine biosynthesis, can be produced by serine hydroxymethyltransferase (*glyA*), which catalyzes the transfer of methyl group from L-serine to tetrahydrofolate producing glycine and 5,10-methylene tetrahydrofolate, or by glycine cleavage system encoded by the genes *gcvT, H, P*, and *lpd* that catalyzes the reversible oxidation of glycine yielding carbon dioxide, ammonia, 5,10-methylene tetrahydrofolate, and a reduced pyridine nucleotide; 10-formyl-tetrahydrofolate, a tetrahydrofolate derivative involved in purine biosynthesis and formylation of initiator formyl-tRNA, is synthesized from 5,10-methylene tetrahydrofolate by bifunctional methylenetetrahydrofolate cyclohydrolase/dehydrogenase encoded by the gene *folD*, whereas 5-methyl-tetrahydrofolate, a tetrahydrofolate derivative involved in L-methionine biosynthesis, is synthesized by reduction of 5,10-methylene tetrahydrofolate catalyzed by 5,10-methylene tetrahydrofolate reductase encoded by the gene *metF*.

Finally, tetrahydrofolate can be also recycled by deformylation of 10-formyl-tetrahydrofolate, which can be catalyzed by two different deformylases, a specific formyltetrahydrofolate deformylase encoded by the gene *purU* and a phosphoribosylglycinamide formyltransferase encoded by the gene *purN* that is involved in purine biosynthesis catalyzing the transfer of formyl group of 10-formyl-tetrahydrofolate to the purine biosynthesis intermediate 5-phosphoribosylglycineamine (GAR) yielding tetrahydrofolate and the purine precursor 5'-phosphoribosyl-N-formylglycineamine

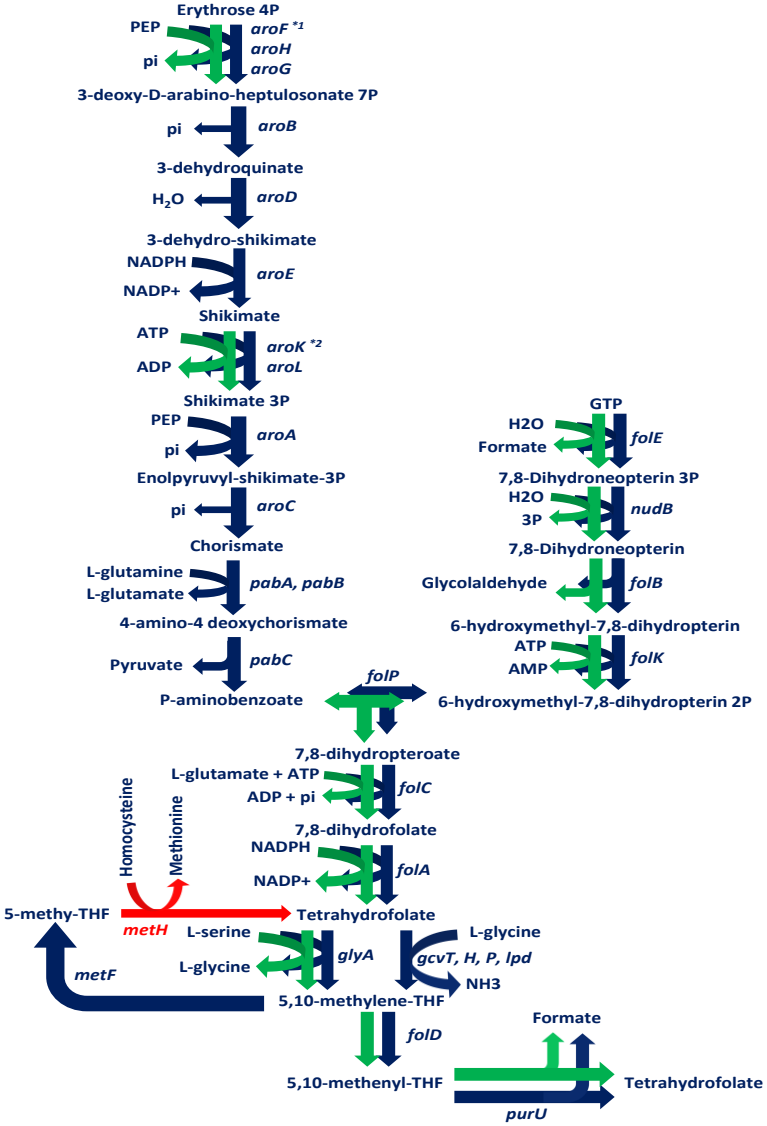


Figure 4.17: *Metabolic profile of tetrahydrofolate biosynthesis and interconversions in S. glossinidius and W. glossinidia. Blue coloured reactions represents S. glossinidius functional genes, whereas green coloured reactions represents W. glossinidia functional genes. Red coloured reactions represents pseudogenized enzymatic activities in S. glossinidius.*

**1: S. glossinidius retains functional genes for the three isozymes corresponding to 3-deoxy-7-phosphoheptulonate synthase, whereas W. glossinidia retains a single aroF gene.*

**2: S. glossinidius retains functional genes for the two isozymes corresponding to shikimate kinase, whereas W. glossinidia retaisn a single aroK gene*

S. glossinidius retains all functional genes for tetrahydrofolate biosynthesis from GTP and chorismate, as well as functional *folD* and *purU* genes for tetrahydrofolate recycling from 5,10-methylene tetrahydrofolate, with pseudogenization affecting to *purN* gene. However, *purN* pseudogenization does not affect purine biosynthesis because it encodes the minor transformylase isozyme, and *S. glossinidius* retains a functional *purT* gene encoding the major GAR transformylase isozyme. In addition, *S. glossinidius* also retains functional glycine cleavage system (*gcvT, H, P*, and *lpd* genes) and functional serine hydroxymethyltransferase (*glyA* gene) for 5,10-methylenetetrahydrofolate biosynthesis, as well as functional 5,10-methylenetetrahydrofolate reductase (*metF*) for 5,10-methylenetetrahydrofolate reduction to 5-methyltetrahydrofolate in L-methionine biosynthesis. However, pseudogenization has affected to *metH* gene that encodes a cobalamin-dependent homocysteine transmethylase that catalyzes the transfer of methyl group of 5-methyltetrahydrofolate to L-homocysteine to produce L-methionine, retaining functional *metE* gene that encodes a second cobalamin-independent isozyme that uses 5-methyltetrahydropteroyltri-L-glutamate as methyl-group donor (Figure 4.17).

In contrast *W. glossinidia* lacks *pabA, pabB* and *pabC* for p-aminobenzoate biosynthesis from chorismate, retaining functional genes for the rest of enzymes of the pathway for tetrahydrofolate biosynthesis from GTP. This is in concordance with their inability for endogenous biosynthesis of chorismate from the pentose-phosphate pathway intermediate erythrose 4-phosphate through the shikimate pathway (Figure 4.17). Chorismate is a common precursor of aromatic aminoacid biosynthesis and isoprenoid biosynthesis, which are also absent in *W. glossinidia*. Exogenous chorismate cannot be transported inside the cells and the only way to synthesize it is from erythrose 4P through shikimate pathway. In contrast, p-aminobenzoate can be freely transport inside the cells and exogenous p-aminobenzoate can support the growth on strains unable to synthesize it endogenously (Green et al., 1992; Green et al., 1996), so is it possible that p-aminobenzoate synthesized by *S. glossinidius* could be exported and captured by *W. glossinidia* for their endogenous biosynthesis of tetrahydrofolate. In fact, mutants of *pabC* in *E. coli* needs the external supply of PABA for their survival (Green et al., 1992). For the rest of enzymatic activities involved in the biosynthesis of one-carbon derivatives of tetrahydrofolate, *W. glossinidia* retains functional *glyA, purU* and *folD* genes, with no signal of *gcv* genes of glycine cleavage system retaining only *lpd* gene that is functional in the context of pyruvate dehydrogenase complex, and no signal for *metF* gene, in concordance with the absence of genes for L-methionine biosynthesis. Both *S. glossinidius* and *W. glossinidia* also retain a functional *thyA* gene that encodes a thymidilate synthase enzyme involved in "de-novo" pyrimidine biosynthesis that catalyzes the methylation of dUTP to dTMP using 5,10-methylenetetrahydrofolate as methyl-group donor, generating the folate precursor 7,8-dihydrofolate that can be converted to tetrahydrofolate by functional folate reductase encoded by the gene *folA* in both bacterial endosymbionts.

4.3.5.6.5 Pantothenate and Coenzyme A biosynthesis

Pantothenate is the precursor of Coenzyme A (CoA) and acyl carrier protein (ACP), essential cofactors that participate in over 100 different reactions in the intermediary metabolism. CoA is the precursor of different CoA thioesters like acetyl Coenzyme A (AcCoA), succinyl CoA (SucCoA) or malonyl CoA that are critical in the metabolization of different carbon sources and in the biosynthesis of fatty acids and membrane phospholipids. The composition of the CoA thioester pool is dependent of the carbon source under the cells are growing on, and under glucose as carbon source, AcCoA is the predominant thioester.

Pantothenate is synthesized from the branched-chain amino acid precursor 2-ketoisovalerate, which is first converted to L-pantoate in two enzymatic steps catalyzed by 3-methyl-2-oxobutanoate hydroxymethyltransferase (*panB*) and 2-dehydropantoate -2-reductase (*panE*). L-pantoate is condensed with β-alanine by pantothenate sythetase (*panC*) to produce pantothenate, and CoA is synthesized from pantothenate in 5 enzymatic steps (*coaA, dfp, coaD* and *coaE* genes). ACP is synthesized by the transfer of the 4-phosphopantotheine moiety of CoA to the precursor apo-ACP (encoded by the gene *acpP*) in a reaction catalyzed by the enzyme holo-ACP-synthase (*acpS*) (Figure 4.18).

S. glossinidius retains all functional genes for pantothenate, CoA, and ACP biosynthesis from 2-ketoisovalerate, which can be produced from both two molecules of pyruvate in the context of branched-chain amino acid biosynthesis and from L-valine and pyruvate by valine-pyruvate aminotransferase (*avtA*) as by-product of L-alanine biosynthesis. In addition, retains functional *panD* gene encoding aspartate-1-decarboxylase for β-alanine biosynthesis by decarboxylation of L-aspartate. In contrast, *W. glossinidia* retains almost all genes for pantothenate biosynthesis although is unable to synthesize 2-ketoisovalerate, in concordance with the absence of genes for L-valine and branched-chain amino acids biosynthesis

In addition, *W. glossinidia* also lacks *panD* gene for β-alanine biosynthesis from L-aspartate (Figure 4.18), pointing out to another possible metabolic complementation between *S. glossinidius* and *W. glossinidia* at pantothenate biosynthesis level. Inactivation studies have revealed that *panD* mutants generate an absolute requirement of β-alanine or pantothenate for growth in *E. coli* and *Salmonella thipimurium* (Cronan, Jr. et al., 1982; Kennedy and Kealey, 2004). In addition, *E. coli* mutants with none of the genes involved in 2-ketoisovalerate biosynthesis have an absolute requirement of branched-chain amino acids and pantothenate for growth (Whalen and Berg, 1982), so the metabolic profile of *W. glossinidia* points out to the necessity of exogenous supply of branched chain amino acids and pantothenate for their survival, that in the case of pantothenate can be supplied by *S. glossinidius*. In fact, it has been documented that over 90% of

pantothenate synthesized in *E. coli* is excreted rather than utilized for CoA biosynthesis (Jackowski and Rock, 1981), so it is possible that pantothenate could be secreted by *S. glossinidius* and incorporated by *W. glossinidia* for CoA biosynthesis.

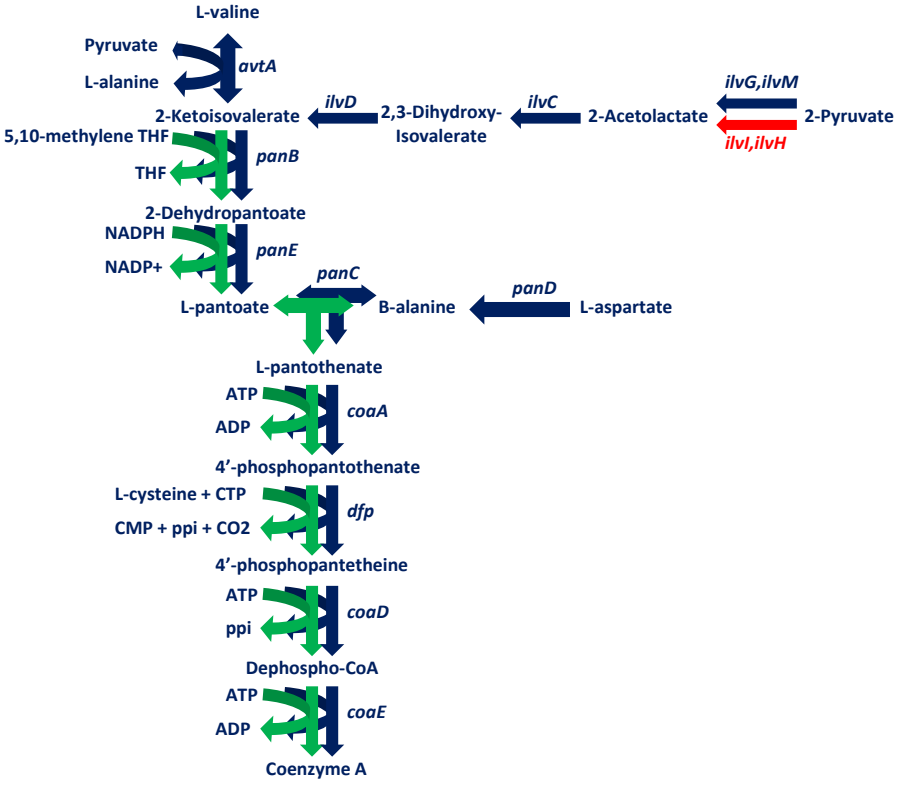


Figure 4.18: *Metabolic profile of Coenzyme A biosynthesis in S. glossinidius and W. glossinidia. Blue coloured reactions represents S. glossinidius functional genes, whereas green coloured reactions represents W. glossinidia functional genes. Red coloured reactions represents pseudogenized enzymatic activities in S. glossinidius*

4.3.5.6.6 Biotin biosynthesis

Biotin is an essential cofactor for carboxyl-group transfer enzymes like AcCoA carboxylase enzyme complex, essential in fatty acid biosynthesis. Biotin is synthesized from pimeloyl-CoA and L-alanine in four enzymatic steps catalyzed by enzymes encoded by the genes *bioF*, *A*, *D* and *B* that are organized in an operon together with *bioC* and *bioH* genes, which encode a predicted methyltransferase and

a carboxyesterase respectively involved in pimeloyl-CoA biosynthesis (DeMoll, 1996). The complete operon is conserved completely functional in *S. glossinidius* and *W. glossinidia*.

4.3.5.6.7 Lipoic acid biosynthesis

Lipoic acid or lipoate is a sulfur-containing cofactor essential for the functionality of several enzyme complexes involved in oxidative decarboxylations like pyruvate dehydrogenase or the glycine cleavage system (DeMoll, 1996). Both *S. glossinidius* and *W. glossinidia* retain functional *lipA* and *lipB* for lipoate biosynthesis from the fatty acid precursor octanoyl-ACP.

4.3.5.6.8 Phorphirin biosynthesis

Porphyrins, also known as tetrapyrrole family of cofactors, are metal-binding cofactors derived all from the common precursor uroporphirinogen III that is synthesized from activated L-glutamyl-tRNAglu in 5 enzymatic steps encoded by the genes *hemA, I, B, C, D*, all of which are functional in both *S. glossinidius* and *W. glossinidia*. Urophorphirinogen III is the common precursor of siroheme, an iron-containing tetrapyrrole that acts as prosthetic group of sulfite and nitrite reductases involved in the conversion of highly oxidized forms or nitrogen and sulfur (nitrite and sulfite) to their fully reduced forms used in biosynthetic reactions (ammonia and sulfide), and different heme groups that are essential for the functionality of cytochrome C oxidases of the respiratory chain (Beale, 1996). Siroheme is synthesized from Urophorphyrinogen III in four enzymatic steps catalyzed by a single multifunctional urophorphirin III C-methyltransferase encoded by the gene *cysG*. The different heme groups are synthesized from the common precursor protoheme IX, which is synthesized in 4 enzymatic steps from urophorphyrinogen III (*hemE, N, F, G, H*). From protoheme IX, heme O group is synthesized by condensation with the isoprenoid farnesyl diphosphate catalyzed by heme O synthase (*cyoE*), whereas heme D group is synthesized by hydroxylation of protoheme IX catalyzed by catalase HPII (*katE*).

S. glossinidius retains all functional genes for heme and siroheme biosynthesis, with pseudogenization affecting only to *hemN* gene that encodes the anaerobic isozyme of coprophorphirinogen III oxidase but retaining functional *hemF* gene that encodes the aerobic isozyme. *W. glossinidia* retains functional genes for heme group biosynthesis from urophorphirinogen III but there is no signal of *cysG* gene for siroheme biosynthesis.

4.3.5.6.9 Molybdenum cofactor biosynthesis

Molybdenum-dependent enzymes catalyze important redox reactions in the global carbon, sulfur, and nitrogen cycles mainly used during anaerobic growth such

as dissimilatory nitrate reductase, formate dehydrogenase, or dimethyl-sulfoxide reductase, with the corresponding substrates acting as terminal electron acceptors in anaerobic respiration (Schwarz, 2005). *S. glossinidius* shows pseudogenization for most of the genes involved in molybdenum cofactor biosynthesis (*moaA, D, E, mobA, B, moeB, mogA*), retaining only functional *moaC* and *moeA* genes, in concordance with the inactivation or absence of most enzymatic activities of anaerobic metabolism like genes encoding dimethylsulfoxide reductase for anaerobic respiration or genes involved in menaquinone and dimethylmenaquinone biosynthesis. In *W. glossinidia* there is no signal for any of the genes of molybdenum cofactor biosynthesis.

4.3.5.6.10 Thiamin biosynthesis

Thiamin biosynthesis pathway is the unique biosynthetic pathway that is inactivated in *S. glossinidius* but has an apparent complete biosynthetic pathway in *W. glossinidia*, which means that not only tsetse flies but also *S. glossinidius* is dependent of *W. glossinidia* for its synthesis. However, a detailed analysis of the thiamin biosynthetic pathway in *W. glossinidia*, although leads to the detection of some previously no annotated genes, also revealed that the pathway was incomplete, lacking the essential *thiI* gene and showing *thiF* as a pseudogene. The outline of the thiamin biosynthesis pathway in both tsetse symbionts is represented in Figure 4.19A. The synthesis of the biologically active form of the coenzyme thiamine diphosphate in bacteria takes place by several pathways. The *de novo* pathway requires the synthesis of two intermediates, a pyrimidine phosphate moiety and a thiazole phosphate moiety. Both are combined by the action of a thiamine phosphate synthase (ThiE) and a thiamine phosphate kinase (ThiL) to produce the active cofactor. (Vander Horn et al., 1993; Begley et al., 1999; Leonardi et al., 2003; Lehmann et al., 2006; Jurgenson et al., 2009).

The analysis of thiamin biosynthesis pathway in *S. glossinidius* showed that the ancestor of this bacterium probably contained the complete set of genes for its synthesis, but as consequence of the massive gene inactivation process the complete pathway for "de-novo" thiamin biosynthesis is inactive, identifying 4 pseudogenes (*thiS, thiF, thiG* and *thiE*) together with an originally annotated gene (*thiH*) that presents a premature stop codon detecting the final part of the ancestral gene as an adjacent pseudogene, probably revealing an ongoing gene inactivation event (Figure 4.19A). For the rest of thiamin biosynthesis genes, there is no signal of *thiD* gene nor as gene or pseudogene, although retains functional *dxs, thiL, thiI* and *iscS* genes, probably due to their essential role in other metabolic reactions, like cysteine desulphurase encoded by the gene *iscS*, which is the main responsible of tRNA modification by nucleoside sulfur addition, or cysteine sulfur transferase encoded by the gene *thiI* that is involved in one of this modifications, the transfer of the cysteine sulfur group to uridine in tRNA molecules to produce thiouridine (Kambampati and

Lauhon, 2000; Lauhon, 2002). The most plausible explanation for the presence of the cofactor in *S. glossinidius* is a salvage pathway through the presence of a functional thiamin ABC transport system encoded by the genes *thiP*, *thiQ* and *tbpA* together with a functional thiamine kinase (*thiK*) and thiamine diphosphate kinase (*thiL*). However, it implies that sufficient amounts of non-phosphorylated thiamine should be available from the host. In addition, *S. glossinidius* retains remnant activities and proteins related with the pathway, including the ability to synthesize thiazole phosphate carboxylate moiety through a salvage pathway (Figure 4.19A).

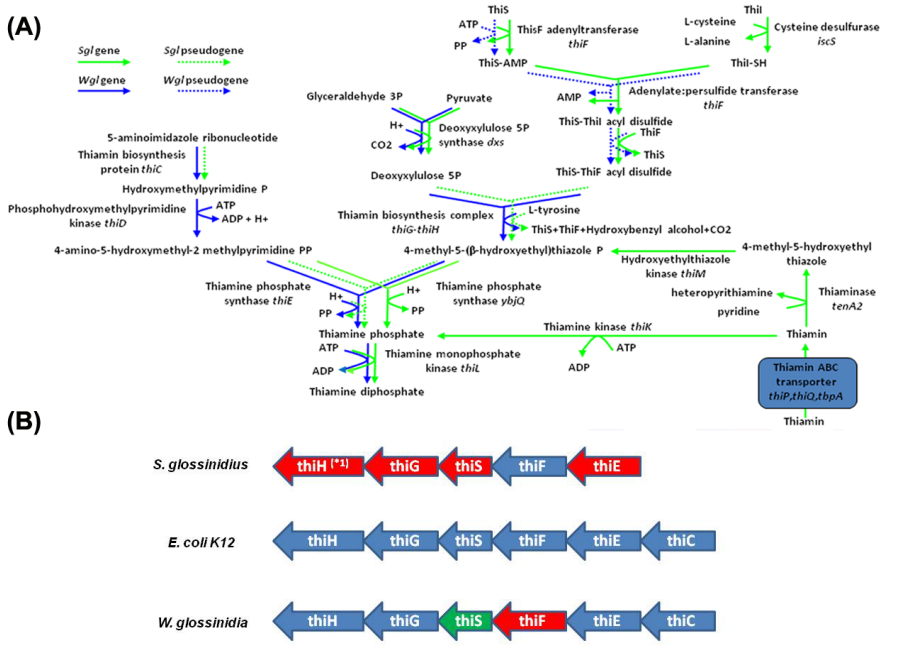


Figure 4.19: *Metabolic profile of thiamine biosynthesis pathway in S. glossinidius and W. glossinidia.*

(A) *Schematic diagram of thiamine biosynthesis pathway. Reactions coloured in blue represents the functional profile of thiamine biosynthesis in W. glossinidia, whereas reactions coloured in green represents the functional profile of thiamine biosynthesis in S. glossinidius; Pseudogenized activities are represented by dashed lines. Functional thiamine ABC transport system, thiamine kinase (thiK) and thiamine phosphate kinase (thiL) allow thiamine diphosphate biosynthesis from exogenous thiamine.*

(B) *Structure of the thiamine biosynthesis operon in S. glossinidius, E. coli K12, and W. glossinidia. In blue are represented functional genes, in red pseudogenes, and in green the thiS gene characterized in W. glossinidia in this study. (*1) thiH in S. glossinidius corresponds to an originally annotated gene (SG0136) with a premature stop codon detected during the reannotation process.*

The analysis of thiamin metabolism in *W. glossinidia* revealed that it retains functional *thiH*, *thiG*, *thiE*, and *thiC* genes, with no signal of *thiI* gene, and with *thiS* and *thiF* genes that appears non-annotated but present in the intergenic region between *thiG* and *thiE* (Figure 4.19B). TBLASTN analysis with ThiF and ThiS proteins from *E. coli* K12 against the complete genome of *W. glossinidia* confirmed the presence of unnannotated *thiS* and *thiF* genes immediately downstream of *thiG* gene. However, whereas *thiS* gene has no stops nor frameshifts, the putative *thiF* gene of *W. glossinidia* contains an internal stop codon that disrupts the translation of the putative ThiF protein at amino acid 165, before the essential cysteine residue at position 184 that is responsible of the disulfide linkage between *ThiS* and *ThiF* that acts as sulfur donor in thiazole phosphate moiety biosynthesis (Xi et al., 2001). The absence of these two genes limits the metabolic capability of *W. glossinidia* to the synthesis of the pyrimidine moiety hydroxymethylpyrimidine pyrophosphate (HMP-PP in Figure4.20).

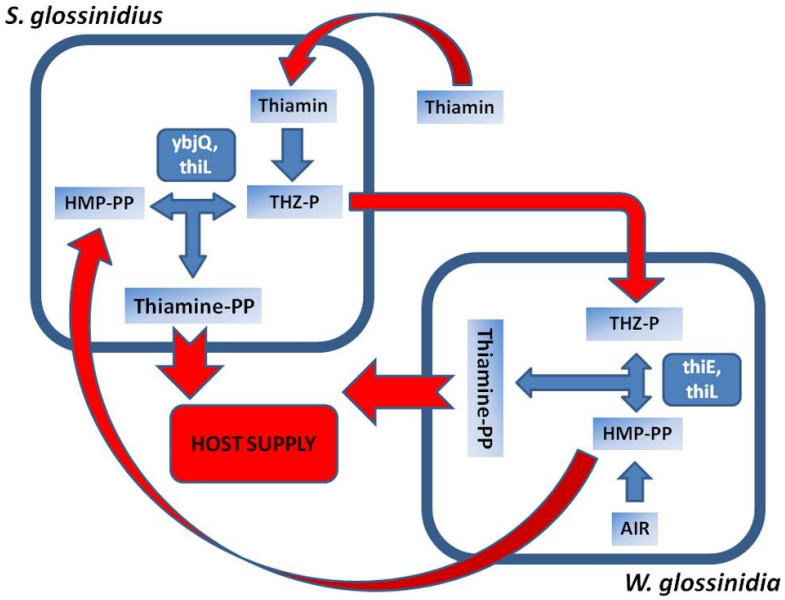


Figure 4.20: *Metabolic complementation between W. glossinidia and S. glossinidius at thiamine biosynthesis level. Thiazole phosphate carboxylate (THZ-P) is synthesized by S. glossinidius from exogenous thiamin through salvage pathway (tenA2, thiM), whereas hydroxymethyl pyrimidine pyrophosphate (HMP-PP) is synthesized by W. glossinidia from 5-aminoimidazole ribonucleotide (AIR) (thiC, thiD). THZ-P and HMP-PP are shared between both bacteria to produce the functional thiamine diphosphate that is provided to the tsetse host.*

Because the metabolic pathways of both endosymbionts were incomplete, we searched for a possible metabolic complementation. This was revealed through the identification of *ybjQ* gene in *S. glossinidius*. The YbjQ protein has been recently demonstrated to contain a thiamin phosphate synthase activity able to rescue thiamin auxotrophy in a mutant *thiE* strain (Morett et al., 2008). This leads to a scenario in which *W. glossinidia* synthesizes HMP-PP moiety and *S. glossinidius* THZ-P. Both intermediates may be shared by the endosymbionts, with *W. glossinidia* synthesizing the active cofactor after the action of ThiE and ThiL and *S. glossinidius* after the action of YjbQ and ThiL (Figure 4.20).

4.3.5.7 Cell envelope metabolism

S. glossinidius retains all functional genes needed for the biosynthesis of all components of a typical gram-negative cell envelope, including the complete set of genes for fatty acid biosynthesis from AcCoA, peptidoglycan biosynthesis and assembly, bacterial lipopolysaccharide (LPS) biosynthesis, and genes encoding different porines and transport systems for solute exchange across membranes. In the next sections I will briefly describe the functionality of the different biosynthetic pathways of cell envelope structures presents in *S. glossinidius*.

4.3.5.7.1 Peptidoglycan biosynthesis

The essential peptidoglycan sacculus, also known as murein sacculus or murein layer, is located in the periplasmic space of gram-negative bacteria forming a thin layer adjacent to the periplasmic side of the cytoplasmic membrane that is responsible of the rigidity of bacterial cell walls and that allows cell protection from rupture (Cooper, 1991; Holtje, 1998). Peptidoglycan layer is composed by chains that contain repeated subunits of a muropeptide known as peptidoglycan unit. This peptidoglycan unit is composed by aminosugars N-acetylglucosamine (GlcNAc) and N-acetylmuramic acid (MurNAc) and aminoacids D-alanine, L-alanine, 2,3 diaminopimelate, and D-glutamate. Peptidoglycan layer is a network of chains of this muropeptide, with the aminosugars forming linear strands of alternating GlcNAc and MurNAc that are crosslinked by short peptides composed of L-alanine-D-glutamate-L-2,3-diaminopimelate-D-alanine(Vollmer and Bertsche, 2008; Gan et al., 2008). A schematic representation of the whole peptidoglycan biosynthesis pathway is included in Figure 4.21.

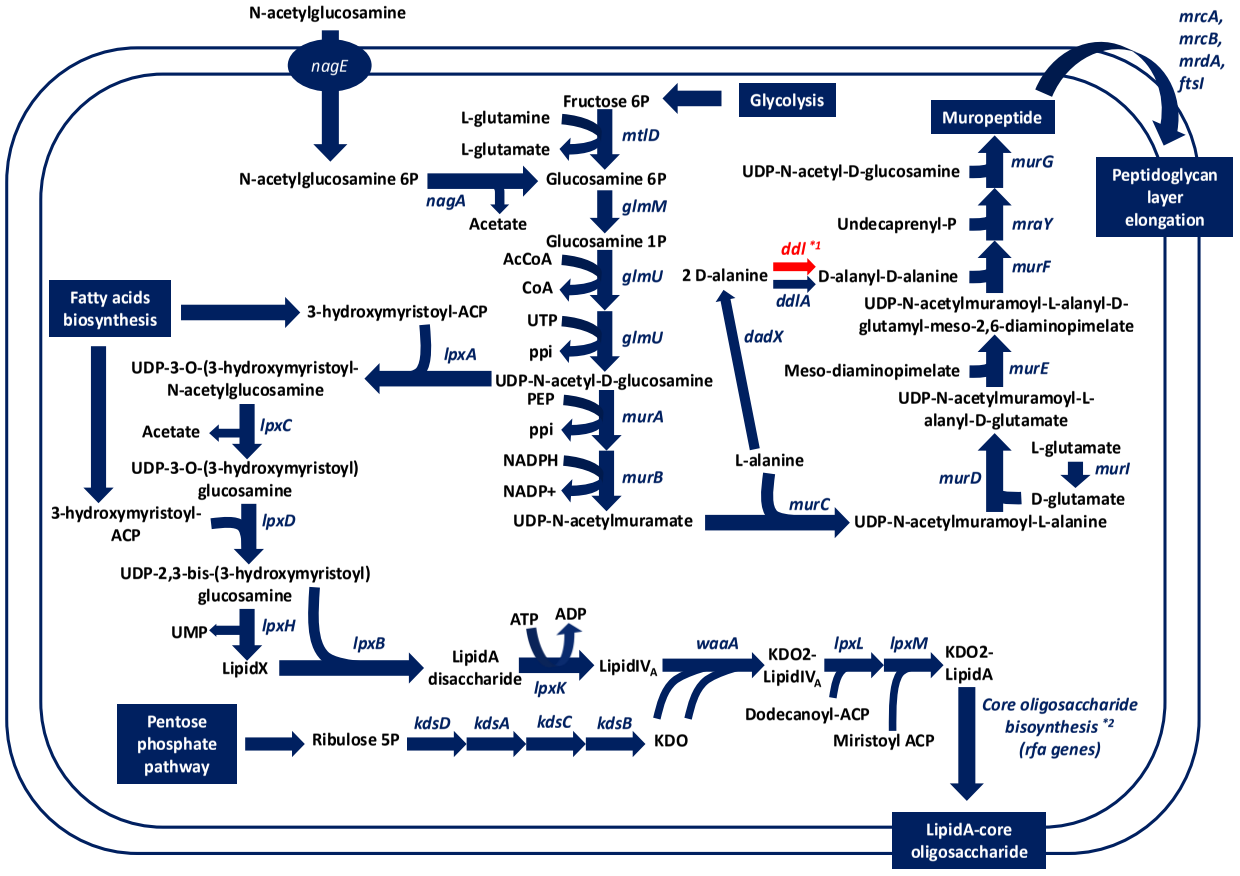


Figure 4.21: *Metabolic profile of cell envelope biosynthesis in S. glossinidius. Red coloured enzymatic reactions represent pseudogenization events, whereas blue coloured ones are those encoded by functional genes.* *[*1] S. glossinidius retains a functional gene (SG0899) and a pseudogene ps_SGL0959 with homology with D-alanine-D-alanine ligase isoform A.* *[*2] Core oligosaccharide of bacterial LPS is synthesized by addition of different sugars to the KDO moiety of LipidA by glycosyltransferases encoded by rfa genes (SG2193-SG2202)*

S. glossinidius retains all functional genes for synthesis and assembly of peptidoglycan layer, with functional *glmM* and *glmU* genes for the biosynthesis of murein precursor UDP-N-acetyl-D-glucosamine from glucosamine-6-phosphate (GlcN6P) and functional *mrvA*, *ddlA*, and *mur* genes (*murA, B, C, D, I, E, F*) for muropeptide assembly by addition of the different amino acids, the isoprenoid undecaprenyl-phosphate, and a second molecule of UDP-N-acetyl-D-glucosamine to UDP-N-acetyl-D-glucosamine (Figure 4.21). In addition, *S. glossinidius* also retains functional genes for the main murein synthases PBP1A (*mrcA*), PBP1B (*mcrB*), PBP2 (*mrdA*) and PBP3 (*ftsI*) responsible for peptidoglycan elongation by addition of murein monomers to the peptidoglycan chains; this is carried out through transglycolation reactions between carbon 1 of the MurNAc residue of the growing chain and carbon 4 of GlcNAc residue of the new muropeptide monomer and transpeptidase reactions linking murein chains through D-alanine residues of one chain and diaminopimelate residue of their adjacent chain (Vollmer and Bertsche, 2008). PBP2 is an essential enzyme for cell elongation, responsible of the maintenance of rod-shape form in *E. coli* K12, whereas PBP3 is essential also for cell division because is responsible of septum formation (van Heijenoort, 1996). Finally, *S. glossinidius* also retains functional *amiA* and *amiB* genes that encode two different N-acetylmuramyl-L-alanine amidases, a group of murein hydrolases involved in the split of the murein septum between cells during cell division catalyzing the removal of the peptide moiety from glycan chains.

4.3.5.7.2 Bacterial lipopolysaccharide biosynthesis

Bacterial lipopolysaccharide (LPS) is a unique constituent of the outer membrane of gram-negative bacteria that contributes to structural integrity of bacterial cells stabilizing the overall membrane structure and protecting the membrane from chemical attacks (Raetz, 1996). *S. glossinidius* retains all functional genes for lipidA biosynthesis from the peptidoglycan precursor UDP-N-acetyl-D-glucosamine (*lpxA, C, D, H, B, K*) together with functional *kds* genes (*kdsD, A, C, B*) for the biosynthesis of 3-deoxy-D-manno-octosulonic acid (KDO) from the pentose phosphate pathway intermediate ribulose-5-phosphate and *waaA* gene encoding KDO transferase for the addition of two KDO monomers to lipidA (Figure 4.21). LipidA, also known as endotoxin, constitutes the hydrophobic anchor of bacterial LPS to the outer membrane, being also responsible of pathogenicity in some Gram negative infections, and the two KDO subunits of lipidA serves as linking point with the core oligosaccharide component of bacterial LPS. Core oligosaccharide is synthesized by the sequential addition of different sugars to the KDO subunits of lipidA. Their composition is heterogeneous, with different sugars that are added in different bacterial strains by glycosyltransferases encoded by the genes *rfa*, which differ between species depending on core oligosaccharide composition. In *E. coli* K12, the core oligosaccharide is composed by the sequential addition of ADP-L-glycero-D-manno-heptose, UDP-glucose, and UDP-D-galactose

catalyzed by 14 *rfa* genes organized in two oppositely transcribed operons, of which only *rfaD*, *rfaF* and *rfaC* are present functional in *S. glossinidius*. The rest of *rfa* genes of *E. coli* K12 have no homology with *rfa* genes of *S. glossinidius*, that have the highest similarities with *rfa* genes from *Serratia proteomaculans* 568 in BLASTP and FASTA searches (Figure 4.22A). *rfaC* and *rfaF* encodes two different heptosyltransferases that catalyze the sequential addition of two molecules of ADP-L-glycero-manno-heptose to the KDO moiety of lipidA, whereas *rfaD* encodes a epimerase that catalyzes the final step in ADP-L-glycero-manno-heptose biosynthesis from the pentose phosphate pathway intermediate D-seudoheptulose-7-phosphate, that is also completely functional in *S. glossinidius* (*lpxA, rfaE, gmhB* and *rfaD*). The rest of *rfa* genes specific of *E. coli K12* encodes glycosyltransferases that catalyzes the addition of UDP-D-glucose and UDP-D-galactose to the growing core oligosaccharide. These two nucleotide sugars are synthesized by interconversion of galactose and glucose through the Leloir pathway that requires the three enzymes galactokinase (*galK*), galactose-1-phosphate uridyltransferase (*galT*) and UDP-4-galactose 4-epimerase (*galE*) (Frey, 1996). *S. glossinidius* retains functional genes for all enzymes of the pathway, including functional *mglA, B, C* genes encoding a D-galactose ABC transport system, but *galK* and *galT* are situations of putatively longer ancestral genes where the absent part of the gene is detected as adjacent pseudogene, reflecting a probable ongoing pseudogenization process (Figure 4.22B). The inactivation of *galK* and *galT* genes in *S. glossinidius* may disrupt the production of UDP-D-galactose and UDP-D-glucose, and this can explain the absence of *rfa* homologs encoding their corresponding glycosyltransferases for their incorporation to the core oligosaccharide, with *S. glossinidius* that would have different glycosyltransferases indicative of different core oligosaccharide composition compared with *E. coli* K12.

Finally, *S. glossinidius* has no *rfb* genes responsible of O-antigen biosynthesis, the most external and variable component of bacterial LPS responsible of antigenic variation, a fact that has been related with the absence of immune response against *S. glossinidius* in the tsetse fly (Toh et al., 2006).

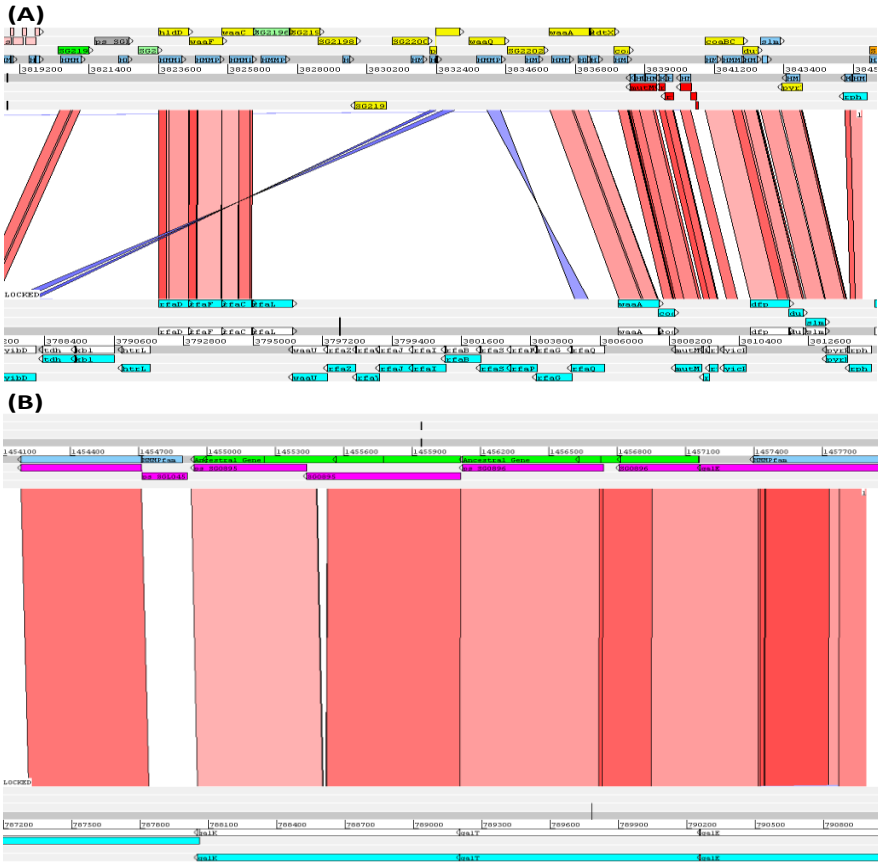


Figure 4.22: *ACT comparisons between S. glossinidius (top) and E. coli K12 (bottom) for the genome regions corresponding to rfa genes for core oligosaccharide biosynthesis (A) and galK and galT genes involved in UDP-glucose and UDP-galactose biosynthesis (B). In (B), SG0895 and SG0896 corresponds to the originally annotated galK and galT genes, whereas adjacent ps_SG0895 and ps_SG0896 corresponds to the corresponding pseudogenes characterized during re-annotation process. The whole ancestral gene are represented in green boxes.*

4.3.5.7.3 Extracellular polysaccharides biosynthesis

In addition to bacterial LPS, many bacteria produce extracellular polysaccharides (EPS) that can remain attached to the cell surface in a capsular form or alternatively can be released outside the cell. These EPS can be serotype-specific like the O-antigen subunit of bacterial LPS and the K-antigen capsular polysaccharide, which constitutes prominent antigens structurally and serologically diverse, or non-specific

EPS like enterobacterial common antigen (ECA) and colonic acid (CA) (Frey, 1996; Whitfield, 2006).

CA is composed by four different nucleotide sugars (UDP-D-glucose, UDP-D-galactose, UDP-D-glucuronate, and GDP-L-fucose) that constitute the CA repeat unit or E-unit, synthesized in the cytoplasmic side of the inner membrane and exported to the periplasmic space where it polymerizes with other E-units and nucleotide sugars to produce the final CA molecule. *S. glossinidius* retains functional *ugd* gene for UDP-D-glucuronate biosynthesis from UDP-D-glucose and *manB*, *manC*, *gmd*, and *fcl* genes for GDP-L-fucose biosynthesis from mannose-6-phosphate, although the Leloir pathway for UDP-D-glucose and UDP-D-galactose biosynthesis appears disrupted by the possible inactivation of *galK* and *galT* genes (see above). *S. glossinidius* also retains functional *wza*, *wzb*, and *wzc* genes that encodes subunits of the capsular polysaccharide export apparatus for EPS translocation across the outer membrane, that forms a cluster with seven genes encoding glycosyltransferases involved in nucleotide sugar addition to EPS, with only three of the seven genes with orthologs in *E. coli* K12 (*wcaJ, L, K*). The presence of functional genes for most of nucleotide sugar components of CA together with functional export apparatus for EPS translocation and functional glycosyltransferases different from *wca* genes in *E. coli* K12 indicates variability at CA molecules between *S. glossinidius* and *E. coli* K12, in concordance with the variability also observed at bacterial LPS core oligosaccharide level.

However, despite differences observed at bacterial LPS core oligosaccharide and CA between *S. glossinidius* and *E.coli* K12, there is clear homology at enterobacterial common antigen level (ECA), an outer membrane EPS common to all members of enterobacteria, with *S. glossinidius* having all genes involved in ECA biosynthesis completely functional (*rffM, wzyE, rffT, wzxE, rffA, C, H, G, D, E, wzzE* and *rfe*) and orthologous with *E. coli* K12.

4.3.5.7.4 Outer membrane structures

S. glossinidius retains functional genes for the major channel-forming proteins that allow the exchange of nutrients and other compounds across the outer membrane (Figure 4.23). This includes functional *ompA*, *ompF* and *ompC* genes encoding the major porin proteins (non-specific diffusion channels), functional TonB and Tol-Pal energy transducing systems for the energy-dependent transport of large solutes like siderophores or vitamin B12, and functional *tolC* gene encoding an outer membrane porin involved in the efflux of different hydrophobic and amphipatic molecules, being essential for the functioning of different multidrug efflux systems responsibles of the exclusion of liphophilic compounds (bile salts, detergents, and fatty acids) outside the cells (Braun, 1995; Lloubes et al., 2001). TonB system is composed by *tonB* gene that encodes a cytoplasmic membrane

protein that transduces the energy from the proton motive force of respiratory chain to outer membrane transporters, and *exbB* and *exbD* genes, which encodes proteins involved in energy transduction from proton motive force to the TonB protein, whereas Tol-Pal system structure is analogous to the TonB-energy transducing system. Of the different multidrug-efflux systems, *S. glossinidius* retains functional arcAB system (*arcA, arcB*) involved in aminoglycosides exclusion (Fralick, 1996; Rosenberg et al., 2000) and functional MacAB system (*macA, macB*) involved in macrolide resistance via active drug efflux (Kobayashi et al., 2001), with pseudogenization having affected *mdtB* and *mdtC* genes of the MdtABC efflux system involved in bile salts exclusion (Nagakubo et al., 2002) and to *emrA* gene of the EmrAB multidrug efflux system.

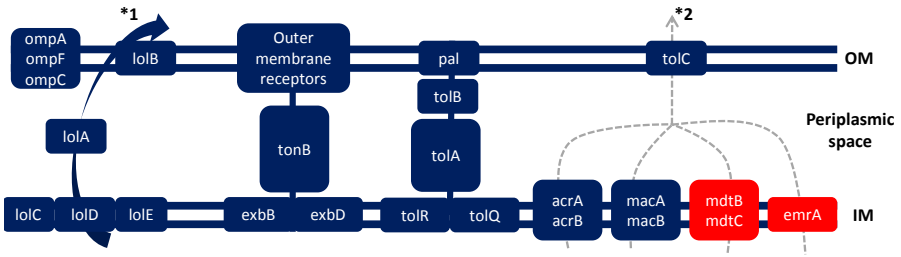


Figure 4.23: *Schematic representation of cell envelop structures in S. glossinidius. Red coloured structures represents pseudogenization events, whereas blue coloured ones are those encoded by functional genes.*
**1: Lol system for periplasmic transport and outer membrane insertion of lipoproteins. ABC transport system encoded by lolC, D, E genes catalyzes the release of lipoproteins to the periplasmic space, where periplasmic chaperone encoded by lolA directs lipoproteins to the outer membrane receptor encoded by the gene lolB responsible of the final location.*
**2 Different mutidrug-efflux systems in inner membrane (IM) shares common outer membrane porin encoded by the gene tolC for the exclusion of different lipophilic compounds*

From a structural point of view, *S. glossinidius* retains functional *lpp* gene encoding the major murein lipoprotein, one of the most abundant proteins in bacterial cells involved in the maintenance of the integrity and stability of bacterial cell envelope, being essential for cell survival. This and other lipoproteins located in the outer membrane are synthesized in the cytoplasm as prolipoprotein precursors with a consensus sequence called lipobox or lipoprotein box around the signal cleavage site, and have to be translocated to the periplasm through different secretion pathways (SEC, TAT, or SRP translocation systems), where they have to be inserted into the outer membrane. *S. glossinidius* retains functional *lgt* gene encoding an essential membrane protein with phosphatidylglycerol-prolipoprotein

diacylglyceryl transferase activity responsible of lipid attachment of the prolipoprotein to the periplasmic side of the cytoplasmic membrane, functional *lspA* encoding a signal peptidase II responsible of the cleavage of the signal peptide, and functional *lnt* gene for the final maturation of the lipoprotein. For periplasmic transport and insertion in the outer membrane *S. glossinidius* retains a completely functional lol system (*lolA, B, C, D, E*) composed by a specific ABC transport system (*lolC, D, E*) for the energy-dependent release of lipoproteins from the outer leaflet of the inner membrane to the periplasmic space, a periplasmic chaperone (*lolA*) for lipoprotein transport across periplasmic space to the outer membrane, and an outer membrane receptor (*lolB*) required for the final localization of lipoproteins in the outer membrane (Narita et al., 2004).

Other outer membrane proteins like porins OmpA, OmpF, and OmpC spans the outer membrane with β-barrel transmembrane domains linked by external loops domains. These specific β-barrel domains are amphipatic and do not cause the retention of proteins in the inner membrane, allowing their diffusion to the periplasmic space, but needs a specific outer membrane assembly complex for their insertion inside the outer membrane at β-barrel domains level (Nikaido, 2003). *S. glossinidius* retains functional genes for this assembly complex (*yaeT, nlpB, yfiO, yfgH*) together with functional *surA* and *skp* genes encoding periplasmic chaperones needed for the proper folding of outer membrane proteins.

4.3.5.7.5 Secretory pathways

For the different secretory pathways involved in protein export across the cytoplasmic membrane to the periplasm, *S. glossinidius* retains a completely functional signal peptide-dependent pathway (*secA, B, E, G, D, F, Y*) that constitutes the major route for protein secretion of unfolded proteins in bacterial cells, together with functional *lspA* and *lep* genes encoding leader peptidases (SPases) II and I respectively for signal peptide cleavage of lipoproteins (SPase II) and other exported proteins (SPase I). *S. glossinidius* also retains a functional TAT export system for protein secretion of macromolecular substrates (*tatA, B, C*), although pseudogenization affecting *tatE* gene can decrease the range of substrates able to be translocated (Sargent et al., 1998), and a completely functional signal recognition particle system (SRP) involved in the recognition and transport of specific proteins to cellular membranes for insertion and secretion composed of *ffh* gene that encodes a protein that binds to the sequence of pre-proteins, *ffs* gene that encodes a 4.5S RNA that targets the nascent polypeptide-ribosome complex to the inner membrane, and *ftsY* gene that encodes a membrane receptor of protein-RNA complexes. The 4.5S-RNA and Ffh protein form the signal recognition particle (SRP), which binds to the signal peptide and targets the nascent protein to the SRP receptor FtsY in the cytoplasmic membrane (Batey et al., 2000). The *ffs* gene encoding the 4.5-RNA

appears unnanotated in the original annotation, and has been characterized during the reannotation process.

For the different protein secretion systems across outer membranes, *S. glossinidius* retains three different type III secretion systems representing three different symbiosis regions that has been extensively characterized in previous studies and in the original genome paper (Dale et al., 2001; Dale et al., 2002; Dale et al., 2005; Toh et al., 2006), whereas there is no signal for genes of the type II and type IV secretion systems.

4.3.5.8 DNA replication and repair

S. glossinidius retains all functional genes for complete replication machinery. This includes the complete DNA polymerase III holoenzyme, composed by the preinitiation complex (*holA, holB, holD, holC* and *dnaX*) that binds to the single-stranded DNA near the RNA primer for initiation of replication, the β-clamp subunit encoded by the gene *dnaN* that links the core polymerase to DNA stimulating its activity and increasing its processivity, and the core polymerase complex (*dnaE, dnaQ* and *holE*) responsible for polymerase and exonuclease activities of DNA polymerase III holoenzyme. Replication proceed by the binding of initiator protein DnaA to the replication origin followed by the binding of DNA helicase (*dnaB*) that carries out the unwinding of the parental double stranded DNA with the help of an accessory protein (*dnaC*) and a DNA gyrase complex (*gyrA* and *gyrB*) that removes DNA supercoiling for the action of DnaB helicase. A primase encoded by the gene *dnaG* carries out the synthesis of small RNA primers for the initiation of replication on the lagging strand through Okazaki fragments in a discontinuous manner, essential for the initiation of replication by DNA polymerase III holoenzyme. The DNA helicase and the primase constitute the primosome complex, which requires additional proteins for their correct assembly (*priA, priB, priC, dnaT* and *dnaC*). Once synthesized, the small RNA primers of the Okazaki fragments must be removed, the gap between fragments must be filled, and the remaining nicks between fragments must be closed in order to complete the replication cycle. This is carried out by multifunctional DNA polymerase I (*polA*) that have also 5'-3' exonuclease activity, RNAase H1 (*rnhA*), and DNA ligase (*ligA*). All these enzymatic activities are functional in *S. glossinidius*.

For the different systems of DNA recombination and repair that constitute the SOS response in bacterial cells, *S. glossindius* retains functional *recA* gene, which encodes a multifunctional protein responsible of DNA strand exchange in homologous recombination, induction of SOS response acting on LexA repressor, and mutagenic bypass of DNA lesions during SOS response. In addition, *S. glossinidius* also retains functional nucleotide excision repair complex (*uvrA, B, C*) for the restoration of single strand breakages in DNA sequences and functional

RecFOR (*recF, O, R*) and RecBCD (*recB, C, D*) complexes that mediates RecA-dependent homologous recombination between homologous DNA regions for DNA repair under double strand breakage through the formation of Holliday junction structures (Sharples, 2009). *S. glossinidius* also retains functional RuvABC resolvasome (*ruvA, B, C*) and RecG helicase (*recG*) that binds to this Holliday junction structures and catalyzes their resolution by strand cleavage and branch migration (Sharples, 2009). *S. glossinidius* also retains functional genes for the methyl-directed mismatch repair pathway (*mutH, L, S*), an alternative pathway for DNA repair that targets mispaired bases that arise by replication errors, during homologous recombination, and as consequence of DNA damage (Schofield and Hsieh, 2003).

In contrast, pseudogenization has affected to *polB* gene, which encodes a bifunctional DNA polymerase and exonuclease involved in replication restart following UV exposure, and *dinB* gene that encodes a DNA polymerase IV that lacks proofreading activity, introducing spontaneous mutations in DNA sequences that increases the mutation rate. There is no signal of *umuC* and *umuD* genes that encodes DNA polymerase V complex, another polymerase that introduces random nucleotides under conditions of massive DNA damage, increasing also the mutation rate (Sharples, 2009). Pseudogenization has also affected to *sbcC* and *sbcD* genes that encodes a double stranded DNA exonuclease complex involved in elimination or repair of DNA secondary structures, and *exoX* gene, which encodes another exonuclease involved in the methyl-directed mismatch repair pathway removing unmethylated strand harboring the mismatch, although retains functional exonuclease I (*sbcB*) and exonuclease VII (*xsaA* and *xsaB*) that catalyzes the same reaction.

4.3.5.9 DNA transcription

S. glossinidius retains completely functional transcriptional machinery. This includes functional *rpoA, rpoB* and *rpoC* genes encoding the α, β and β' subunits respectively of the core RNA polymerase holoenzyme, together with an almost complete set of sigma (σ) factors for promoter recognition under different cell conditions that includes the major σ^{70} subunit (*rpoD*) required for transcription of most genes involved in fundamental cell functions under normal growth conditions, σ^{32} (*rpoH*) and σ^{24} (*rpoE*) required for transcription of genes involved in heat-shock response, σ^{54} (*rpoN*) required for transcription of genes involved in nitrogen assimilation, and σ^{28} (*fliA*) required for transcription of genes involved in motility. Pseudogenization has affected to *rpoS* gene encoding a σ^{38} factor required for transcription of genes involved in cell growth under stationary phase (Record et al., 1996). *S. glossinidius* also retains *rsd* gene encoding an anti-sigma factor that binds specifically to σ^{70} factor blocking its activity during the transition from exponential growth to stationary phase.

For the different transcription factors, *S. glossinidius* retains functional *nusA*, *nusB* and *nusG* genes encoding elongation factors that modulate the rates of transcript elongation and participate in the formation of hairpin structures at attenuator sequences downstream of operon promoters that prevent transcription termination at internal terminator sequences, and functional *greA* and *greB* genes encoding elongation factors that cleaves the 3' end of RNA transcripts in stalled or arrested transcriptional complexes preventing transcriptional arrest of RNA polymerase core holoenzyme. *S. glossinidius* also retains functional *mfd* gene encoding a transcription-repair coupling factor that dissociates arrested RNA polymerase complexes consequence of DNA damage, and functional *rho* gene encoding the major transcription terminator factor Rho involved in the release of newly synthesized RNA from its complex with RNA polymerase and DNA template (Greenblatt, 1996).

4.4 DISCUSSION

S. glossinidius is a non-pathogenic facultative endosymbiont of tsetse flies and is especially attractive from an evolutionary point of view because is one of the few completely genomes of a bacterial endosymbiont at the very beginning of the transition from free-living to a host-dependent lifestyle. Their genome sequence reveals an ongoing process of massive genome reduction with a high number of pseudogenes but without reduction in genome size in comparison with their free-living relatives, indicative of a very recent symbiotic association with their tsetse host that has been previously confirmed in previous phylogenetic studies in which there is observed no co-evolution between phylogenetic trees of *S. glossinidius* and their corresponding *Glossina* host species characteristic of ancestral symbiotic associations (Aksoy et al., 1997; Chen et al., 1999). The detailed survey of *S. glossinidius* intergenic regions based on BLASTX searches carried out in this chapter allows to increase the number of pseudogenes from 972 described but not annotated in the original genome paper (Toh et al., 2006) to 1501 pseudogenes characterized at both nucleotide and amino acid level. The differences in the number of pseudogenes compared with the original annotation can be explained by the different methodologies used for pseudogene identification in this work. In the original genome paper, pseudogene identification is based on the results of different gene prediction programs, considering as pseudogenes all CDSs with less than half the length of its functional homologs in BLASTP searches (Toh et al., 2006). This approach limits pseudogene identification to recent inactivation events that conserve start or end positions along the open reading frame. By contrast, pseudogene identification based on BLASTX searches with raw nucleotide sequences of

intergenic regions against protein databases allows to characterize highly degraded pseudogenes whose open reading frames are not predicted by *ab-initio* gene prediction methods due to their advanced stages of degradation, as well as pseudogenes originated by insertion of IS elements where different frames of the ancestral gene are located upstream and downstream of the IS. Similar approaches for pseudogene identification identifies at least 100 additional pseudogenes in the genomes of the *E. coli/Shigella* clade(Lerat and Ochman, 2004), 118 additional pseudogenes in the genome of *Yersinia pestis* CO92 (Lerat and Ochman, 2005), or 6895 potential pseudogenes over 64 prokaryotic genomes (Liu et al., 2004). In addition, detailed analysis of genes and pseudogenes limits combined with sequence similarity searches allows to identify 142 originally annotated genes that are susceptible to be considered as potential pseudogenes because corresponds to situations of CDSs shorter than their corresponding database homologs and is possible to detect the absent part of the ancestral gene as an adjacent pseudogene.

The results of the functional re-annotation of genes and pseudogenes indicate a massive presence of genes related to mobile genetic elements (831 CDSs), being also the functional class most affected by the pseudogenization (441 out of 831 CDSs as pseudogenes). Mobile genetic element expansion has been traditionally associated with initial stages of bacterial adaptation to host dependent lifestyle as consequence of the relaxed selective pressures over large segments of the genome that becomes non-essential in the new environment associated with the insect host, a fact that allows massive proliferation of mobile genetic elements across the genome without detrimental effects to the bacterial endosymbiont (Moran and Plague, 2004; Bordenstein and Reznikoff, 2005; Dale and Moran, 2006). The characterization of the complete set of IS elements present in the genome of *S. glossinidius* revealed that IS elements represents only 2.72% of the genome sequence, much lower than the estimate loads of IS elements for the genomes of *Sitophilus oryzae* and *Sitophilus zeamays* primary endosymbionts (SOPE and SZPE respectively), that are the closest relatives of *S. glossinidius* in phylogenetic reconstructions (Plague et al., 2008; Dougherty and Plague, 2008; Gil et al., 2008). All three bacterial endosymbionts form a distinct clade within the γ-proteobacteria in close proximity with pathogenic and free-living enterobacteria, and are characterized by the recent association with their corresponding insect hosts, dated some 50-100 million years ago (Heddi et al., 1998; Dale et al., 2002). However, SOPE and SZPE, in contrast to *S. glossinidius*, are obligated mutualistic bacteria that reside exclusively inside specialized bacteriocyte cells in their weevil hosts (Heddi et al., 1999; Heddi et al., 2001). BLASTN comparisons with the consensus sequence of the 5 different IS elements characterized in the genome of *S. glossinidius* against the recently described sequences of four IS elements from the genome of SOPE (Gil et al., 2008) showed high similarity between ISSgl1 of *S. glossinidius* and ISsope1 from SOPE that can be explained by a common evolutionary origin, with 82% of identity at nucleotide level, whereas there is no similarity between the rest of IS elements,

possibly reflecting independent acquisitions of different set of IS elements posterior to the divergence of both lineages. In addition, a detailed survey of IS flanking regions by BLASTX searches revealed that only 18 of the 1501 identified pseudogenes of *S. glossinidius* were originated by IS insertion, reflecting that IS transposition has not been a major force in the process of gene inactivation in *S. glossinidius*, with the majority of pseudogenes generated by multiple frameshift mutations or premature stop codons. This indicates that gene inactivation in *S. glossinidius* has been produced by multiple single gene inactivation events that generates a large proportion of non-functional DNA that will be eliminated gradually due to the inherent mutational deletional bias associated with bacterial genomes and the lack of selective pressure for the maintenance of these non-functional regions (Andersson and Andersson, 1999b; Mira et al., 2001; Silva et al., 2001).

The high amount of mobile genetic elements detected in the genome of *S. glossinidius* is consequence of the massive presence of phage-related CDSs. Bacteriophages are particularly abundant in bacterial pathogens associated with the transmission of virulence traits through lysogenic conversion of the recipient genome, and are responsible of most of the strain-specific DNA between closely related bacteria (Perna et al., 2001; Ohnishi et al., 2001; Brussow and Hendrix, 2002; Thomson et al., 2004; Canchaya et al., 2004). In the context of bacterial endosymbionts, bacteriophage elements have been identified in recent symbiotic associations like bacteriophages APSE-1 and APSE-2 in the genome of the secondary endosymbiont of aphids *ca. Hamiltonella defensa*, where are associated with the protection activity carried out by this secondary endosymbiont killing parasitoid wasp larvae (Oliver et al., 2003; Moran et al., 2005a; Degnan and Moran, 2008b; Degnan and Moran, 2008a), as well as in parasitic *Wolbachia spp.*, where a bacteriophage WO that was originally associated with the induction of cytoplasmic incompatibility on invertebrate hosts, has been recently proposed to be beneficial for the invertebrate host allowing to control the loads of bacterial cells through prophage lytic development (Bordenstein and Wernegreen, 2004; Bordenstein et al., 2006). This is the situation of *S. glossinidius*, in which 17.7% of their 3932 CDSs (genes and pseudogenes) corresponds to prophage proteins. In addition, TBLASTX comparison of *S. glossinidius* against the phage subdivision of GenBank database allowed the characterization of two complete Mu-like prophage elements with differential incidence of pseudogenes although both revealing an inactive profile, and almost 11 prophage regions that shows homology with different domains of completely sequenced phage genomes. In addition, 2 of this prophage domains show homology with phage element epsilon 15 (NC_004775.1) that has been postulated as one of the precursors of the extrachromosomally replicating element pSG3 from *S. glossinidius* str. *Morsitans* together with phage element HK620 (NC_002730.1) through homologous recombination between this related ancestral prophages (Clark et al., 2007), possibly reflecting a common origin of the extrachromosomal element

pSG3 and this 2 genome prophage regions of *S. glossinidius* genome in the tsetse host *Glossina morsitans morsitans*. In addition, transposases from ISSgl5 family have 75% of identity at amino acid level with a transposase encoded in the genome of pSG3, reinforcing the possible flux of genetic material between the extrachromosomal element pSG3 and the bacterial chromosome. Is important to take into account that, whereas extrachromosomal plasmids elements of *S. glossinidius* (pSG1, pSG2, and pSG4) are conserved between strains from different tsetse species, more plasticity has been detected associated with the bacteriophage-like element pSG3, with differences in size and structure between *S. glossinidius* from *Glossina morsitans morsitans* and *Glossina palpalis palpalis*, being absent in *S. glossinidius* from *Glossina austeni* (Darby et al., 2005; Clark et al., 2007), so it would be expected to observe considerable differences in the structure and content of prophage elements in the genome sequence of *S. glossinidius* from different tsetse species.

Finally, the analysis of *S. glossinidius* metabolism revealed a metabolic profile that is closer to free-living bacteria than to obligate mutualists, both in terms of energy production from different carbon sources and in terms of biosynthetic capabilities for most essential metabolites and macromolecules, with the unique exception of L-arginine and thiamine. Our analyses disagree with those previously reported for amino acid biosynthesis that indicated that *S. glossinidius* was able to synthesize all amino acids except L- alanine (Toh et al., 2006), and showed that L-arginine was the unique amino acid lacking a complete biosynthetic pathway. This is consequence of pseudogenization events at four different steps, a fact that also affect to other related biosynthetic pathways like putrescine, L-lysine and peptidoglycan biosynthesis, and that supposes the necessity of an external supply of L-arginine from the host that may be responsible of the loss of extracellular stage in *S. glossinidius* lifecycle. We have also observed a putative metabolic complementation between *S. glossinidius* and *W. glossinidia* to produce the active cofactor thiamine pyrophosphate based on the gene repertoire of both bacteria. *S. glossinidius* was unable to produce thiamine (Toh et al., 2006), while *W. glossinidia* genome was described as having the potential to synthesize it (Akman et al., 2002). The screening for the complete set of genes required for thiamine biosynthesis in *W. glossinidia* showed lack of two essential components (*thiI* and *thiF*) which would avoid the *de novo* biosynthesis of thiamine. However, a detailed revision of the proteins encoded by both endosymbiont species revealed that each endosymbiont was able to synthesize one of the two moieties that the enzyme thiamine phosphate synthase combines in the pathway for the synthesis of the active cofactor thiamine diphosphate (Begley et al., 1999; Jurgenson et al., 2009). While *W. glossinidia* has the capability of synthesizing the pyrimidine moiety (HMP-PP), *S. glossinidius* has the capability of synthesizing the thiazole moiety (THZ-P). Considering that both bacterial symbionts are able to acquire the missing metabolite from outside, they would be able to synthesize the cofactor thiamine diphosphate after the action of

thiamine phosphate synthase and thiamine kinase. The genes encoding these enzymes (*thiE* and *thiL*, respectively) are present in the *W. glossinidia* genome, but only the second is present in the genome of *S. glossinidius*. We have found that *S. glossinidius* contains the gene *yjbQ* described recently as a thiamine phosphate synthase gene homolog, having an alanine in position 89 of the protein that has been demonstrated experimentally in *E. coli* that increases enzyme activity (Morett et al., 2008). However, it is important to consider that this hypothesis is based on the unique two available genomes sequences of both tsetse endosymbionts, that unfortunately come from different tsetse host species, and as a consequence may not reflect the real metabolic scenario in each tsetse host. In addition, *W. glossinidia* is essential for host reproduction and fitness, whereas *S. glossinidius* have less detrimental effects upon their specific removal from tsetse host due to their more recent association, although it produces a marked decrease in tsetse longevity, so although this potential complementation could be possible in light of genome sequences, it appears not to play a major role in tsetse physiology (Dale and Welburn, 2001).

Just a few cases of complementation between the metabolisms of two insect endosymbionts for mutual benefit have been described. Two types of complementation may be distinguished. In one case, both endosymbionts produce different end products of the metabolism (i.e. amino acids). In the other case, each endosymbiont controls part of the biosynthetic pathway, and the combined effort of both is required in order to produce the final product. The case of *Sulcia muelleri* and *Baumannia cicadellinicola*, endosymbionts of the sharpshooter *Homalodisca coagulata*, illustrates very well both types of complementations (Wu et al., 2006; McCutcheon and Moran, 2007). Thus, many amino acids are synthesized by *S. muelleri* and provided to *B. cicadellinicola* and to the host, except methionine and histidine, which are synthesized by *B. cicadellinicola* and provided to the co-endosymbiont and the host. The biosynthesis of one of these amino acids, methionine, is shared between both endosymbionts with *S. muelleri* providing the intermediate homoserine to *B. cicadellinicola*.

Another case of by biosynthetic pathway sharing is the tryptophan biosynthesis by the two endosymbionts of the aphid *Cinara cedri* (Gosalbes et al., 2008). *Buchnera aphidicola* performs the first enzyme step producing anthranilic acid. It is uptaken by the co-endosymbiont *Serratia symbiotica* which has the genes that encode the remnant steps of the pathway. Then, the produced tryptophan may be supplied to the co-endosymbiont and to the insect host.

A similar complementation explanation arises from the analysis of folate and Coenzyme A biosynthetic pathways, both completely functional in *S. glossinidius* but impaired in *W. glossinidia* due to their inability to produce 2,3-ketoisovalerate and p-aminobenzoate (see results). Interestingly, the complementation at 2,3-

ketoisovalerate level has been also observed between the bacterial endosymbionts of sharpshooters, in which *B. cicadellinicola* is able to produce Coenzyme A from 2,3-ketoisovalerate but is not capable to produce 2,3-ketoisovalerate itself, which is produced by *S. muelleri* in the pathway for L-valine biosynthesis. However, sharpshooters endosymbionts inhabits in close proximity inside the bacteriome organ, whereas tse-tse flies endosymbionts are located in different tissues, with *W. glossinidia* being located inside bacteriocytes that forms the bacteriome in the anterior midgut and *S. glossinidius* having a more wider tissue tropism, although recently has been described the presence of an extracellular population of *W. glossinidia* in the lumen of milk-gland tissue of female tse-tse flies that co-inhabits with *S. glossinidius*, being both transmitted to the progeny through milk-gland secretions that nourishes the developing descendents in the mother´s uterus (Attardo et al., 2008; Pais et al., 2008). It´s also worth to notice that the sequenced genomes of *S. glossinidius* and *W. glossinidia* comes from different tse-tse host species, with sequenced *W. glossinidia* that came from *Glossina brevipalpis* (Fusca group) and sequenced *S. glossinidius* that came from *Glossina morsitans morsitans* (Morsitans group), so metabolic inferences about possible complementation between both endosymbionts have to be taken with caution. The complete genome sequences of the complementary endosymbionts (*W. glossinidia* of *G. morsitans morsitans* and *S. glossinidius* from *G. brevipalpis*) would give a more realistic picture of the symbiotic association of tse-tse flies and their bacterial endosymbionts in order to compare the degree of genome stability associated to bacterial endosymbionts at different time-scales of the symbiotic association and to test if the metabolic profile associated with each member of the same symbiotic association gives support to the metabolic complementation hypothesis.

4.5 CONCLUSIONS

A complete re-analysis of the genome sequence of *S. glossinidius* has revealed novel insights in the initial stages of the transition from free-living to a host dependent lifestyle in this bacterial genome. The relaxed selective pressures over non-essential genome regions in the more stable environment of the tsetse host have lead to a massive proliferation of mobile genetic elements in the form of prophages and IS elements, although their impact in the process of gene inactivation is miminal, with most of the pseudogenes generated by frameshift mutation or premature stop codons and with only 18 pseudogenes generated by insertion of an IS element. A detailed survey of intergenic regions led to the characterization of a total number of 1501 pseudogenes, significatively higher than the 972 genes reported in the original annotation, pointing out to the importance of sequence analysis in the characterization of highly degraded pseudogenes that is not possible to identify by “ab-initio" gene prediction methods. In addition, 142 originally annotated genes

shows also hallmarks of an ongoing process of gene inactivation that can be affecting their functionality.

A detailed survey of the metabolic capabilities of genes and pseudogenes together with a comparison with the metabolic profile of tsetse fly primary endosymbiont *W. glossinidia* revealed novel inactive pathways consequence of the pseudogenization process previously undescribed in *S. glossinidius*, like L-arginine biosynthesis pathway, as well as the functionality of L-alanine biosynthesis that was reported as the unique amino acid biosynthesis pathway inactivated in *S. glossinidius* (Toh et al., 2006). A possible phenomenon of metabolic complementation between both tsetse endosymbionts at thiamine biosynthesis level is also described, based on the inability of "de-novo" thiamine biosynthesis in both tsetse endosymbionts in light of genomic data. This possible metabolic complementation together with the incomplete pathways for Coenzyme A and Folate coenzymes in *W. glossinidia*, both completely functional in *S. glossinidius*, would explain the co-existence of both bacterial endosymbiont in the context of the tsetse host.

5. Reconstruction and functional analysis of the metabolic networks of Sodalis glossinidius, the secondary endosymbiont of tsetse flies: A systems biology approach to reductive evolution

5.1 INTRODUCTION

The availability of complete genome sequences for hundreds of organisms together with the development of high-throughput experimental techniques to analyze cellular functions at whole genome level have revealed the increasing complexity of living systems. The multigenic nature of cellular function presents a significant challenge in extracting phenotypic information from the genome sequence. In the context of cellular metabolism, this can be accomplished with the reconstruction of detailed metabolic networks from annotated genome sequences that describe the functional potential of the corresponding organism and the analysis of its systemic properties by different approaches that include flux-balance analysis (FBA) or pathway analysis. The ultimate goal of computational analysis of metabolic networks would be the development of dynamic models that allow the complete simulation of metabolic systems, although this approach is hampered by the lack of detailed kinetic information on the dynamics and regulation of many metabolic reactions, limiting the dynamic modeling of complete cellular systems with the exception of human red blood cells (Edwards and Palsson, 2000c; Edwards et al., 2002). However, in the absence of kinetic information, is still possible to accurately evaluate the metabolic potential and functional capabilities of a whole cell system through steady-state analysis of metabolic networks. Steady-state analysis is based uniquely on the stochiometry of metabolic reactions, which is a structural invariant of metabolic networks, without considering kinetic information (Varma and Palsson, 1994a; Schilling et al., 2000a; Lee et al., 2006).

Stochiometric analysis of genome-scale metabolic networks can be focused on two main objectives, the inference of optimal metabolic phenotypes under different environmental conditions through FBA and the inference of metabolic pathways or enzyme subsets through the related concepts of elementary flux modes and extreme pathways (Schilling et al., 2000a; Klamt and Stelling, 2003; Planes and Beasley, 2008). Both approaches share a common mathematical framework that is based on the principles of convex analysis and that relies on the basic stochiometry of the reactions of the metabolic network. Metabolic networks are a collection of enzymatic reactions and transport processes that describes the drain and production of cellular metabolites in the cellular system under study (Schilling et al., 2000b). This set of reactions can be inferred from the information harbored in the annotated genome sequence complemented with additional biochemical information from experimental studies and database searches (Rocha et al., 2008; Durot et al., 2009). An example of a simple metabolic network is represented in Figure 5.1.

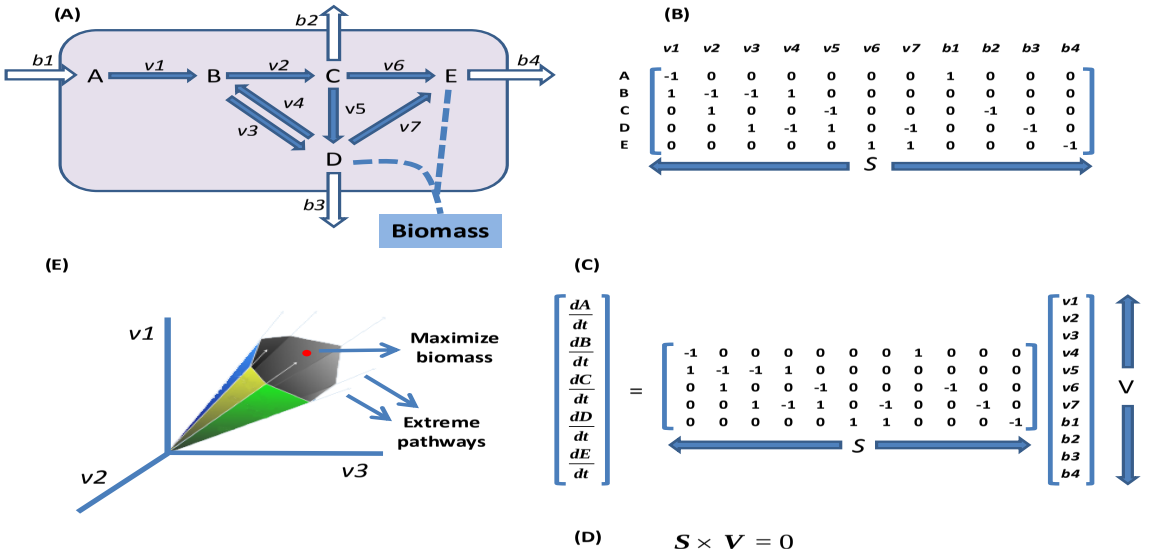


Figure 5.1: *Example of a metabolic network (A) with 7 internal reactions (v1-v7), 5 metabolites (A-E) and 4 exchange fluxes (b1-b4). The basic structural properties of the metabolic system can be represented in the form of a stochiometric matrix S with columns representing reactions, rows representing metabolites, and the indices representing the stochiometric coefficient of each metabolite in each reaction (B). Mass balance equations accounting for all reactions and transport mechanisms can be formulated for each metabolite representing the change in metabolite concentrations over time. These equations are formulated in matrix form as the product of the stochiometric matrix S by the vector of reaction fluxes V (C). Under the steady state assumption, all reaction fluxes producing and consuming a particular metabolite must be balanced, reducing the set of differential equations to a system of linear homogeneous equations (D). This generates a space of feasible solutions (flux distributions) in the n-dimensional space of reactions that are represented as a flux cone that describes the metabolic capability of the system, represented in (E) simplified for three reaction fluxes. The edges of the cone (grey arrows in (E)) correspond to the extreme pathways or convex basis. Elementary flux modes can be inferred from the set of extreme pathways. In FBA, an optimal flux distribution that maximizes a particular objective function (in the example the production of biomass composed of metabolites E and D) is found, that corresponds to a single point in the space of feasible solutions represented in the flux cone.*

The whole system is closed to the passage of certain metabolites, while others are allowed to enter and/or exit to the system based on external sources or sinks that are operating on the network as a whole, and that define the external conditions under which the system operates in terms of nutrients and oxygen availability. External sources or sinks, also named pseudoreactions (Schilling et al., 2000b), are represented as exchange fluxes in the context of the metabolic network and define the system boundaries around the entire set of cellular and transport reactions, that become defined as internal fluxes. Exchange fluxes are not strictly physical biochemical conversions or transport processes like those of internal fluxes but can be better considered as the inputs and outputs to the system. The activity of exchange fluxes can be considered positive if the metabolite is exiting or being produced by the system and negative if the metabolite is entering or consumed by the system. Once defined the set of internal reactions and exchange fluxes, the analysis of metabolic systems begin with the definition of their structural characteristics or invariant properties, that are those that are not depend neither on the external conditions nor on the internal state of the system, but only on its structure (Reder, 1988). The most important structural property of metabolic networks is the stochiometry of their metabolic reactions, which describes the architecture and topological properties of the system, being essential for the subsequent functional analysis. Stochiometry refers to the molar ratios in which substrates are converted into products in a chemical reaction, like the conversion of one mole of glucose and one mole of ATP into one mole of glucose-6-phosphate and one mole of ADP by glucokinase during glycolytic process (Schilling et al., 2000b). These ratios remain constant over changing reaction conditions that could alter kinetic parameter and rate reaction as a function of time. Stochiometry of metabolic networks is mathematically represented in the form of a Stochiometric matrix *S* with *m* rows and *n* columns, where *m* corresponds to the number of metabolites and *n* to the number of reactions (Figure 5.1). With the defined stochiometry of all metabolites comprised in the metabolic network, dynamic mass balances can be written around every metabolite in the system that describes the net change in metabolite concentration as function of time, and that is equal to the differences in the sum of all fluxes that produces the metabolite and those which consume it. This becomes defined by the flux vector *v* (Figure 5.1):

$$\frac{dX}{dT} = S \times v$$

This set of differential equations for each metabolite incorporates enzyme kinetics through the time derivative, with vector of reaction fluxes (v) that depends on the concentration of metabolites and a number of kinetic parameters. Due to the lack of detailed kinetic information in many reactions of the system, it is reasonable to place the metabolic system into a steady state (Schilling et al., 2000a; Schilling et al., 2000a). This is the most important constraint in FBA and pathway analysis

(mass balance constraint), and is based on the assumption that metabolic transients in internal metabolites are generally more rapid than both cellular growth rates and dynamic changes in the organism environment (Varma and Palsson, 1994a). As a consequence, metabolic fluxes leading to the formation and degradation of any particular metabolite must be balanced (Figure 5.1):

$$S \times v = 0$$

The steady-state conditions transform the set of initial differential equations into a set of linear homogeneous equations from which is possible to calculate flux distributions (v vector in equation above). This system of equations is typically undetermined because the number of reactions exceeds the number of participating metabolites in metabolic networks, and as consequence no single solution in the form of a single flux vector could be defined because multiple solutions (flux distributions) that satisfy the steady-state condition imposed to the metabolic system could be possible. Mathematically, the undetermined nature of the system of linear equations defined in steady-state generates a null space of possible solutions (flux distributions) that can be explored mathematically through convex analysis (Rockafellar, 1970). The set of solutions to any system of linear inequalities is a convex set, and corresponds geometrically to a convex polyhedral cone in the *n*-dimensional space defined by the *n*-reactions of the system (Figure 5.1). This convex cone can be also referred as steady-state flux cone or flux space, and within it lies all the possible steady-state solutions and hence flux distributions under which the system is able to function, representing the functional capabilities or metabolic phenotypes of a given metabolic network (Varma and Palsson, 1994a; Schilling et al., 2000a; Edwards et al., 2002; Famili and Palsson, 2003). The flux space determines what the system can and cannot do, what building blocks and biomass constituents is able to produce, how efficiently the system generates ATP and biomass constituents under alternative carbon sources, or where are critical links or essential genes in the system (Schilling and Palsson, 2000; Reed et al., 2003; Forster et al., 2003b; Blank et al., 2005; Kuepfer et al., 2005; Borodina et al., 2005). Efficient ways to explore and analyze the flux distributions conferred within the steady-state flux cone are necessary in order to determine their structural characteristics and their functional potential from a metabolic perspective. This can be achieved by either a pathway-oriented perspective through the computational identification of the set of Elementary Flux Modes (EFM) and Extreme Pathways (EP) that defines possible metabolic pathways present in the system and from a reaction-based perspective through Flux Balance Analysis (FBA) that allows the identification of the optimal flux distribution in order to achieve a particular metabolic objective.

5.1.1 Stochiometric analysis of Metabolic Pathways

Metabolic pathways like glycolysis or TCA cycle are comprised of groups of enzymes that act coordinately towards the production or breakdown of a certain metabolite or group of metabolites,and can be defined by qualitative identification based on historical group of reactions in a database settings (Overbeek et al., 2000; Kanehisa et al., 2004; Keseler et al., 2009; Caspi et al., 2009) or alternatively through a rigorous quantitative an systemic computational identification of metabolic modules based on mathematical principles such as linear algebra and convex analysis (Schilling et al., 2000a; Schilling et al., 2000b; Schuster et al., 2000; Papin et al., 2004). In the context of metabolic networks, metabolic pathways have been described under two similar mathematical concepts named elementary flux modes (EFMs) and extreme pathways (EP). EFMs establish a link between structural analysis of metabolic networks and metabolic flux analysis, and are defined as the set of non-decomposable pathways that operates at pseudo-steady state in the metabolic networks fulfilling the stochiometric constraints of the system (Schuster et al., 1999; Pfeiffer et al., 1999). The set of EFM is unique for a given network structure, and enables to investigate the space of physiological states meaningful for the cell in the long term perspective (Stelling et al., 2002). In contrast, EPs are a subset of EFMs that fulfill the additional requirement of systemic independence, where no EP can be written as non-trivial non-negative linear combination of other EP (Schilling et al., 2000b; Planes and Beasley, 2008). In the mathematical context of the convex analysis, the set of EPs represents the edges of the cone, similar to the edges of a pyramid, and all possible metabolic flux distributions included within the cone can be represented as non-negative linear combination of these EPs (Schilling et al., 2000a). The set of EPs for a given particular system, also known as convex basis, is unique and represents the absolute minimal set of pathways that can be used to describe all feasible steady-state flux distributions (Schilling et al., 2000a; Klamt and Stelling, 2003). However, pathway analysis based on EFM and EP has major computational problems when dealing with highly complex metabolic networks such as genome-scale networks. As the metabolic network increases in size (number of reactions), the number of possible EFM and EP increases in a combinatorial fashion, making the complete enumeration of these pathways impracticable (Planes and Beasley, 2008). As an example, pathway analysis over a metabolic network representing the central metabolism of *E. coli* (110 reactions) yields a total number of 27.099 EFM with glucose as external carbon source, whereas 507.632 EFM are inferred in the same metabolic network under different carbon sources (Stelling et al., 2002; Klamt and Stelling, 2003). This problem is usually referred as “combinatorial explosion”, and limits the application of this approach to moderately sized metabolic networks with a limited number of reactions, being impossible to apply to genome-scale metabolic networks of moderate size like that of *E. coli* due to the difficulty associated with the computation and interpretation of the whole set of possible pathways..

5.1.2 Quantitative assessment of metabolic phenotypes: Flux Balance Analysis (FBA)

The performance capabilities of any metabolic system becomes defined by the convex space of feasible flux distributions defined under steady state condition, and can be further restricted by the imposition of additional constraints to the systems behavior in the form of thermodynamic constraints based on reaction reversibility, topological constraints based on spatial restriction of metabolites to specific cellular compartments, or environmental constraints based on specific conditions of nutrient and oxygen availability (Price et al., 2003; Price et al., 2004; Becker et al., 2007). In addition, maximum and minimum ranges of metabolic fluxes or even specific values of reaction fluxes experimentally measured by Metabolic Flux Analysis (MFA) can be specified (Christensen and Nielsen, 2000a; Christensen and Nielsen, 2000b; Fischer and Sauer, 2005). Pathway analysis allows to explore metabolic functions through the set of computationally defined EFM and EP, but in order to analyze quantitatively the functional capabilities of any metabolic system under different external conditions, an alternative approach named FBA is applied (Varma and Palsson, 1994a; Edwards and Palsson, 2000b; Edwards et al., 2002; Kauffman et al., 2003). FBA is a mathematical approach that uses linear optimization techniques to determine the optimal flux distribution within the convex space of feasible solutions that optimize a particular objective function, allowing to explore metabolic phenotypes in a quantitative manner (Figure 5.1) (Schilling et al., 2000a; Edwards et al., 2002; Durot et al., 2009).

A critical point in FBA is the definition of an objective function that represents cellular behavior in a realistic manner. Objective functions are usually formulated as physiologically meaningful objectives that quantitatively describe cellular growth in terms of biomass production, but also as specific design objectives focused on the exploitation of a given cellular system for bioengineering purposes through the maximization of the rate of production of a given metabolite. Alternatively, objective functions focused on the minimization of the rates of ATP production or nutrient consumption can be also formulated if the objective is to determine optimal phenotypes in terms of energy efficiency or determine conditions of optimal growth consuming minimal amounts of nutrients (Varma et al., 1993a; Edwards and Palsson, 2000a; Ramakrishna et al., 2001). From an evolutionary point of view, micro-organisms will choose the metabolic flux distribution that enhances their own survival in defined environmental conditions, which becomes defined by the external metabolites available to the system through exchange reactions. This optimal phenotype is expected to optimize cellular growth rates in wild-type microbial cells under nutrient-rich conditions (Price et al., 2003; Price et al., 2004). In fact, prediction of growth phenotypes under different environmental conditions is one of the primary uses of FBA over genome-scale metabolic models. Quantitative assessment of growth phenotypes can be achieved through FBA with the inclusion

in the metabolic network of an additional biomass equation as objective function to maximize that represents the cellular requirements of energy and biomass constituents in order to ensure cellular growth. Determining the biomass composition of the organism under study is therefore essential in order to obtain reliable predictions of growth phenotypes by FBA. The biomass function is typically written to reflect the needs of the cell in order to make one gram of cellular dry weight (Joyce and Palsson, 2008). This is normally achieved by examining the relevant literature regarding experimental measurements of biomass constituents of the organism under study (Reed et al., 2003; Forster et al., 2003a) or alternatively adapting the biomass equation of related organisms if the experimental information of the organism under study is scarce (Durot et al., 2009; Zhang et al., 2009). Alternatively, bi-level optimization approaches also allow to predict the most plausible objective function based on actual flux observations inferred experimentally from metabolomic data (Burgard and Maranas, 2003).

FBA over a metabolic network with maximizing biomass production as objective function computes the maximal growth yield achievable given a set of bounded uptake rates of external substrates. It relies on the strong assumption that bacteria have optimized their growth performance under a subset of possible environmental conditions during their evolution, so the maximization of biomass production can be considered a driving principle of metabolic operation (Varma and Palsson, 1994a). In fact, FBA prediction of cellular growth phenotypes with biomass production as objective function are consistent with experimental data 60% of the time for *Helicobacter pylori* (Schilling et al., 2002) and 86% of the time for *E. coli* (Edwards and Palsson, 2000d), rising to 91% of the time for *E. coli* when transcriptional regulation is considered (Covert and Palsson, 2002). FBA has been applied to study the physiology of ATP production during fatty acid biosynthesis by adipocytes (Fell and Small, 1986), ethanol secretion by yeast (Sonnleitner and Kappeli, 1986) and acetate secretion in E. coli (Varma et al., 1993b; Varma and Palsson, 1994b), where FBA predictions of cellular growth, glucose consumption rates and acetate secretion in aerobic and anaerobic conditions with biomass production as objective functions coincides with experimental observations in cell-batch cultures. Similar results are also obtained analyzing membrane transport fluxes by FBA over *E. coli* metabolic network, with FBA predictions being coincident with experimental results using biomass production as objective function (Edwards et al., 2001).

Since the publication of the first genome-scale metabolic network for *Haemophilus influenzae* (Schilling and Palsson, 2000), 29 metabolic networks of bacterial microorganisms have been reconstructed (http://gcrg.ucsd.edu/In_Silico_Organisms/Other_Organisms), far from the impressive number of bacterial genomes completely sequenced to date, a feature that reflects the difficulties in the modeling of metabolic networks, that is far from being an automated process from annotated genome sequences. Available genome-scale

metabolic networks range from the important model organism *E. coli* (Edwards and Palsson, 2000a; Reed et al., 2003; Feist et al., 2007) to pathogenic microbes such as *Helycobacter pylori* (Schilling et al., 2002; Thiele et al., 2005), *Staphylococcus aureus* (Becker and Palsson, 2005), or *Pseudomonas putida* (Oberhardt et al., 2008). In addition, genome-scale metabolic networks of bacterial species of biotechnological interest allow to investigate *in-silico* their potential for bioenergetic purposes or their potential as systems for the production of particular metabolites. This is the case of the genome-scale metabolic network of *Geobacter sulforreducens,* that has been utilized to investigate their potentiality in organic matter oxidation by reduction of a variety of radioactive and toxic ion metals (Mahadevan et al., 2006), or genome-scale metabolic network of *Streptomyces coelicor,* that has been used to study their potential capabilities for antibiotic production (Borodina et al., 2005). In addition to bacterial microorganisms, genome-scale metabolic models have been also developed for Achaea and Eukaryotes, being specially relevant in the case of the baker´s yeast *Saccharomyces cerevisiae* due to their extensive refinement through successive model revisions and modifications (Forster et al., 2003a; Duarte et al., 2004; Kuepfer et al., 2005).

5.1.3 Genome-scale metabolic networks as model system to predict gene essentiality and bacterial evolution

One of the most important applications of genome-scale metabolic models is to predict the effects of gene deletion on growth phenotypes through the limitation of metabolic flux to zero in reactions catalyzed by simulated knockout genes. This is achieved adding a layer of Gene Protein Reaction associations usually called GPR to the metabolic network (Reed et al., 2003). This type of analysis over *E. coli*, *S. cerevisiae*, or *P. putida* shows that model predictions of gene essentiality was highly coincident with experimental data, whereas discrepancies between model predictions and experimental results allow to refine the model in order to better explain the observed phenotypes (Edwards and Palsson, 2000d; Duarte et al., 2004; Oberhardt et al., 2008). This approach has been also applied successfully in drug target discovery over a metabolic model of the mycolic acid pathway in *Mycobacterium tuberculosis* in order to identify essential genes affecting the biosynthesis of mycolic acids that were also validated experimentally (Raman et al., 2005), as well as to identify a core of essential metabolic reactions shared by *E. coli*, *H. pylori* and *S. cerevisiae* metabolic networks that correspond with the targets of many antibiotics that interfere with bacterial metabolism (Almaas et al., 2005). In addition, two alternative mathematical approaches derived from FBA and based on a minimal adjustment metric (MOMA and ROOM) have been developed to evaluate knockout systems based on the assumption that metabolism in a knockout mutant operates as closely as possible to the metabolic state of the wild-type strain instead maximize the particular objective function (Segre et al., 2002; Shlomi et al., 2005). However, although these alternative approaches would slightly improve metabolic predictions

immediately after the gene deletion event, experimental evolution approaches on *E. coli* measuring cell growth in glycerol minimal medium have demonstrated that although initial growth yield of knockout strains was suboptimal, it progressively evolved to maximize cellular growth as described by FBA maximizing biomass production (Ibarra et al., 2002).

The capability of Genome-scale models to predict the effect of gene deletions over metabolic phenotypes is also applied in metabolic engineering to select those gene deletions that would provide the greatest benefit for a given metabolite production objective. This is achieved through a bi-level optimization approach that simultaneously optimizes two objective functions in order to achieve the metabolic flux distribution that maximizes the production of the desired metabolite while also maximizes cellular growth in terms of biomass production (Burgard et al., 2003). This approach has been applied successfully to increase lycopene and lactate production in *E. coli* (Alper et al., 2005a; Alper et al., 2005b; Fong et al., 2005). Additional computational approaches have been also developed in order to evaluate the combined effect of deletions and additions of reactions for overproduction of targeted compounds (Pharkya et al., 2004). However, this kind of predictions from constraint-based analysis of metabolic models are limited because completely overlooked the essential role of gene regulation and enzyme kinetics in controlling the efficiency of metabolite production (Durot et al., 2009).

Finally, genome-scale metabolic models have been used to study the process of bacterial evolution, where gene content of highly streamlined genomes of *Buchnera aphidicola* and *Wigglesworthia glossinidia* could be predicted with high accuracy by successive gene deletions over *E.coli* metabolic network under external conditions that mimics nutrient availability found by each bacteria in their corresponding insect hosts (Pal et al., 2006b). In addition, FBA over *E. coli* metabolic network combined with comparative genomics and gene content evolution analysis reveals the predominant role of horizontal gene transfer over gene duplication in *E. coli* network evolution in the last 100 million years since their divergence from *Salmonella* lineage, as well as the essential role of horizontally transferred genes under specific external conditions and their preferential integration in peripheral nodes of the metabolic network (Pal et al., 2005a; Pal et al., 2005b).

In the context of this thesis, *S. glossinidius* provides one of the few examples of a bacterium in the very beginning of the process of genome reduction, and the massive presence of pseudogenes functionally annotated in the previous chapter provides a hallmark of the gene content and functional capabilities of the hypothetical free-living ancestor. In order to complement the description of the metabolic capabilities of this bacterium provided by the functional re-annotation of their whole genome sequence, in this chapter a complete reconstruction of the Genome-scale metabolic networks of *S. glossinidius* is carried out in order to study its complete evolution

from a systems-biology perspective. The quantitative assessment of metabolic phenotypes of ancestral and functional metabolic networks of *S. glossinidius* will be analyzed in comparison with their free-living relative *E. coli* in order to evaluate the effect of pseudogenization over the metabolic phenotype of the system under different external conditions and to determine the evolutionary events that could have determined the ecological transition from a free-living to a host-dependent lifestyle. In addition, reductive simulations over functional metabolic network of *S. glossinidius* will be carried out in order to analyze their possible future evolution in the context of the genome reduction process under different external conditions to determine possible coincidences with the reductive evolution pattern experimented by tsetse primary endosymbiont *W. glossinidia*. Finally, the patterns of sequence evolution in essential an non-essential genes defined by their presence in minimal metabolic networks generated from *S. glossinidius* functional network will be analyzed in order to determine if the essential character of genes defined from a network-based perspective corresponds with differential patterns of sequence evolution.

5.2 MATERIAL AND METHODS

5.2.1 Orthology identification

In order to reconstruct *S. glossinidius* metabolic network, we take *E. coli K12* as reference genome because is its closest free-living relative from which a completely reconstructed metabolic network is available (Edwards and Palsson, 2000a; Reed et al., 2003). The metabolic network of *E. coli K12* JR904 (Reed et al., 2003) was considered as starting point to reconstruct *S. glossinidius* metabolic network and for subsequent metabolic comparisons over reconstructed networks. The first step in the study was the identification of orthologous genes between *S. glossinidius* (2431 genes and 1501 pseudogenes characterized in the previous chapter of this thesis) and *E. coli* strain K12 substrain MG1655 (accession number NC_000913) with the program OrthoMCL based on BLASTP searches with a maximum e-value cutoff of 10^{-20} and a minimum of amino acid identity of 80% (Li et al., 2003). This analysis produces a total set of 2257 orthologous clusters between both genomes, from which 169 clusters were *E. coli* K12-specific, 193 clusters were *S. glossinidius*-specific, and 1895 clusters were composed by *E. coli* K12 and *S. glossinidius* genes. The set of 1895 clusters of orthologous genes between both genomes includes 1827 clusters composed by 2 genes (one for each genome) together with 68 clusters composed by more than 2 genes, corresponding to gene duplicates in one or both genomes. In order to differentiate between true orthologs and paralogs originated by gene duplication in this last set of 68 clusters with more than 2 genes, the synteny between *S. glossinidius* and *E. coli* K12 were examined based on TBLASTX comparisons between both genomes followed by manual inspection with Artemis Comparison Tool software version 6 (Carver et al., 2005). This renders 43 clusters

of truth orthologs between both genomes that combined with the 1827 clusters previously identified by OrthoMCL produces a final set of 1870 orthologous clusters between *S. glossinidius* and *E. coli* with one gene per genome.

Orthologous genes between *S. glossinidius* and *W. glossinidia* were also identified with OrthoMCL based on BLASTP searches with amino acid sequences of *S. glossinidius* (2431 CDSs) and *W. glossinidia* (617 CDSs) with the same cutoffs described above in order to compare the gene content of minimal networks generated by reductive evolution simulations over *S. glossinidius* functional network with the gene content of *W. glossinidia*. Clusters for 591 orthologous genes between both genomes were identified, over which essential and non-essential genes in *S. glossinidius* minimal networks under different external conditions were mapped.

5.2.2 Metabolic networks reconstruction and computational inference of cellular behaviour through Flux Balance Analysis (FBA)

In order to reconstruct the ancestral and functional metabolic networks of *S. glossinidius*, the metabolic network JR904 of *E. coli* K12 was taken as reference (Reed et al., 2003). This network comprises 904 genes, 931 internal reactions including metabolic reactions and transport processes, and 143 exchange fluxes that define the system boundaries of the whole network in terms of nutrient availability. The process of network reconstruction can be broadly divided into three main steps, the identification of orthologous genes of JR904 network genes in *S. glossinidius*, the characterization of additional gene deletion events of genes of JR904 network in *S. glossinidius* evolution, and the final curation step that comprises the addition of *S. glossinidius* genes and pseudogenes related with metabolic functions with complete stochiometry information that have not been incorporated by sequence similarity with *E. coli* K12 genes and the functional assessment of the network reconstructions by FBA.

In an initial step, the 904 genes of *E. coli* K12 JR904 network were mapped over the 1870 clusters of orthologous genes between *S. glossinidius* and *E. coli* K12, producing a preliminary set of 597 *S. glossinidius* CDS´s (468 genes and 129 pseudogenes) orthologous to JR904 network genes. In a second step, the 2062 *S. glossinidius* CDSs without orthology with *E. coli* K12 by OrthoMCL were analyzed in order to incorporate genes and pseudogenes related with metabolic functions based on the whole genome functional re-annotation carried out in the previous chapter of this thesis, from which 27 pseudogenes and 13 genes that harbor a complete EC number assigned were incorporated to the preliminary network reconstruction. Finally, in order to incorporate gene deletion events in the evolution of *S. glossinidius* metabolic network, the gene context of the 307 genes of JR904 network without orthology with *S. glossinidius* were analyzed based on ACT

comparisons between both genomes, in order to identify syntenic genome regions around this genes. This yields a preliminary set of 116 JR904 genes absent in *S. glossinidius* but with a conserved genome context between both genomes. This preliminary set of 116 absent genes could have evolved by gene deletion events during *S. glossinidius* evolution or alternatively could be the result of horizontal gene transfer events in the lineage leading to *E. coli* K12 and as consequence cannot be considered as present in the ancestor of *S. glossinidius*. In order to differentiate between these evolutionary scenarios, these 116 genes of *E. coli* JR904 network were mapped in the related genomes of *Serratia marsecens* (http://www.sanger.ac.uk/Projects/S_marcescens/), *Serratia proteomaculans* (NC_009832), *Erwinia carotovora* (NC_004547), *Yersinia pestis CO92* (NC_003143), *Enterobacter sp. 638* (NC_009436), and additional *E. coli* genomes of strains O6:H1 CFT073 (NC_004431), O157:H7 EDL933 (NC_002655) and APEC O1 (NC_008563) with OrthoMCL based on BLASTP searches of all against all proteins with a maximum e-value cutoff of 10^{-20} and a minimum amino acid identity of 80%. Genes that were present in the 4 *E. coli* genomes and at least 2 additional genomes were considered as effectively removed during *S. glossinidius* evolution. This analysis yields 41 additional genes of *E. coli* JR904 metabolic network specifically deleted during *S. glossinidius* evolution.

The result of this analysis is an ancestral gene set of 481 genes, 156 pseudogenes, and 41 deleted genes that can be considered as functional in the ancestral metabolic network of *S. glossinidius*. In order to reconstruct the ancestral metabolic network of *S. glossinidius* the metabolic and transport reactions catalyzed by these genes needs to be defined. In this context, different types of reactions can be defined in the metabolic network (Figure 5.2):

1) Reactions catalyzed by a single enzyme encoded by a single gene. If the gene is present as gene, pseudogene or deleted gene the reaction is included in the ancestral metabolic network. 520 reactions are included in the ancestral network.

2) Reactions catalyzed by a single enzymatic complex encoded by different genes. If all the complex subunits are present as gene, pseudogene or deleted gene the corresponding reaction is included in the network. 82 reactions are included in the ancestral network.

3) Reactions catalyzed by different isozymes or isoenzymatic complexes. If at least one of the isozymes or isoenzymatic complexes is present as gene, pseudogene or deleted gene the corresponding reaction is included in the ancestral network. 73 reactions are included in the ancestral network.

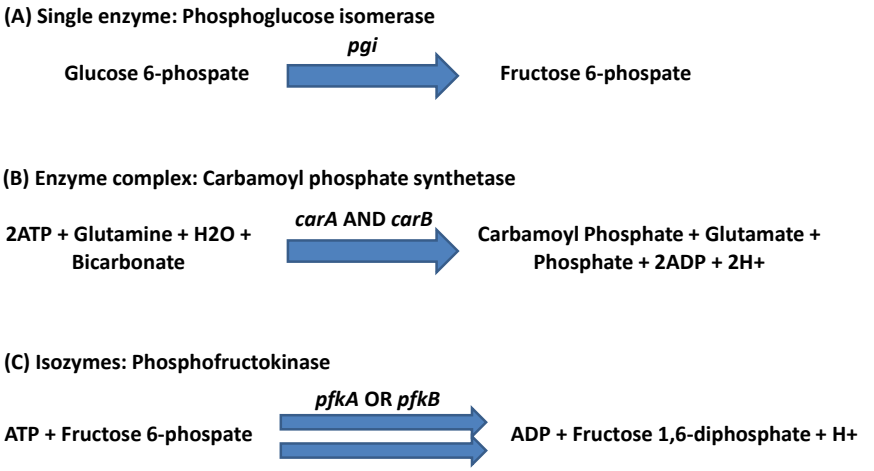


Figure 5.2: *Reactions included in the metabolic network. The simplest case corresponds to a reaction catalyzed by a single enzyme (A). Reactions catalyzed by an enzymatic complex (B) are represented by Boolean operator AND in gene-protein-reaction associations (GPR). Reactions catalyzed by different isozymes or isoenzymatic complexes are represented by Boolean operator OR in GPR (C).*

This renders a set of 682 reactions in the ancestral network that are catalyzed by 479 genes, 148 pseudogenes and 41 deleted genes. This suppose the removal of 8 pseudogenes and 2 genes with orthology with *E. coli* JR904 metabolic genes from final ancestral network of *S. glossinidius* due to their association to reactions catalyzed by enzymatic complexes where not all genes of the complex are detected as genes, pseudogenes or deleted genes, so these reactions and their corresponding genes are not considered as functional in the ancestral metabolic network of *S. glossinidius*. In addition, 58 reactions without gene-association are also included in the ancestral network based on their presence in JR904 metabolic network, that raises the number of reactions in *S. glossinidius* ancestral network to 740 internal reactions including metabolic and transport reactions (See Supplementary File 5.1). In order to define the boundaries of the ancestral network, 143 exchange fluxes included in *E. coli* JR904 metabolic network representing the potential inputs and outputs to the system were also included in the ancestral network. As is described in the introduction of this chapter, these exchange fluxes or pseudoreactions are not strict metabolic reactions, but will define the external conditions under which the system operates in order to optimize objective function. As objective function, biomass equation defined for *E. coli* JR904 metabolic network is incorporated to *S. glossinidius* ancestral network (Reed et al., 2003). The composition of biomass equation is depicted in Table 5.1, and represent a combination of the metabolic

precursors, energy and redox potential required for the production of viable phenotype in terms of cellular growth for a free-living bacterium like *E. coli* K12.

Compound	Stochiometry	Compound	Stochiometry	Compound	Stochiometry	Compound	Stochiometry
5-MTHF	0.05	dGTP	0.0254	L-methionine	0.146	Spermidine	0.007
AcCoA	0.00005	dTTP	0.0247	NAD	0.00215	Succinyl Coenzyme A	0.000003
L-Alanine	0.488	FAD	0.00001	NADH	0.00005	L-threonine	0.241
AMP	0.001	L-glutamine	0.25	NADP	0.00013	L-tryptophan	0.054
L-Arginine	0.281	L-glutamate	0.25	NADPH	0.0004	L-tyrosine	0.131
L-Asparragine	0.229	L-glycine	0.582	Phosphatidylethanolamine	0.001935	UDP-glucose	0.003
L-Aspartate	0.229	GTP	0.203	Peptidoglycan	0.0276	UTP	0.136
ATP	45.7318	H2O	45.5608	Phosphatidylglycerol	0.000464	L-valine	0.402
Cardiolipin	0.000129	L-histidine	0.09	L-phenylalanine	0.176	Glycogen	0.154
Coenzyme A	0.000006	L-isoleucine	0.276	L-proline	0.21	ADP	45.5608
CTP	0.126	L-leucine	0.428	Phosphatidylserine	0.000052	H+	45.56035
L-Cysteine	0.087	Bacterial LPS	0.0084	Putrescine	0.035	Pi	45.5628
dATP	0.0247	L-lysine	0.326	L-serine	0.205	Ppi	0.7302
dCTP	0.0254						

Table 5.1: *Metabolites included in the biomass equation of E. coli JR904 metabolic networks (Reed et al., 2003)*

S. glossinidius functional network is inferred from ancestral metabolic network removing 181 reactions catalyzed by the 148 pseudogenes and 41 deleted genes. The procedure to remove reactions is the same as described above for ancestral network reconstruction:

1) Reactions catalyzed by a single enzyme encoded by a single pseudogene or deleted gene are removed from functional network. 25 reactions catalyzed by deleted genes and 121 reactions catalyzed by pseudogenes are removed from *S. glossinidius* functional network.

2) Reactions catalyzed by a single enzymatic complex where a deleted gene or a pseudogene is involved are removed from functional network. 28 reactions catalyzed by enzymatic complexes that includes deleted genes and pseudogenes are removed from *S. glossinidius* functional network.

3) Reactions catalyzed by different isozymes or isoenzymatic complexes where all the possible isoforms are encoded by deleted genes or pseudogenes are removed from functional network. 7 reactions where all isoforms correspond to deleted genes and pseudogenes are removed from *S. glossinidius* functional network.

This transition involves the elimination of 21 genes of *S. glossinidius* ancestral network associated with enzymatic complexes in which one of the components appears as deleted or pseudogenized. Finally, we obtain a functional metabolic network of *S. glossinidius* composed of 458 genes, 560 internal reactions including metabolic and transport reactions and biomass equation, and 143 exchange fluxes that defines the system boundaries as in *S. glossinidius* ancestral network (See Supplementary File 5.2).

Metabolic networks are represented in SBML format (System Biology Markup Language), an XML-based format designed for computational models of metabolic networks that is accepted as the standard for representation of such models (Hucka et al., 2003). The SBML files generated in this work for each *S. glossinidius* metabolic network (ancestral and functional) contain the following types of data following the standards outlined in (http://sbml.org/Documents/Specifications):

- **Compartments:** Defines the bounded space in which the metabolites are located. In the present study, two different compartments are defined as *"Cytosol"* and *"Extra_organism"* in the context of the network reconstruction that represents metabolic reactions and transport processes within the cell.

- **Metabolites (Format 'M_<metab_abbreviation>_<compartment>')**: Includes metabolite name, compartment, charge, and formula. Metabolites are listed in the <*listOfSpecies*> section of the SBML file in the format described in Figure 5.3A. If a particular metabolite is present in multiple compartments (for example, glucose is captured by cells from the extracellular environment to the cytosol), a separate row for the metabolite in each compartment is defined in the metabolite list.

- **Reactions (Format 'R_<*reaction_abbreviation*>'):** Includes reaction name, reversibility, stochiometry, gene-protein-reaction association (GPR), subsystem, and E.C._number. Reactions are listed in the <*listOfReactions*> section of the SBML file in the format described in Figure 5.3B. In GPR associations, the Boolean operators AND and OR are employed to define enzymatic complexes and isozymes respectively. Upper and lower bounds across network reactions defines the range of flux values achievable across the reaction in FBA simulations..

(A)

```
<species id="M_dhf_c" name="M_7_8_Dihydrofolate_C19H19N7O6" compartment="Cytosol" charge="-2" boundaryCondition="false"/>
<species id="M_h_c" name="M_H_H" compartment="Cytosol" charge="1" boundaryCondition="false"/>
<species id="M_nadph_c" name="M_Nicotinamide_adenine_dinucleotide_phosphate___reduced_C21H26N7O17P3" compartment="Cytosol" charge="-4"
boundaryCondition="false"/>
<species id="M_nadp_c" name="M_Nicotinamide_adenine_dinucleotide_phosphate_C21H25N7O17P3" compartment="Cytosol" charge="-3"
boundaryCondition="false"/>
<species id="M_thf_c" name="M_5_6_7_8_Tetrahydrofolate_C19H21N7O6" compartment="Cytosol" charge="-2" boundaryCondition="false"/>
```

(B)

```
<reaction id="R_DHFR" name="R_dihydrofolate_reductase" reversible="true">
        <notes>
                <html:p>GENE_ASSOCIATION: SG0421</html:p>
                <html:p>PROTEIN_ASSOCIATION: FolA</html:p>
                <html:p>SUBSYSTEM: S_Cofactor_and_Prosthetic_Group_Biosynthesis</html:p>
                <html:p>PROTEIN_CLASS: 1.5.1.3</html:p>
        </notes>
        <listOfReactants>
                <speciesReference species="M_dhf_c" stoichiometry="1.000000"/>
                <speciesReference species="M_h_c" stoichiometry="1.000000"/>
                <speciesReference species="M_nadph_c" stoichiometry="1.000000"/>
        </listOfReactants>
        <listOfProducts>
                <speciesReference species="M_nadp_c" stoichiometry="1.000000"/>
                <speciesReference species="M_thf_c" stoichiometry="1.000000"/>
        </listOfProducts>
        <kineticLaw>
                <math xmlns="http://www.w3.org/1998/Math/MathML">
                        <apply>
                                <ci> LOWER_BOUND </ci>
                                <ci> UPPER_BOUND </ci>
                                <ci> OBJECTIVE_COEFFICIENT </ci>
                                <ci> REDUCED_COST </ci>
                        </apply>
                </math>
                <listOfParameters>
                        <parameter id="LOWER_BOUND" value="-999999.000000" units="mmol_per_gDW_per_hr"/>
                        <parameter id="UPPER_BOUND" value="999999.000000" units="mmol_per_gDW_per_hr"/>
                        <parameter id="OBJECTIVE_COEFFICIENT" value="0.000000"/>
                        <parameter id="REDUCED_COST" value="0.000000"/>
                </listOfParameters>
        </kineticLaw>
</reaction>
```

Figure 5.3: *Example of a metabolic reaction included in S. glossinidius ancestral and functional metabolic network in SBML format. The reaction corresponds to the reduction of dihydrofolate to tetrahydrofolate in tetrahydrofolate biosynthesis pathway catalyzed by dihydrofolate reductase encoded by the gene folA. In (A) is represented the metabolites included in the reaction, and in (B) is represented the corresponding reaction block.*

Flux Balance Analysis simulations over *S. glossinidius* ancestral and functional networks in SBML format using biomass production as objective function are carried out with the COBRA toolbox (Becker et al., 2007) within the MATLAB numerical computation and visualization environment (http://www.mathworks.com/) using linear optimization algorithm provided by LP_solve toolkit (http://sourceforge.net/projects/lpsolve/). The metabolic network of *E. coli* K12 JR904 is included in the analysis for comparative purposes with network behavior of ancestral and functional networks of *S. glossinidius* with the same biomass equation as objective function.

5.2.3 Robustness analysis

Network robustness over *E. coli* K12 JR904 and *S. glossinidius* (ancestral and functional) metabolic networks is evaluated in two different ways. First, the effect of reducing the metabolic flux through a single reaction on cellular growth is evaluated for reactions involved in the central metabolic pathway of the glycolysis over the three metabolic networks. This analysis allows evaluating how a particular objective of interest, in this case cellular growth defined as biomass production, changes as the flux over a specific reaction of interest varies in magnitude. This is carried out with the function *robustnessAnalysis* of the COBRA toolbox. Second, the effect of gene deletion on cellular growth is also compared over the three metabolic networks with the function *singleGeneDeletion* of the COBRA toolbox, limiting the flux over the reactions catalyzed by the deleted gene to zero and inferring cellular growth through FBA maximizing biomass production.

5.2.4 Reductive evolution simulations

In order to predict the possible future evolution of *S. glossinidius* in the evolutionary context of the reductive evolution process, simulations of reductive evolution were carried out with *S. glossinidius* functional metabolic network under different environmental conditions. Starting with the functional metabolic network of *S. glossinidius*, a randomly chosen reaction is removed from the network by setting the metabolic flux through this reaction to zero, and the impact of the deletion on the biomass production rates over the reduced network is evaluated by FBA. If the biomass production rate over the deleted network is above a given cutoff, the reaction is considered as non-essential and is removed permanently together with their corresponding genes from *S. glossinidius* functional network. In contrast, if the biomass production rate over the deleted network is below a given cutoff, we consider the reaction as essential and it´s retained together with their corresponding genes in the reduced network. This procedure is repeated until all network reactions have been evaluated. Three different cutoffs in terms of the biomass production rate for defining a reaction as essential or non-essential are evaluated (0.1, 0.01 and 0.05 of the biomass production rate of original network),

and for each cutoff 500 reductive evolution simulations are carried out. In order to have independent evolutionary outcomes in each simulation, the vector of network reactions is randomly permuted in each reductive evolution simulation.

This experiment is carried out under two different environmental conditions in terms of the external metabolites available for uptake, nutrient-limited conditions with only glucose and arginine as external metabolites, and nutrient-rich conditions allowing the influx of 41 metabolites present inside an insect cell and for which transport reactions are present in the functional metabolic network of *S. glossinidius* (see Table 5.2). In both environmental conditions, maximum oxygen uptake rate is fixed to 20 mmol gr. DryWeigth^{-1} hr^{-1} by fixing the lower bound of their exchange reaction to -20 mmol gr.DryWeight^{-1} hr^{-1}, whereas unconstrained uptake for ammonia, water, phosphate, sulfate, potassium, sodium, iron (II), carbon dioxide and protons is allowed. This is achieved by fixing lower and upper bounds to their corresponding exchange fluxes to $1x10^{30}$ and $-1x10^{30}$ mmol gr. DryWeigth^{-1} hr^{-1} respectively. All other external metabolites, except for the carbon source and additional metabolites under nutrient-rich condition simulations, were only allowed to leave the system under minimal conditions by fixing the lower bounds of their corresponding exchange reaction to zero. At the end of the simulations, 1500 minimal networks in each condition were obtained. From each minimal network, its corresponding genes were extracted and the gene content of minimal networks was compared in order to identify two different gene sets. Genes present in all minimal networks in each condition, that can be considered as essential genes for the functionality of the metabolic system, and non-essential genes or disposable genes that correspond to genes absent in all minimal networks, that can be removed from the network without affecting the functionality of the system in terms of biomass production.

The gene content of simulated minimal networks was compared with the gene content of *W. glossinidia* in terms of the fraction of positive hits, specificity and sensitivity. Sensitivity is defined as the fraction of essential genes in all minimal networks present in the genome of *W. glossinidia* and specificity is defined as the fraction of disposable genes in all minimal networks absent in the genome of *W. glossinidia*. The fraction of positive hits represents the fraction of true positives (essential genes present in the genome of *W. glossinidia*) and true negatives (disposable genes absent in the genome of *W. glossinidia*) over the total number of essential and disposable genes in all minimal networks in each condition

Component
4-Aminobutanoate
Acetaldehyde
Acetate
Adenosine
Cytidine
Deoxyadenosine
D-Fructose
D-Galactose
D-Glucose
D-Mannose
Ethanol
Formate
Glycerol
Glycine
Guanine
Hypoxanthine
Lactose
L-Alanine
L-Arginine
L-Asparagine

Component
L-Aspartate
L-Cysteine
L-Fucose
L-Glutamate
L-Histidine
L-Isoleucine
L-Leucine
L-Lysine
L-Methionine
L-Phenylalanine
L-Serine
L-Tryptophan
L-Tyrosine
L-Valine
NAD
NMN
Pyruvate
Thymidine
Urea
Uridine
Xanthine

Table 5.2: *Metabolites available for uptake in reductive evolution simulations under nutrient-rich conditions.*.

5.2.5 Evolutionary analysis over essential and disposable genes in minimal metabolic networks

In order to evaluate if the essential character of genes present in *S. glossinidius* functional network based on reductive evolution simulations are correlated with differential patterns of sequence evolution, the values of Codon Adaptation Index (CAI) for *S. glossinidius* genes and the number of synonymous substitutions per synonymous site (*dS*) and nonsynonymous substitutions per nonsynonymous site (*dN*) in essential an disposable genes under different external conditions for the evolutionary lineage comprised between *S. glossinidius* and *E. coli* K12 are compared. To carry out these analysis, CAI for *S. glossinidius* genes were extracted from HEG-DB based on the codon usage of a set of highly expressed genes that are identified based on an iterative process starting with the set of genes encoding

ribosomal proteins (Puigbo et al., 2008b). For the inference of *dS* and *dN* values for each gene, orthologous genes between *S. glossinidius* and *E. coli* K12 were extracted based on reciprocal best-hits on FASTA searches (Pearson, 1990), their amino acid sequences were aligned with ClustalW program with the default options (Larkin et al., 2007) and the codon alignments were obtained from the amino acid alignment and the corresponding nucleotide sequences with the program TRANALIGN from the EMBOSS package (Rice et al., 2000). Pairwise *dN* and *dS* estimates between *S. glossinidius* and *E. coli* K12 orthologous genes were calculated by the approximate method of Yang and Nielsen (Yang and Nielsen, 2000) using the *yn00* program in the PAML version 4.3 software package (Yang, 2007). We tested for statistically significant differences between the mean of CAI, *dN*, and *dS* values in essential and non-essential genes defined by the reductive evolution simulations under different external conditions. Correlation between CAI and pairwise *dN* and *dS* estimates were also analyzed to test if there is any relation between the expression level of a network gene (associated with CAI values) and the conservation of amino acid sequences (approximated by *dN* values), and to test the degree of purifying selection on synonymous sites indicative of the degree of codon usage bias (Sharp and Li, 1987a; Sharp and Li, 1987b; Sharp, 1991).

Pairwise estimates of *dN* and *dS* between *S. glossinidius* and *E. coli* K12 are calculated for the whole evolutionary branch comprised between both species. In order to test if conclusions inferred from pairwise estimates are reproduced in each independent lineage since their divergence from their common ancestor, specific *dN* and *dS* values for the evolutionary branches of *S. glossinidius* and *E. coli* K12 were calculated based on the incorporation of an external outgroup common to both species. We choose *Vibrio cholerae* O1 (accession number NC_002505) as external outgroup of *S. glossinidius* and *E. coli* K12, based on phylogenetic reconstruction obtained in Chapter 3 of this thesis. Orthologous genes between *S. glossinidius*, *E. coli* K12 and *V. cholerae* O1 were extracted with OrthoMCL based on BLASTP searches with a maximum e-value cutoff of 10^{-20} and a minimum amino acid identity of 80%. Codon alignments and *dN* and *dS* values were obtained as described above for *S. glossinidius* and *E. coli* K12 orthologous pairs. In order to calculate specific *dN* and *dS* values for *S. glossinidius* (*A*) and *E. coli* K12 (*B*) lineages since the divergence from their common ancestor (*O*), *dN* and *dS* values between *S. glossinidius* and *E.coli* K12 are considered as the sum of substitutions accumulated in the evolution of *S. glossinidius* since their divergence from the ancestor *O* ($dN_{sgl\text{-}O}$; $dS_{sgl\text{-}O}$) and substitutions accumulated in the evolution of *E.coli* K12 since their divergence from the same common ancestor *O* ($dN_{eco\text{-}O}$ and $dS_{eco\text{-}O}$) (Figure 5.4).

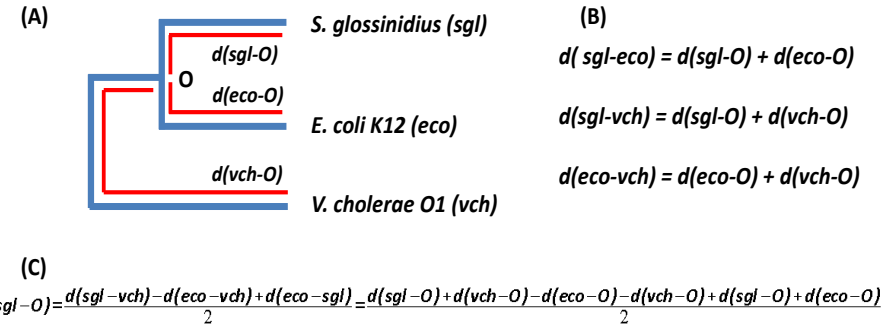


Figure 5.4: *Inference of lineage-specific dN and dS values for S. glossinidius and E. coli K12 since the divergence from their common ancestor O based on the introduction of V. cholerae O1 as common outgroup (A). Pairwise estimates of nucleotide substitutions per site (synonymous or non-synonymous) can be decomposed as the sum of the number of substitutions taking place in each individual lineage since their divergence from the common ancestor O (B). An example of the calculation of substitution distances in the lineage of S. glossinidius since the divergence from O based on combination of pairwise distances is represented in (C).*

Following this reasoning and considering *V. cholerae* O1 as external outgroup of *S. glossinidius* and *E. coli* K12, specific *dN* and *dS* values for *S. glossinidius* and *E. coli* K12 lineages can be calculated with the following formulas:

$$dN_{sgl} = \frac{dN_{(sgl-vch)} - dN_{(eco-vch)} + dN_{(eco-sgl)}}{2}$$

$$dNeco = \frac{dN_{(eco-vch)} - dN_{(sgl-vch)} + dN_{(eco-sgl)}}{2}$$

$$dS_{sgl} = \frac{dS_{(sgl-vch)} - dS_{(eco-vch)} + dS_{(eco-sgl)}}{2}$$

$$dSeco = \frac{dS_{(eco-vch)} - dS_{(sgl-vch)} + dS_{(eco-sgl)}}{2}$$

5.2.6 Statistical analysis

Statistical analysis to compare CAI, *dN* and *dS* values between different gene sets were carried out with the program PASW statistics v17.

5.3 RESULTS

5.3.1 The metabolic networks of *S. glossinidius* at different stages of the genome reduction process

S. glossinidius represents one of the few examples of a completely sequenced bacterial genome at initial stages of the genome reduction process. In the present chapter, the ancestral and functional metabolic networks of *S. glossinidius* have been reconstructed based on the re-annotation of genes and pseudogenes carried out in the previous chapter of this thesis. For comparative purposes and to facilitate the interpretation of the results, the metabolic network of *E. coli* JR904 is taken as reference network of a free-living bacterium. The reconstruction process can be summarized as follows: (1) An initial mapping of orthologous genes between *S. glossinidius* ancestral proteome (genes and pseudogenes) and *E. coli* JR904 genes leading to a first backbone of the ancestral metabolic network of S. glossinidius; (2) Identification of deletion events of metabolic genes during the evolution of *S. glossinidius* since their divergence from a free-living ancestor by means of comparative genomics; (3) Model completion and validation stage that includes the incorporation of *S. glossinidius*-specific genes and pseudogenes related to metabolism and without homology at sequence level with *E. coli* K12 and validation of network reconstructions by means of FBA using biomass production as objective function. The final ancestral metabolic network of *S. glossinidius* contains 668 gene products, 741 internal reactions including transport processes, cytoplasmic reactions, and the biomass equation used as objective function, and 690 metabolites, of which 547 are cytoplasmic and 143 are extracellular. Of the 668 gene products, 479 correspond to functional genes, 148 correspond to pseudogenes, and 41 correspond to genes deleted during the evolution of *S. glossinidius* from their free-living ancestor. The 627 CDSs represents 16% of the total number of CDS (genes and pseudogenes) of the genome, which is in the range of other genome-scale metabolic network reconstructions (See Table 5.3). Of the 741 internal reactions included in the network, 683 reactions (92.17 %) have at least one assigned gene, pseudogene, or deleted gene in the gene-protein-reaction (GPR) relationships. 58 reactions non gene-associated in *E. coli* JR904 metabolic network were also incorporated to *S. glossinidius* ancestral network.

Table 5.3: *Main compositional features of metabolic networks available from http://gcrg.ucsd.edu/In_Silico_Organisms/Other_Organisms . The number of genes, metabolites and reactions are included. S. glossinidius ancestral and functional networks are incorporated at the bottom of bacterial metabolic networks. E. coli K12 JR904 and S. glossindius ancestral and functional networks analyzed in this study are highlighted*

Organism	Total genes	Model genes	% of the genome represented	Nº metabolites	Nº reactions	Reference
Sodalis glossinidius ancestral	**3932 (*1)**	**668 (*2)**	**15.95**	**547**	**741**	
Sodalis glossinidius functional	**2431**	**458**	**18.84**	**481**	**560**	
Acinetobacter baumanii AYE	3760	650	17.29	778	891	(Kim et al 2010)
Acinetobacter baylyi ADP1	3287	774	23.55	701	875	(Durot et al 2008)
Bacillus subtilis	4114	844	20.52	988	1020	(Oh et al 2007)
Bacillus subtilis	4114	534	12.98	456	563	(Goelzer et al 2008)
Bacillus subtilis	4114	1103	26.81	1138	1437	(Henry et al 2009)
Buchnera aphidicola APS	574	196	34.15	240	263	(Thomas et al 2009)
Clostridium acetobutylicum ATCC 824	3848	474	12.32	422	552	(Senger and Papoutsakis 2008)
Clostridium acetobutylicum ATCC	3848	432	11.23	479	502	(Lee et al 2008)
Corynebacterium glutamicum	3002	227	7.56	423	502	(Shinfuku et al 2009)
Escherichia coli K12 MG1655	4405	660	14.98	438	627	(Edwards and Palsson 2000)
Escherichia coli K12 MG1655	**4405**	**904**	**20.52**	**625**	**931**	**(Reed et al 2003)**
Escherichia coli K12 MG1655	4405	1260	28.60	1039	2077	(Feist et al 2007)
Geobacter metallireducens	3532	747	21.15	769	697	(Sun et al 2009)
Geobacter sulfurreducens	3530	588	16.66	541	523	(Mahadevan et al 2006)
Haemophilus influenzae Rd	1775	296	16.68	343	488	(Edwards and Palsson 1999)
Haemophilus influenzae Rd	1775	400	22.54	451	461	(Schilling and Palsson 2000)
Helicobacter pylori 26695	1632	291	17.83	340	388	(Schilling et al 2002)
Helicobacter pylori 26695	1632	341	20.89	485	476	(Thiele et al 2005)
Lactobacillus plantarum WCFS1	3009	721	23.96	531	643	(Teusink et al 2006)
Lactococcus lactis ssp. Lactis	2310	358	15.50	422	621	(Oliveira et al 2005)
Mannheimia succiniciproducens	2384	335	14.05	332	373	(Hong et al 2004)

Table 5.3 (Continuation)

Organism	Total genes	Model genes	% of the genome represented	Nº metabolites	Nº reactions	Reference
Mannheimia succiniciproducens	2384	425	17.83	519	686	(Kim et al 2007)
Mycobacterium tuberculosis H37Rv	4402	661	15.02	828	939	(Jamshidi and Palsson 2007)
Mycobacterium tuberculosis H37Rv	4402	726	16.49	739	849	(Beste et al 2007)
Mycoplasma genitalium G-37	521	189	36.28	274	262	(Suthers et al 2009)
Neisseria meningitidis serogroup B	2226	555	24.93	471	496	(Baart et al 2007)
Pseudomonas aeruginosa PA01	5640	1056	18.72	760	883	(Oberhardt et al 2008)
Pseudomonas putida KT2440	5350	746	13.94	911	950	(Nogales et al 2008)
Pseudomonas putida KT2440	5350	815	15.23	886	877	(Puchalka et al 2008)
Rhizobium etli CFN42	3168	363	11.46	371	387	(Resendis-Antonio et al 2007)
Rhodoferax ferrireducens	4770	744	15.60	790	762	(Risso et al 2009)
Salmonella typhimurium LT2	4489	1083	24.13	774	1087	(Raghunathan et al 2009)
Salmonella typhimurium LT2	4489	945	21.05	1036	1964	(AbuOun et al 2009)
Staphylococcus aureus N315	2588	619	23.92	571	641	(Becker and Palsson 2005)
Staphylococcus aureus N315	2588	551	21.29	604	712	(Heinemann et al 2005)
Staphylococcus aureus N315	2588	546	21.10	1431	1493	(Lee et al 2009)
Streptococcus thermophilus	1889	429	22.71		522	(Pastink et al 2009)
Streptomyces coelicor A3(2)	8042	700	8.70	500	700	(Borodina et al 2005)
Thermatoga maritima MSB8	1917	478	24.93	503	562	(Zhang et al 2009)
Yersinia pestis 91001	4037	818	20.26	825	1020	(Navid and Almaas 2009)

***1: 2431 genes and 1501 pseudogenes**
***2: 479 genes, 148 pseudogenes, 41 deleted genes**

S. glossinidius ancestral network includes 27 pseudogenes that have no sequence similarity with *E. coli* JR904 genes but that are related with precise metabolic functions in terms of reaction stochiometry based on the results of the functional-reannotation carried out in chapter 4 of this thesis. These pseudogenes and their corresponding reactions are represented in Supplementary Table 5.3. These *S. glossinidius* specific pseudogenes involved the addition of 7 reactions to ancestral metabolic network absent in the metabolic network of *E. coli* JR904 representing characteristic features of the ancestral metabolism of *S. glossinidius*. This includes the capability of growth under formate as carbon sources by the presence of two different pseudogenes (ps_SGL0061c and ps_SGL0873) encoding formate dehydrogenases (EC:1.2.1.2; K00122) that are incorporated as isozymes of reaction R_FDH in ancestral network, the capability of growth also under glycerol as carbon source by the presence of a pseudogene (ps_SGL1175) encoding a glycerone phosphotransferase (EC:2.7.1.29; K00863) absent in *E. coli* JR904 network that combined with pseudogenized glycerol dehydrogenase (ps_SGL0095c) allows the cells to growth anaerobically on glycerol as carbon source by their two-steps conversion to the glycolytic intermediate dihydroxyacetone phosphate (Rush et al., 1957), or the presence of a pseudogene (ps_SGL1185c) that encodes a homoserine-O-acetyltransferase (EC:2.3.1.31; K00641) that catalyzes the acetylation of L-homoserine with AcCoA to produce O-acetyl-L-homoserine in L-methionine biosynthesis in fungi and many clinically important bacterial species like *Pseudomonas aeruginosa*, *Haemophilus influenzae*, and *Mycobacterium tuberculosis* (Mirza et al., 2005). This reaction is equivalent to that catalyzed by homoserine O-succinyltransferase encoded by the gene *metA* in the first step of L-methionine biosynthesis from L-homoserine in most enteric bacteria that is also functional in *S. glossinidius* (gene SG2152). In this reaction succinyl group of succinyl CoA is transferred to L-homoserine to produce O-succinyl-L-homoserine, and has been reported that the next enzyme of methionine biosynthesis pathway, Cystathionine synthase encoded by the gene *metB*, is able to synthesize methionine precursor cystathionine from both homoserine esters, although the enzyme reacts slowly with O-acetyl-L-homoserine (Nagai and Flavin, 1967; Greene, 1996). In order to consider this possibility in the ancestral genome, two additional reactions are incorporated to the ancestral metabolic network corresponding to homoserine O-acetyltransferase associated to pseudogene ps_SGL1185c (R_HSAT) and O-acetylhomoserine lyase associated to gene SG2152 for L-cystathionine biosynthesis from O-acetyl-L-homoserine (R_AHSL1). In other cases, *S. glossinidius* specific pseudogenes corresponds to different isoenzymatic forms of a functional enzyme, like the presence of three different pseudogenes (ps_SGL0103c, ps_SGL0919c, ps_SGL1476c) encoding isozymes of dihydropicolinate synthase (*dapA* gene) responsible of the first step of L-lysine biosynthesis from L-aspartate (EC:4.2.1.52; K01714) that are added as isozymes together with the functional gene (SG1726) in reaction R_DHDPS of the ancestral metabolic network, a pseudogene (ps_SGL0722) encoding an isozyme of agmatinase (EC:3.5.3.11; K01480) involved

in putrescine biosynthesis from L-arginine that is incorporated as isozyme of the functional gene (SG2071) in R_AGMT of ancestral network, or a pseudogene (ps_SGL0797) encoding the isozyme phosphofructokinase II (EC:2.7.1.11; K00850) involved in the glycolytic pathway that is incorporated as isozyme of the functional phosphofructokinase I (SG2178) in R_PFK of ancestral network. The metabolic network of *E. coli* JR904 includes the two isozymes of phosphofructokinase (*pfkA* and *pfkB* genes) associated to the phosphorylation of fructose-6-phosphate to fructose-1,6-biphosphate during the glycolysis, and although *pfkA* gene encoding the major isozyme phosphofructokinase I have a functional ortholog in *S. glossinidius* genome (SG2178), the pseudogene ps_SGL0797 have no homology at sequence level with *pfkB* gene of E. coli, showing homology only with phosphofructokinases of *Erwinia carotovora*, *Klebsiella pneumoniae* and *Serratia marsecens* in FASTA and BLASTP searches.

In the transition to the functional metabolic network of *S. glossinidius*, all reactions catalyzed by enzymes encoded by the 148 pseudogenes and the 41 deleted genes were removed from the ancestral network. This renders a functional metabolic network composed by 458 gene products, 560 internal reactions including transport processes, cytoplasmic reactions and the biomass equation used as objective function, and 624 metabolites, of which 481 are cytoplasmic and 143 are extracellular (see Table 5.3). Of the 560 internal reactions, 502 have assigned at least one gene in GPR associations (90%), whereas the 58 reactions without associated gene included in the ancestral network and in *E. coli* JR904 network were maintained in the functional network of *S. glossinidius*. The 458 genes included in the functional network represents 18.84% of the total number of genes of *S. glossinidius* genome, which is in the range of other metabolic reconstructions and higher than the fraction of CDS included in the ancestral metabolic network (see Table 5.3). In the transition to the functional metabolic network, 21 functional genes included in the ancestral network are removed due to their inclusion in enzymatic complexes where at least one of the components was encoded by a pseudogene or a deleted gene. 13 genes of *S. glossinidius* that were initially not mapped with OrthoMCL were incorporated to both the ancestral and functional metabolic networks of *S. glossinidius* based on the information of functional re-annotation (see Supplementary Table 5.4). Most of these genes are associated to reactions present in the metabolic network of *E. coli* JR904 that are not identified initially because their corresponding genes in *S. glossinidius* and *E. coli* K12 show no homology at sequence level. This is the case of the genes encoding the TCA cycle enzyme isocitrate dehydrogenase (EC:1.1.1.42; K00031), with no homology at sequence level between the corresponding genes in *S. glossinidius* (SG0700) and *E. coli* K12 (b1136). Based on subunit structure, bacterial isocitrate dehydrogenases can be broadly classified into monomeric and homodimeric enzymes, with significant homology at sequence level within each group but with no similarity between groups (Eikmanns et al., 1995; Sahara et al., 2002). *S. glossinidius* isocitrate

dehydrogenase shows higher identity with monomeric enzymes like that of Pseudomonadales *Azotobacter vinelandii* or *Pseudomonas syringae pv. Tomato DC3000*, whereas most enteric bacteria including *E. coli* K12 have homodimeric forms of isocitrate dehydrogenase, that explains the absence of this enzymatic activity in the initial mapping of orthologous genes between *S. glossinidius* and *E. coli* K12. Another example is the genes encoding glycosyl hydrolases responsible of sugar addition to the core oligosaccharide in bacterial lipopolysaccharide (LPS) biosynthesis. In *E. coli* JR904 metabolic network, this reaction is associated with genes *rfaC*, *rfaF*, *rfaG*, *rfaL*, *rfaJ* and *rfaI* (b3621, b3620, b3631, b3622, b3626, and b3627 respectively), that are organized in two oppositely transcribed operons in *E. coli* K12 genome (Raetz, 1996). *S. glossinidius* retains orthologs at sequence level for *rfaC* and *rfaF* genes (SG2195 and SG2194) that encode heptosyltransferases for the sequential addition of two molecules of ADP-L-glyceromanno-heptose to the KDO-Lipid A molecule, and for *rfaG* gene (SG2202) that encodes a LPS-glycosyltransferase involved in the next steps of nucleotide sugar addition to the core oligosaccharide. For the rest of *rfa* genes there is no similarity at sequence level although are located in a syntenic region between both genomes and corresponds to genes encoding glycosyl-hydrolases (see Figure 5.5), that are indicative of differences in composition of core oligosaccharide between both genomes, a feature that has been also observed even between closely related *E. coli* strains and *Salmonella* (Heinrichs et al., 1998). In order to cope with these differences, *S. glossinidius* genes SG2197, SG2199 and SG2200 are incorporated together with SG2194, SG2195 and SG2202 to the reaction for bacterial LPS biosynthesis (R_LPSSYN in ancestral and functional metabolic networks).

Other genes incorporated to the functional and ancestral networks of *S. glossinidius* are those with orthology with *E. coli* K12 genes that were not present in *E. coli* JR904 metabolic network. This is the case of two genes (SG0528 and SG1603) that encodes broad-specificity nucleosidases *YbfR* and *SurE* respectively present also in *E. coli* K12 (b2291 and b2744 respectively). *YbfR* and *SurE*, together with similar nucleotidase encoded by the gene *ushA* (b0480), are involved in the degradation of nucleotides to phosphate and their corresponding nucleosides (Proudfoot et al., 2004). In *E. coli* JR904 network there is 11 reactions corresponding to nucleotidases that are associated to *ushA* gene, that is absent in *S. glossinidius*. Based on the broad substrate specificity characterized for *SurE* and *YbfR* nucleotidases (Proudfoot et al., 2004), the corresponding orthologs of *S. glossinidius* (SGL0528 and SGL1603) were associated to these nucleotidase reactions. In addition, *ushA* gene appears also associated to a nucleotide sugar hydrolase reaction (R_USHD) by which the nucleotide sugar UDP-2,3-bis (3-hydroxymyristoyl) glucosamine is hydrolyzed producing 2,3-bis (3-hydroxymyristoyl)-β-D-glucosaminyl 1-phosphate, also known as Lipid X, that is the precursor of the Lipid A subunit of bacterial lipopolysaccharide. However, this reaction has been experimentally demonstrated that is catalyzed by a specific UDP-

2,3-diacylglucosamine hydrolase encoded by the gene *lpxH* (b0524) in *E. coli* (Babinski et al., 2002a; Babinski et al., 2002b). *S. glossinidius* retains a functional ortholog of *lpxH* (SG0703), so this gene is associated to R_USHD reaction in both ancestral and functional metabolic networks.

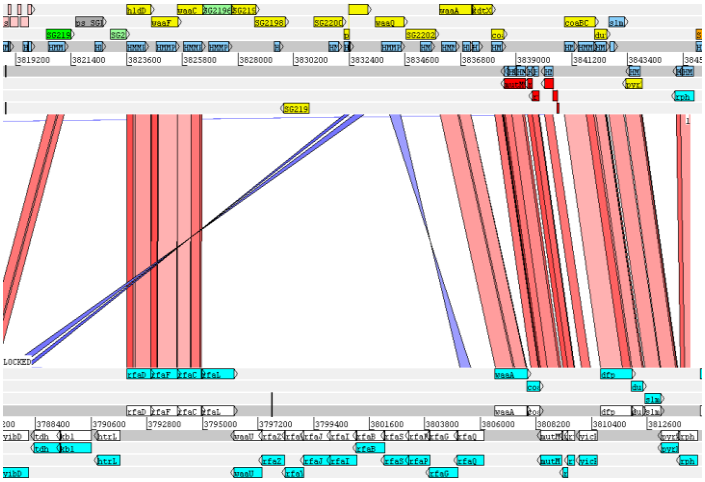
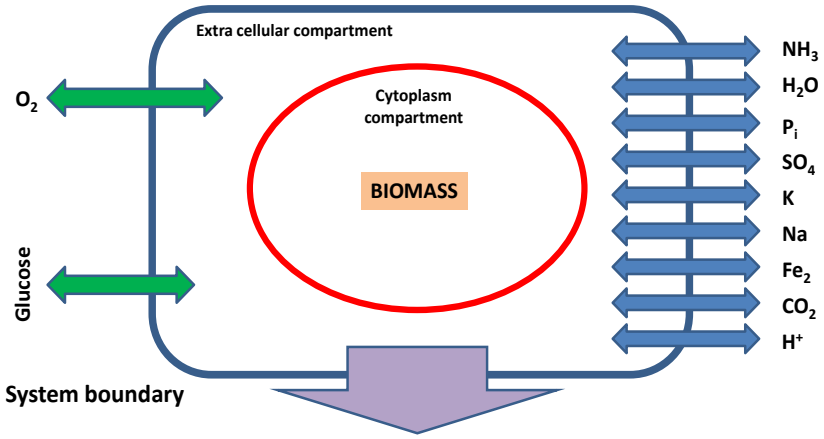


Figure 5.5: *ACT comparison between S. glossinidius and E. coli K12 in the region corresponding to rfa genes involved in bacterial LPS biosynthesis. Orthology relationships can be inferred for rfaC (b3621-SG2195), rfaF (b3620-SG2194) and rfaG (b3631-SG2202) based on OrthoMCL. S. glossinidius specific glycosyl hydrolases included in ancestral and functional metabolic networks associated to bacterial LPS biosynthesis are encoded by SG2197, SG2199, and SG2200*

5.3.2 Analysis of *S. glossinidius* metabolic networks through FBA: Transitions to host-dependent lifestyle

In order to evaluate the functionality of *S. glossinidius* ancestral and functional metabolic networks in comparison with the reference network of a free-living organism like *E. coli* K12, the *in-silico* biomass yield of these metabolic networks under different minimal media in steady-state conditions was calculated using FBA. As objective function, the biomass equation formula designed for *E. coli* JR904 metabolic network is incorporated to both ancestral and functional metabolic networks of *S. glossinidius*. The composition of the biomass equation is described in Table 5.1, and reflects the basic requirements in terms of essential metabolites (amino acids, nucleotides, phospholipids, cofactors) in order to ensure cell survival of a free-living bacterium like *E. coli* K12. The recent transition of *S. glossinidius* to

a host-dependent lifestyle can be tested evaluating the functionality of their ancestral and functional metabolic networks with a biomass equation defined for a free-living bacteria like *E. coli* K12 in a minimal medium with only glucose as external carbon source from which to produce all biomass constituents. In order to simulate this minimal medium the upper and lower bounds of exchange reactions defining the boundaries of the system must be fixed (Figure 5.6).



	Lower bound	Upper bound
NH_3, H_2O, P_i, SO_4, K, Na, Fe_2, CO_2, H^+	-1x10^{30} mmol gr. DryWeight^{-1} hr^{-1}	1x10^{30} mmol gr. DryWeight^{-1} hr^{-1}
O_2	-20 mmol gr. DryWeight^{-1} hr^{-1}	1x10^{30} mmol gr. DryWeight^{-1} hr^{-1}
Glucose	-6 mmol gr. DryWeight^{-1} hr^{-1}	1x10^{30} mmol gr. DryWeight^{-1} hr^{-1}
Other external metabolites	0 mmol gr. DryWeight^{-1} hr^{-1}	1x10^{30} mmol gr. DryWeight^{-1} hr^{-1}

Figure 5.6: *Schematic diagram of metabolites allowed entering and exiting to the metabolic system through their corresponding exchange reactions in nutrient-limited conditions. Metabolites for which unconstrained import and export across system boundaries are allowed are represented by blue arrows (9 metabolites). Maximum uptake rates for molecular oxygen and glucose are restricted by fixing the lower bounds of their corresponding exchange reactions to -20 mmol gr. DryWeigth^{-1} hr^{-1} and -6 mmol gr. DryWeigth^{-1} hr^{-1} respectively (green arrows). The rest of external metabolites (132 metabolites) are only allowed to leave the system by fixing the lower bound of their corresponding exchange reactions to zero (purple arrow). External metabolites are first incorporated to the extracellular compartment through exchange reactions, from where they must pass to the cytoplasmic compartment through transport reactions in order to produce biomass constituents.*

If the flux over this exchange reactions have positive values, the corresponding external metabolite leaves the system, whereas if this exchange fluxes have negative

values the corresponding external metabolites enter the system and, as consequence, the set of cytoplasmic and transport reactions defined in the metabolic network can use them to produce all biomass constituents. As consequence, minimal medium with only glucose as external carbon source is simulated by fixing the lower bound of the exchange reaction for glucose to -6 mmol gr. Dry Weight^{-1} hr^{-1}, as described in *E. coli* JR904 metabolic network (Reed et al., 2003).

The results of FBA simulations over the three metabolic networks under minimal aerobic medium with only glucose as carbon source are represented in Figure 5.7. Under these conditions, the ancestral metabolic network of *S. glossinidius* shows a biomass production rate of 0.545 gr. Dry Weight (mmol Glucose)$^{-1}$, very similar to the biomass yield for *E. coli* JR904 metabolic network (0.5391 gr. Dry Weight (mmol Glucose)$^{-1}$). These results confirm the very recent transition of *S. glossinidius* from a free-living to a host-dependent lifestyle, given that its expected phenotype in terms of biomass production rates is completely equivalent to that of a typical free-living bacteria like *E. coli* K12, being able to synthesize all biomass constituents under a minimal environment with only glucose as external carbon source. By contrast, the functional network of *S. glossinidius* comprised only by functional genes shows a lethal phenotype in terms of biomass production under this minimal conditions (0 gr. Dry Weight (mmol Glucose)$^{-1}$).

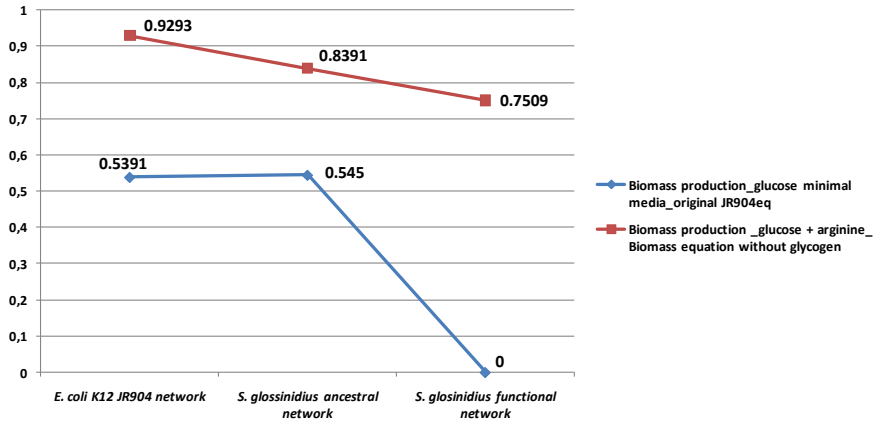


Figure 5.7: *Results of FBA simulations over E. coli K12 JR904 and S. glossinidius ancestral and functional metabolic networks in nutrient limited conditions with original biomass equation and only glucose as external carbon source (blue diagram). Under these conditions, no viable phenotype in terms of biomass production can be obtained for S. glossinidius functional network. It is necessary to add an external source of L-arginine and remove glycogen from biomass equation in order to obtain a viable phenotype for S. glossinidius functional network (red diagram).*

This is due to two main pseudogenization events affecting biosynthetic pathways that have taken place during *S. glossinidius* evolution. First, the pseudogenization of genes *glgA* and *glgB* (ps_SGL1448c and ps_SGL1450 respectively) encoding glucose 1-phosphate adenyltransferase and glycogen synthase respectively prevents glycogen biosynthesis from ATP and α-D-glucose 1-phosphate. Glycogen is a constituent of biomass equation, and as consequence its production is an essential requirement in order to obtain a viable phenotype in terms of biomass production, so the absence of these two reactions in functional network of *S. glossinidius* renders a lethal phenotype. In fact, even the elimination of these two genes over metabolic network of *E. coli* JR904 renders the same lethal phenotype. The elimination of glycogen from biomass reaction solves this problem and gives a functional phenotype in terms of biomass production. Second, the pseudogenization of genes *argA*, *argG*, *argD* and *argC* (ps_SGL1220, ps_SGL0928, ps_SGL1434, and ps_SGL1382c respectively) prevents L-arginine biosynthesis from L-glutamate. L-arginine is also a component of biomass equation, and as consequence the inactivation of L-arginine biosynthesis pathway renders a lethal phenotype in terms of biomass production over functional network of *S. glossinidius* under these minimal conditions. However, in contrast with the situation described with glycogen, removing L-arginine from biomass equation still renders a lethal phenotype in terms of biomass production. This is because the inactivation of L-arginine biosynthesis pathway has also additional consequences in the functional metabolic network of *S. glossinidius*, affecting also to the pathways of putrescine and spermidine biosynthesis. As is described in the previous chapter of this thesis, putrescine is a polyamine that can be synthesized through two different pathways, from L-arginine in two enzymatic steps catalyzed by arginine decarboxylase (*speA* gene) and agmatinase (*speB* gene) or in a single enzymatic step catalyzed by ornithine decarboxylase (*speC* gene) from the L-arginine precursor L-ornithine. From putrescine, spermidine is synthesized by condensation with S-adenosyl-L-methioninamine by spermidine synthase (*speE* gene). Putrescine and spermidine are included in the biomass equation of *E. coli* JR904, so they have to be produced by the system in order to obtain a functional phenotype in terms of biomass production. *S. glossinidius* have pseudogenized the gene *speC* for putrescine biosynthesis from L-ornithine (ps_SGL1267c), but retains functional *speA*, *speB* and *speE* genes (SG2018, SG2017 and SG0479 respectively), so the unique pathway for putrescine and spermidine biosynthesis is through exogenous L-arginine. However, in this pathway, urea is generated as by-product of the activity of agmatinase, which hydrolyzes agmatine to putrescine and urea (Satishchandran and Boyle, 1986). Urea, as by-product of the biomass constituent putrescine, is always produced by the system, and in the context of stationary-state analysis, must be consumed or exported out of the system in order to obtain a viable phenotype in terms of biomass production in FBA simulations. Urea can be metabolized to ammonia and carbon dioxide in a single enzymatic reaction catalyzed by urease enzyme complex or in

two steps catalyzed by urea carboxylase and allophanate hydrolase (Mobley and Hausinger, 1989; Lee et al., 1992; Kanamori et al., 2004), but any of this pathways are functional neither in *S. glossinidius* nor in *E. coli* K12, with *S. glossinidius* that have two pseudogenes corresponding to the alpha subunit of urease enzyme complex *UreC* (ps_SGL1048c) and to allophanate hydrolase (ps_SGL0718c), so urea has to be exported to the extracellular compartment of the system, where it is taken out of the system boundaries. In *E. coli* JR904 metabolic network, urea is exported to the extracellular compartment through a transport reaction associated with glycerol facilitator protein channel encoded by the gene *glpF* (b3927), that is able to mediate also facilitate diffusion of urea (Heller et al., 1980). In *S. glossinidius*, their corresponding ortholog by sequence and gene order comparison appears pseudogenized (ps_SGL1388), but retains a functional gene (SG1858) encoding a propanediol diffusion facilitator that shows 54.57% of identity at nucleotide level with *glpA* gene of *E. coli*, so this gene has been associated to urea exchange in ancestral and functional metabolic networks of *S. glossinidius* (Supplementary Table 5.4). When L-arginine is incorporated as external metabolite together with glucose, a functional phenotype in terms of biomass production is obtained for *S. glossinidius* functional metabolic networks (See Figure 5.7). In addition, the presence of L-arginine as external metabolite together with glucose increases the biomass production over all metabolic networks including *E. coli* JR904 and *S. glossinidius* ancestral network in comparison with the situation with only glucose as external metabolite (see Figure 5.7). A comparison of the reaction fluxes between ancestral and functional networks of *S. glossinidius* reveals that L-arginine is also necessary for the viability of the functional network through putrescine degradation to the TCA cycle intermediate succinate (Figure 5.8).

This is consequence of the pseudogenization of the gene *ppc* encoding the anaplerotic enzyme phosphoenolpyruvate carboxylase (ps_SGL1383), a key enzymatic activity in central metabolism that catalyzes the carboxylation of the glycolytic intermediate phosphoenolpyruvate to the TCA cycle intermediate oxaloacetate (R_PPC in *E. coli* K12and *S. glossinidius* ancestral network). In *E. coli* K12 and *S. glossinidius* ancestral networks with only glucose as external metabolite without external arginine, metabolic flux proceeds through PPC reaction connecting glycolytic pathway to TCA cycle, with putrescine being produced from L-ornithine by ornithine decarboxylase *SpeC*, that appears in *S. glossinidius* ancestral network as pseudogene (ps_SGL1267c). In absence of *ppC* gene, metabolic flux is redirected through the glyoxylate bypass in *E. coli* JR904 network, in concordance with experimental results in *E. coli ppC* knockout mutants (Peng et al., 2004; Peng and Shimizu, 2004) (Figure 5.8).

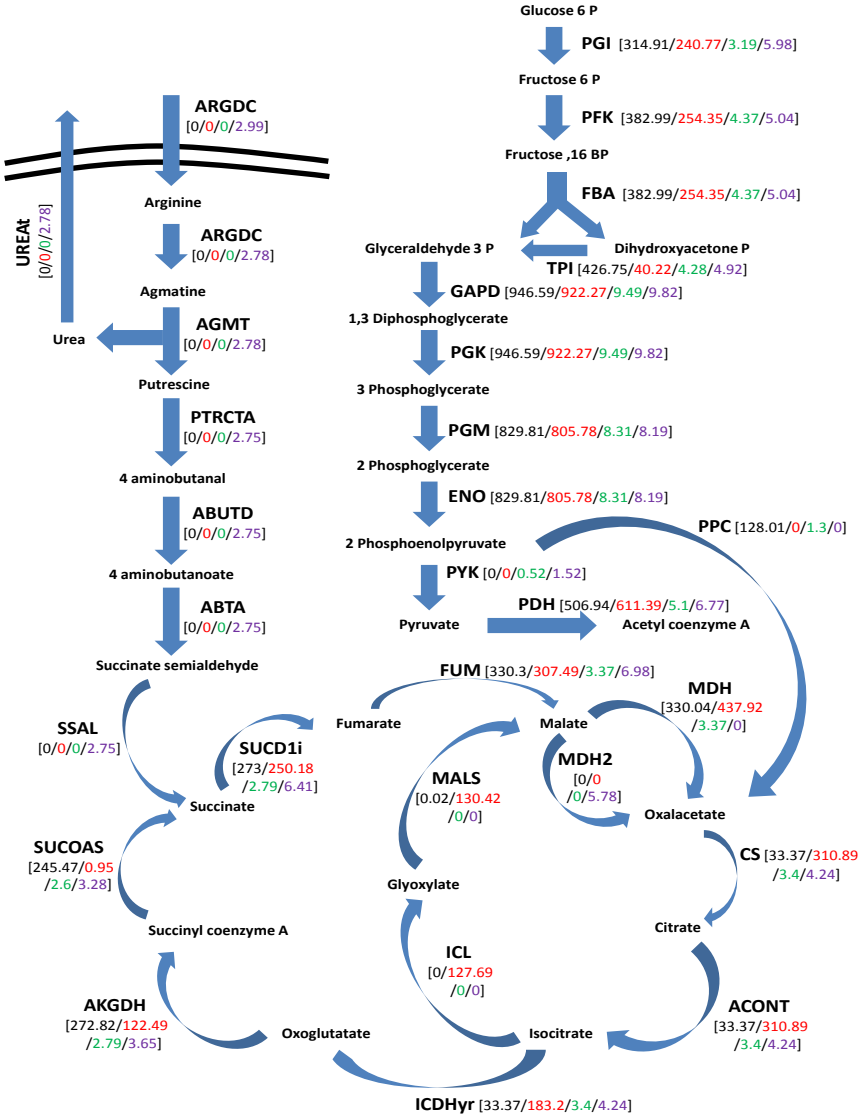


***Figure5.8**: Effects of pseudogenization of phosphoenolpyruvate carboxykinase (PPC reaction) over E. coli K12 JR904 and S. glossinidius metabolic networks. Reactions of glycolysis, TCA cycle and transport and metabolization of external L-arginine are represented together with their corresponding reaction fluxes in FBA simulations. Black values correspond to reaction fluxes in E. coli K12 JR904 network with only glucose as external carbon source. Red values correspond to reaction fluxes in E. coli JR904 network under the same conditions but removing PPC reaction. Green values correspond to reaction fluxes in S. glossinidius ancestral network with only glucose as external carbon source. Purple values correspond to reaction fluxes in S. glossinidius functional network with glucose and arginine as external metabolites. Reaction fluxes are expressed in mmol gr. DryWeigth^{-1} hr^{-1}.*

However, in *S. glossinidius* there is no signal of genes encoding isocitrate lyase and malate synthase responsible of glyoxylate bypass and, as consequence, the single deletion of *ppC* pseudogene in *S. glossinidius* ancestral network in a minimal environment with only glucose as external carbon source renders a lethal phenotype in terms of biomass production (1.5077xe-14 gr. Dry Weight (mmol Glucose)$^{-1}$). This situation is reverted adding an external source of arginine in addition to glucose, where arginine is used to produce putrescine and spermidine, and part of putrescine is degraded to the TCA cycle intermediate succinate (Figure 5.8), increasing the biomass yield in both *S. glossinidius* ancestral network and *E. coli* JR904 network in comparison with the situation with only glucose as carbon source. As consequence, L-arginine is needed for the viability of the functional metabolic network of *S. glossinidius* not only due to their role as biomass constituent and as precursor of polyamines putrescine and spermidine but also as energy supply through putrescine degradation to succinate consequence of the pseudogenization of the central metabolic enzyme phosphoenolpyruvate carboxylase and the absence of genes of the glyoxylate bypass.

In order to obtain a functional phenotype in terms of biomass production for the functional network of *S. glossinidius*, is necessary to remove glycogen from biomass equation (as consequence of the pseudogenization of genes involved in glycogen biosynthesis) and to add an external source of arginine in addition to glucose (as consequence of the pseudogenization of four of the genes involved in L-arginine biosynthesis together with the pseudogenization of *ppC* gene encoding phosphoenolpyruvate carboxylase). Fixing a lower bound of -6 mmol gr. DryWeigth^{-1} hr^{-1} for L-arginine exchange reaction (the same as glucose exchange reaction) and removing glycogen from biomass equation is it possible to obtain a viable phenotype for functional metabolic network of *S. glossinidius*, with a biomass yield of 0.7509 gr. Dry Weight (mmol Glucose)$^{-1}$ (Figure 5.7).

FBA was also used to predict the behavior of *E. coli* K12 and *S. glossinidius* metabolic networks with different minimal media in aerobic conditions differing uniquely in the external carbon source. The objective was to compare the *in-silico* results of FBA simulations with experimental data from *S. glossinidius* cell cultures (Dale and Maudlin, 1999). 27 different carbon sources were tested using the same conditions as described above in glucose simulations. In each case, the lower bound of their corresponding exchange reaction and L-arginine exchange reaction is fixed to 6 mmol gr. DryWeigth^{-1} hr^{-1}, due to the non-functionality of *S. glossinidius* ancestral network in absence of external source of L-arginine. As negative control, FBA simulations with only arginine as external metabolite are also carried out. The results of these simulations are represented in Figure 5.9.

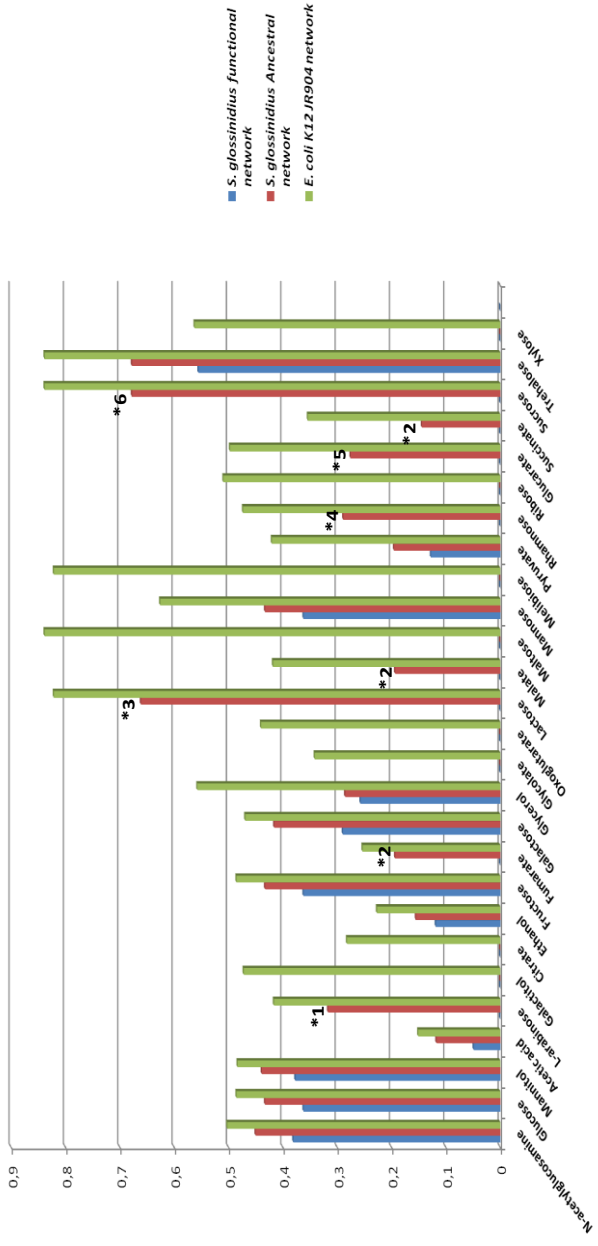


***Figure 5.9**: FBA simulations with E.coli K12 JR904 and S. glossinidius metabolic networks in nutrient-limited conditions with L-arginine and different carbon sources. The differences in biomass production rates in external conditions with arginine and the corresponding carbon source and external conditions with only arginine as external metabolites are represented in order to reflect the capability of metabolic systems to metabolize the corresponding carbon source.*
**1: Pseudogenization of arabinose ABC transport system (ps_SGL0829 (araF), ps_SGL0830 (araG), ps_SGL0831 (araH))*
**2: Deletion of dcuA gene (b4138) encoding dicarboxylate transporter: Associated to transport of fumarate, malate and succinate*
**3: Pseudogenization of β-galactosidase (ps_SGL1067 (lacZ)) and β-D-glucoside glucohydrolase (ps_SGL0469 (bglX))*
**4: Pseudogenization of rhamnose transporter (ps_SGL1264c (rhaT))*
**5: Pseudogenization of glucarate transporter (ps_SGL0291c (gudP))*
**6: Pseudogenization of Sucrose PTS subunit (ps_SGL1075c (scrA))*

With only arginine as external metabolite, all networks produces biomass, although at levels that are about half of the observed with glucose as external carbon source. This is consequence of the utilization of L-arginine for putrescine production and their concomitant degradation to TCA cycle intermediate succinate described above, from which all biomass constituents can be produced. In order to evaluate the specific effect of the external carbon source on biomass production rates separately from the effects of the external supply of L-arginine, the differences in biomass production with external supply of arginine plus carbon source and with only external arginine are represented. In concordance with the experimental observations on *S. glossinidius* cell cultures carried out by Dale and Maudlin (Dale and Maudlin, 1999), *S. glossinidius* functional network is functional in terms of biomass production with N-acetylglucosamine, glucose and mannitol as external carbon sources, whereas non-functional phenotypes are observed for galactitol, citrate, fumarate, glycolate, oxoglutarate, lactose, malate, maltose, melibiose, rhamnose, ribose, glucarate, succinate, sucrose, xylose, and arabinose. Of this non-assimilable carbon sources in *S. glossinidius* functional network, ancestral network appears functional under L-arabinose, fumarate, lactose, malate, rhamnose, glucarate, succinate and sucrose, indicating metabolic capabilities that has been lost as consequence of pseudogenization and gene deletion events during the transition to a host-dependent lifestyle (Figure 5.9). This is the case of the pseudogenization of the genes *araH*, *araG*, and *araF* encoding arabinose ABC transport system (psSGL0829, psSGL0830, and psAGL0831 respectively), the pseudogenization of the gene *lacZ* (ps_SGL1067) encoding a β-galactosidase for lactose degradation, the pseudogenization of the gene *rhaT* (ps_SGL1264c) encoding a rhamnose transporter, or the deletion of the gene *dcuA* that in *E. coli* JR904 network appears associated with the transport of succinate, malate and fumarate. Finally, discordant results between FBA simulations and experimental data are observed for acetate, ethanol, fructose, galactose, glycerol, mannose, pyruvate and trehalose, where biomass production is observed in FBA simulations in *S. glossinidius* functional network whereas no growth is reported in experimental observations (Dale and Maudlin, 1999). For simulations with acetate, pyruvate and ethanol, biomass production rates in FBA simulations are significant lower than under N-acetylglucosamine, glucose or mannitol, whereas for the cases of galactose and mannose metabolism, experimental observations contradicts the expected metabolic phenotype based on genome sequence, that is coincident with the predicted phenotype in terms of biomass production in FBA simulations (Figure 5.9). For example, *S. glossinidius* retains a completely functional pathway for galactose transport and metabolization through their transport by functional *mglB*, *mglA* and *mglC* genes (SG0963, SG0964 and SG0965 respectively) encoding a galactose ABC transport system and their concomitant degradation to the glycolytic intermediate glucose 6-phosphate by functional galactokinase *GalK* (SG0895), galactose 1-

phosphate uridyltransferase *GalT* (SG0896) and phosphoglucomutase *Pgm* (SG0866). However, both *galK* and *galT* genes correspond to a situation of a putatively longer ancestral gene where the absent part of the gene is detected as an adjacent pseudogene (described in the previous chapter), that in the case of *galT* represents 66.44% of the hypothetical ancestral gene, probably pointing out to an ongoing pseudogenization process.

The situation with mannose metabolism is different. As we have described in the previous chapter of this thesis, *S. glossinidius* retains a completely functional PTS system for mannose transport and phosphorylation (SG1325, SG1326 and SG1327) together with functional mannose 6-phosphate isomerase *ManA* (SG1461) for the conversion of mannose 6-phosphate to the glycolytic intermediate fructose 6-phosphate. As consequence, FBA simulations with mannose as external carbon source render a functional phenotype in functional metabolic network of *S. glossinidius*. By contrast, *S. glossinidius* have pseudogenized PTS systems for glucose and fructose, but mannose PTS system has been reported as a broad-specificity PTS capable to transport both glucose and fructose (Curtis and Epstein, 1975; Postma et al., 1993; Kornberg, 2001). This broad substrate-specificity of mannose PTS transport system has been incorporated to *S. glossinidius* functional network, where mannose PTS system (SG1325, SG1326 and SG1327) has been associated with transport and phosphorylation of glucose and fructose to their 6-phosphate derivatives. As consequence, functional network of *S. glossinidius* renders a functional phenotype in terms of biomass production with fructose, mannose and glucose in FBA simulations (see Figure 5.9). However, experimental observations of Dale and Maudlin report functional growth for *S. glossinidius* cell cultures under glucose but not under mannose and fructose as carbon sources (Dale and Maudlin, 1999), reflecting a possible case of novel metabolic speciation where mannose PTS transport system has been specialized in transport and phosphorylation of glucose instead of their natural substrate mannose.

A similar situation is found for trehalose metabolization. In *E. coli* K12, it has been characterized two different trehalases that catalyze the hydrolysis of trehalose to two molecules of glucose, a periplasmic trehalase encoded by the gene *treA* and a cytoplasmic isoform encoded by the gene *treF* (Elbein et al., 2003). *S. glossinidius* retains a single trehalase-encoding gene annotated as periplasmic trehalase *TreA* (SG1886) that shows higher identity with periplasmic trehalases including *E. coli* K12 *TreA* in BLASTP and FASTA searches. As consequence it appears associated with periplasmic trehalase reaction in *S. glossinidius* functional network (R_TREHe). Periplasmic trehalase in *S. glossinidius* degrades trehalose to their glucose monomers that can be transported to the cytoplasm through mannose PTS system as described above. In concordance with these expectations, *S. glossinidius* functional network renders a viable phenotype in terms of biomass production under trehalose as unique carbon source on FBA simulations (Figure 5.9), whereas

experimental observations of Dale and Maudlin reported non-growth for *S. glossinidius* under trehalose as carbon source (Dale and Maudlin, 1999).

5.3.3 Robustness analysis in *S. glossinidius* and *E. coli* K12 JR904 metabolic networks

The genetic robustness of *E. coli* JR904 and *S. glossinidius* metabolic networks was explored from two different perspectives. First, exploring network robustness to gene deletion events by removing single genes from the metabolic network, thereby setting zero fluxes through any reaction for which that particular gene product is essential. Second, exploring how biomass production rates changes as the metabolic flux across a particular reaction of interest varies in magnitude. The objective is to evaluate the effect of gene inactivation process in the robustness of metabolic networks to gene deletion events and changes in enzymatic activities of the central glycolytic pathway. The results of the gene deletion analysis over the three metabolic networks are represented in Figure 5.10.

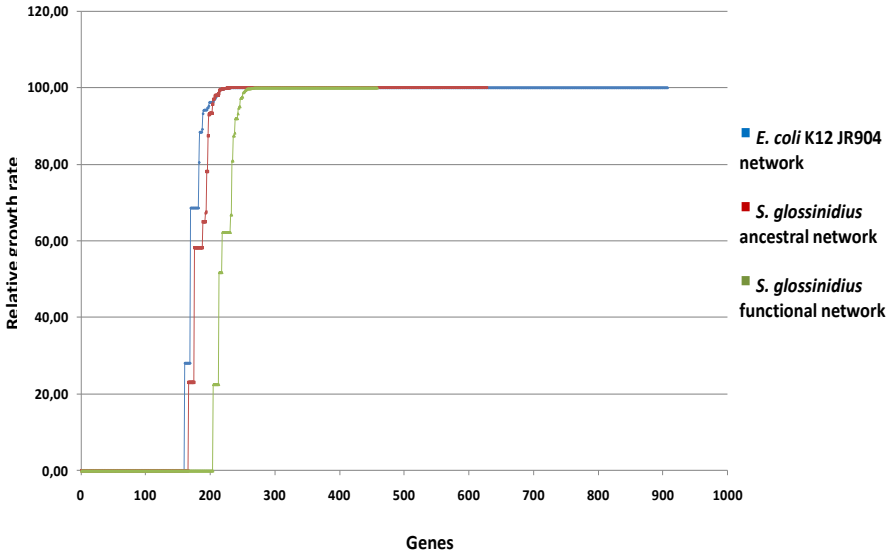


Figure 5.10: *Results of single knockout simulations for E. coli K12 JR904 and S. glossinidius metabolic networks using FBA. The relative growth rate for each knockout in comparison with the original network is represented*

The simulations were carried out in aerobic conditions with only arginine and glucose as external metabolites, due to the non-functionality of *S. glossinidius* functional network without external arginine supply. The metabolic network of *E. coli* JR904 provides a robust metabolic network for comparison (Reed et al., 2003). Under these conditions, the behavior of *S. glossinidius* ancestral network is similar to that of *E. coli* JR904 metabolic network in the fraction of essential genes or genes

whose deletion renders a lethal phenotype in terms of biomass production, with 160 and 166 genes in *E. coli* JR904 and *S. glossinidius* ancestral network respectively whose deletion renders a biomass production rate of 0 gr. Dry Weight (mmol Glucose)$^{-1}$ in FBA simulations. Both networks differs mainly in the number of genes whose deletion does not affect the functionality of the system in terms of biomass production, with 675 genes in *E. coli* JR904 network and 396 genes in *S. glossinidius* ancestral network whose deletion renders the same phenotype in terms of biomass production as their corresponding original network without the deletion. The differences in the number of essential and non-essential genes are more pronounced when the functional network of *S. glossinidius* is included in the comparison, where 204 genes appears as essential in single gene deletion experiments (0 gr. Dry Weight (mmol Glucose)$^{-1}$ in FBA simulations) whereas 190 genes can be deleted without affecting the biomass production rate of the network (0.7509 gr. Dry Weight (mmol Glucose)$^{-1}$ in FBA simulations, the same as original network). This suppose a strong decrease in the robustness of metabolic networks to gene deletion events as consequence of the gene inactivation process since the hypothetical free-living ancestor represented by *S. glossinidius* ancestral network to their actual state represented by their functional network, where the number of essential genes or genes whose deletion renders a lethal phenotype in terms of biomass production overcomes the number of non-essential genes or genes whose deletion does not affect the functionality of the metabolic system.

This is better represented when the fraction of essential and non-essential genes of the different metabolic networks are compared (Figure 5.11). If consider as essential gene a gene whose particular deletion results in a more than 1% decrease in the biomass production rate over original network, the fraction of essential genes increases from *E. coli* JR904 network (23.59 %) to *S. glossinidius* functional network (55.02 %), where its higher than the fraction of non-essential genes (44.98 %). If we compare this results with the results of the same simulations over the metabolic network of *Buchnera aphidicola* from the pea aphid *Acyrthosiphon pisum* carried out by Thomas and collaborators (Thomas et al., 2009), the fraction of essential genes increases to 84 % of the genes included in the network (Figure 5.11), revealing a common trend in reductive evolution related with a strong decrease of the robustness of metabolic system to gene deletion events associated with the transition to a host-dependent lifestyle, with *S. glossinidius* functional network representing initial stages of the association with the insect host whereas *B. aphidicola* representing advanced stages of the association.

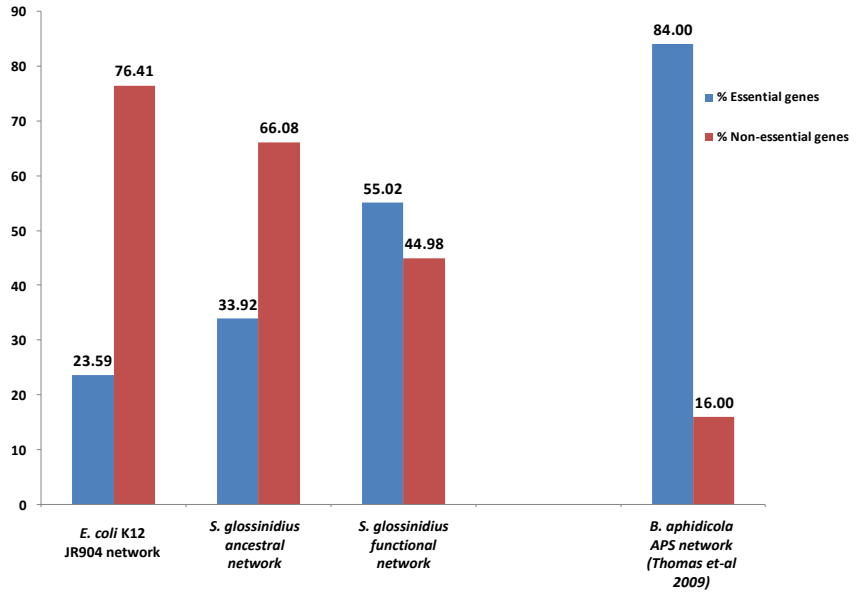


Figure 5.11: *Fraction of essential and non-essential genes in single knockut simulations over metabolic networks analyzed. Essential genes are defined as those genes whose deletion results in a decrease of more than 1% of the original biomass production rate. Values for B. aphidicola APS metabolic network are incorporated from (Thomas et al., 2009), where the same analysis is carried out*

Similar results were found in the robustness analysis over metabolic flux variations through particular reactions of the metabolic network (Figure 5.12). The reactions of the glycolysis were selected to carry out this analysis due to their essential role as central metabolic pathway that supplies precursors for the production of many biomass constituents, so variations in reactions of glycolytic pathway is expected to affect the biomass production rate of the whole system in FBA simulations. As we can observe in Figure 5.12, the range of metabolic flux values across glycolytic reactions over which is possible to produce a viable phenotype in terms of biomass production in FBA simulations is reduced since *E. coli* JR904 network to *S. glossinidius* functional network in all instances. This behavior can be interpreted as an increase in the sensitivity of metabolic networks to changes in metabolic fluxes as consequence of the gene inactivation process, with the functional metabolic network of *S. glossinidius* that have a more restricted range of metabolic fluxes over which the system can produce a functional phenotype in terms of biomass production compared with their ancestral network and *E. coli* JR904 network.

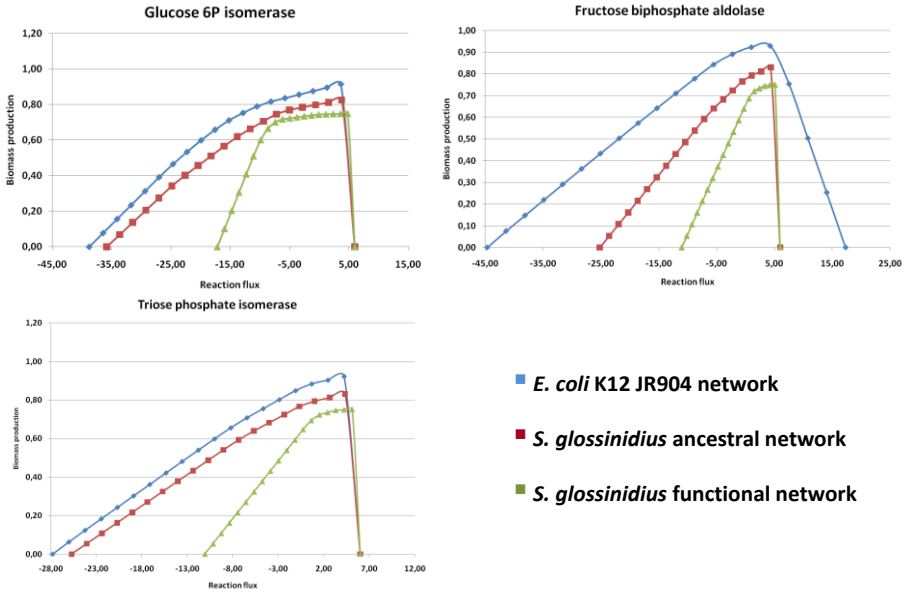


Figure 5.12: *Results of robustness analysis to metabolic flux variations in three reactions of the glycolysis for E. coli K12 JR904 and S. glossinidius metabolic networks. Similar results are obtained for all glycolytic reactions*

5.3.4 Reductive evolution simulations over functional network of *S. glossinidius*

In order to predict the possible future evolution of *S. glossinidius* in the context of the reductive evolution process, reductive evolution simulations are carried out over *S. glossinidius* functional network (458 genes, 560 internal reactions) that consist in the sequential random removal of each reaction of the original metabolic network evaluating at each step the functionality of the deleted network in terms of biomass production; if this rate is nearly unaffected, the reaction and their corresponding genes are permanently removed from the network, otherwise, the reaction and their corresponding genes are restored to the network and the next reaction is evaluated, proceeding repeatedly until all reactions have been evaluated. Three different cutoffs for defining a reaction as essential or non-essential were evaluated and for each cutoff 500 simulations were carried out randomizing in each one the order of reactions to be tested. Initially, simulations were carried out under minimal aerobic conditions with only glucose and arginine as external metabolites assuming a maximum uptake rate of 6 mmol gr. DryWeigth^{-1} hr^{-1} for both metabolites, whereas unconstrained uptake for ammonia, water, phosphate, sulfate, potassium, sodium, iron (II), carbon dioxide and protons was allowed. These conditions are the minimal needed to obtain a viable phenotype for *S. glossinidius*

functional network, as we have observed in the previous section. In addition, similar simulations were also carried out under nutrient-rich conditions that consist in allow the external influx of 41 metabolites adapted from a similar study of Pal and collaborators, where similar reductive evolution simulations over *E. coli* JR904 network with a set of external metabolites available for uptake were used to model *W. glossinidia* reductive evolution process (Pal et al., 2006b). In order to model *S. glossinidius* reductive evolution in the context of its ecological association with the tsetse host, lower bounds of exchange reactions for 41 metabolites of the list proposed by Pal and collaborators for which functional transport system is present in the functional network of *S. glossinidius* were modified to allow a maximum uptake rate of 6 mmol gr. DryWeigth^{-1} hr^{-1} for each one. The results of these simulations over functional network of *S. glossinidius* are represented in Table 5.4.

Reductive evolution simulations under nutrient-limited conditions	10% original biomass	5% original biomass	1% original biomass	Total 1500 minimal networks
Mean gene number (+- SE)	291.47 (+- 0.17)	292.15 (+- 0.16)	291.64 (+- 0.14)	291.75 (+- 0.09)
Mean reaction number (+- SE)	280.03 (+- 0.12)	280.51 (+- 0.13)	279.78 (+- 0.12)	280.11 (+- 0.07)
Common genes 500 min networks	258	257	261	255
Absent genes 500 min networks	129	128	127	127
Common reactions 500 min networks	240	239	242	237
Absent reactions 500 min networks	208	205	203	200

Reductive evolution simulations under nutrient-rich conditions	10% original biomass	5% original biomass	1% original biomass	Total 1500 minimal networks
Mean gene number (+- SE)	291.47 (+- 0.17)	292.15 (+- 0.16)	291.64 (+- 0.14)	237.25 (+- 0.26)
Mean reaction number (+- SE)	280.03 (+- 0.12)	280.51 (+- 0.13)	279.78 (+- 0.12)	231.9 (+- 0.18)
Common genes 500 min networks	258	257	261	139
Absent genes 500 min networks	129	128	127	111
Common reactions 500 min networks	240	239	242	141
Absent reactions 500 min networks	208	205	203	185

Table5.4: *Results of reductive evolution simulations over nutrient-limited and nutrient-rich conditions*

Under nutrient-limited conditions, the three cutoffs evaluated renders very similar results, with an average network size over 1500 minimal networks of 280.11 internal reactions and 291.75 genes. In addition, 237 reactions and 255 genes were commonly present in all minimal networks whereas 200 reactions and 127 genes were commonly absent from all minimal networks. This supposes that 84.61% of the reactions and 87.4% of the genes are shared over all minimal networks. Under

nutrient-rich conditions, the three cutoffs tested produces also similar results, with an average network size of 231.9 internal reactions and 237.25 genes in the 1500 minimal networks. In addition, 141 internal reactions and 139 genes were always present in all minimal networks whereas 185 reactions and 111 genes were always absent in all minimal networks under nutrient-rich conditions. This supposes that 60.86% of the reactions and 58.69% of the genes were common to all minimal networks under nutrient-rich conditions, reflecting a higher degree of plasticity compared with the minimal networks obtained with nutrient-limited conditions with only glucose and arginine as external metabolites, where the range of potential outcomes in terms of gene and reaction content of minimal networks are further restricted (See Figure 5.13).

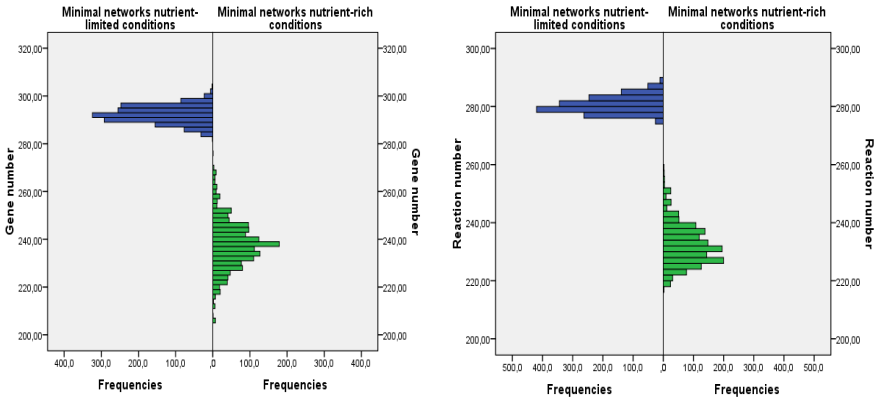


Figure 5.13: *Distribution of the number of genes and reactions in 1500 minimal networks under nutrient-limited and nutrient-rich conditions*

Minimal networks can be also compared in terms of the number of genes and reactions commonly present and absent in all minimal networks under original and nutrient-rich conditions (Table 5.5). When this analysis is carried out with essential genes (genes present in all minimal networks), we see that 138 out of 139 essential genes under nutrient-rich conditions were also essential under original conditions, whereas all 141 essential reactions in nutrient-rich conditions were also essential under original conditions. This conforms to what one would expect in the sense that a reaction that is essential in all reductive evolution simulations under a nutrient-rich environment would be also essential in a more limited environment with only glucose and arginine as external metabolites.

	Minimal networks nutrient-limited conditions	Minimal networks nutrient-rich conditions	Genes	Reactions
Essential genes in all minimal networks	Present	Present	138	141
	Present	Absent	117	96
	Absent	Present	1	0
Disposable genes in all minimal networks	Present	Present	109	172
	Present	Absent	18	28
	Absent	Present	2	13

Table 5.5: *Comparison of essential and disposable genes in minimal networks under nutrient-limited and nutrient-rich conditions. The set of essential genes in each condition includes genes present in 1500 minimal networks, whereas the set of disposable genes includes genes of S. glossinidius functional network absent in 1500 minimal networks in each condition.*

The unique gene always present in all minimal networks under nutrient-rich conditions but not under minimal conditions was SG0465, that encodes an aromatic amino acid permease *AroP* responsible of the transport of histidine, phenylalanine, tyrosine and tryptophan in both *E. coli* JR904 and *S. glossinidius* metabolic networks. These four amino acids are included in the biomass equation and are available for uptake in reductive evolution simulations under nutrient-rich conditions. As consequence, the removal of any reaction involved in the biosynthesis of any of these amino acids becomes *aroP* as essential gene in order to obtain a component of the biomass equation needed to produce a viable phenotype in terms of biomass production. The whole set of 138 genes and 141 reactions essential under all conditions are represented in Supplementary Table 5.5. These genes and reactions are mainly involved in the biosynthesis of biomass constituents that cannot be assimilated from the environment in nutrient-rich conditions. These includes the complete biosynthetic pathways for putrescine and spermidine biosynthesis from external L-arginine together with Arginine ABC transport system, the complete pathways for membrane phospholipids (phosphatidylglycerol, phosphatidylethanolamine, and cardiolipin), peptidoglycan and bacterial lipopolysaccharide biosynthesis, the complete pathways for the biosynthesis of methylenetetrahydrofolate, coenzyme A, FAD and NAD biosynthesis, the complete pathway for L-proline biosynthesis from L-glutamate, or *glnA* gene for glutamine biosynthesis from L-glutamate. Other essential genes in all minimal networks under all conditions are involved in the biosynthesis of intermediate metabolites essential for the biosynthesis of biomass constituents, like the gene *dadX* encoding alanine racemase responsible of the isomerization of L-alanine to D-alanine, essential for peptidoglycan biosynthesis, the gene *prsA* encoding phosphoribosyl pyrophosphate synthetase responsible of 5-phosphoribosyl 1-pyrophosphate (PRPP) biosynthesis,

common precursor of histidine, tryptophan, purine and pyrimidine nucleotides and NAD biosynthesis and salvage pathways, the gene *metK* encoding a methionine adenosyltransferase responsible of S-adenosylmethionine biosynthesis from L-methionine, essential precursor for spermidine biosynthesis, or all genes involved in meso-diaminopimelate biosynthesis from L-aspartate, essential for peptidoglycan and L-lysine biosynthesis. Finally, reactions like urea and glycolate transport appears as essential in all minimal networks under original and nutrient-rich conditions because are needed to eject these metabolites that are always produced by the system associated to putrescine and methylenetetrahydrofolate biosynthesis respectively in order to obtain a functional phenotype in FBA simulations under the steady-state assumption that governs the behavior of the system (Schilling et al., 2000a).

There are also 117 genes and 96 reactions that appear as essential in reductive evolution simulations under original conditions with only arginine and glucose as external metabolites but not under nutrient-rich conditions. These genes can be interpreted as susceptible to be lost by *S. glossinidius* in the context of their association with the tsetse host. These genes and reactions are represented in Supplementary Table 5.6. Most of these genes and reactions correspond to amino acid biosynthesis genes, most of them also absent in *W. glossinidia* (Akman et al., 2002; Zientz et al., 2004). Amino acids are components of biomass equation, and under original conditions with only arginine and glucose as external metabolites these genes are always present because they represents the unique way to produce these biomass constituents, whereas under nutrient-rich conditions their presence in minimal networks will depend on which reactions (transport or biosynthesis) are eliminated first; if transport reaction for a particular amino acid is eliminated first, all genes involved in their biosynthetic pathway becomes essential whereas if a single biosynthetic reaction is eliminated first, the system can still produce a viable phenotype in terms of biomass production if the amino acid are available for uptake (lower bound of their corresponding exchange reaction lower than zero) and there is a functional transport reaction for their incorporation to the cytoplasmic compartment, that will become essential. As consequence, all genes involved in asparagine, aspartate, alanine, cysteine, serine and glycine, histidine, methionine, threonine, lysine, tyrosine, tryptophan, phenylalanine, valine, leucine, and isoleucine appear as essential only under original conditions but not under nutrient-rich conditions. For all these amino acids *S. glossinidius* retains a functional transport system so they can be synthesized *de-novo* or be acquired from external sources in order to contribute to biomass production. For L-lysine biosynthesis, only the gene *lysA* (SG1988) encoding a diaminopimelate decarboxylase responsible of the conversion of meso-diaminopimelate to L-lysine appears included in this list, whereas the rest of genes for meso-diaminopimelate biosynthesis from L-aspartate appears essential under all conditions due to the essential role of meso-diaminopimelate as peptidoglycan precursor. This has been also observed in

experimental essentiality analysis over *E. coli* K12 genes under nutrient-rich environment, where all genes involved in meso-diaminopimelate biosynthesis appears as essential based on experimental knockouts by transposon mutagenesis, whereas *lysA* gene inactivation appears as viable under nutrient-rich environment (Gerdes et al., 2003). In a similar way, glycerol 3-phosphate dehydrogenase encoded by the gene *gpsA* (SG2184) appears also as essential under original but not under nutrient-rich conditions. This enzyme catalyzes the reduction of the glycolytic intermediate dihydroxyacetone phosphate to glycerol 3-phosphate, essential for membrane phospholipids biosynthesis, but under nutrient-rich conditions, glycerol 3-phosphate can be produced through the entry of external glycerol through functional glycerol facilitator protein *GlpF* (SG1858) followed by their phosphorylation by functional glycerol kinase *GlpK* (SG2172). A particular situation is found with the pathway for aromatic amino acids biosynthesis from pentose phosphate pathway intermediate erythrose 4-phosphate through chorismate intermediate, where all genes of the shikimate pathway for chorismate biosynthesis from erythrose 4-phosphate appears present in all minimal networks under original and nutrient-rich conditions, whereas the remaining steps from chorismate to tryptophan, phenylalanine and tyrosine appears only essential under original but not under nutrient-rich conditions. This is because aromatic amino acids can be incorporated from external sources through functional transporter *AroP* (SG0465) whereas chorismate is essential for methylenetetrahydrofolate biosynthesis, and cannot be incorporated from external sources (See Figure 5.14).

Similar situation is found with the pathways for purine and pyrimidine nucleotides biosynthesis, where final steps of the pathways from the nucleotide monophosphates (UMP, AMP, GMP) to the nucleotide and deoxyribonucleotide triphosphate derivatives appear always present in all minimal networks under original and nutrient-rich conditions, whereas initial steps of the pathways for de-novo purine and pyrimidine nucleotides biosynthesis until the production of nucleotide monophosphate derivatives appear always present only under original conditions but not under nutrient-rich conditions. This can be explained by the presence of functional guanine phosphoribosyltransferases *GpT* and *Hpt* (SG0598 and SG0482 respectively) that allows the synthesis of GMP from exogenous guanine, and by the presence of functional uridine kinase *Udk* (SG0973) and adenosine kinase *Adk* (SG0693) that allows exogenous uridine and adenosine nucleosides to be converted to UMP and AMP respectively. Additional genes and reactions only essential under original conditions includes TCA cycle enzymes (malate:quinone oxidoreductase, citrate synthase, aconitase, isocitrate and succinate dehydrogenase and fumarase), NADH dehydrogenase, or ATP synthase enzyme complexes for oxidative phosphorylation.

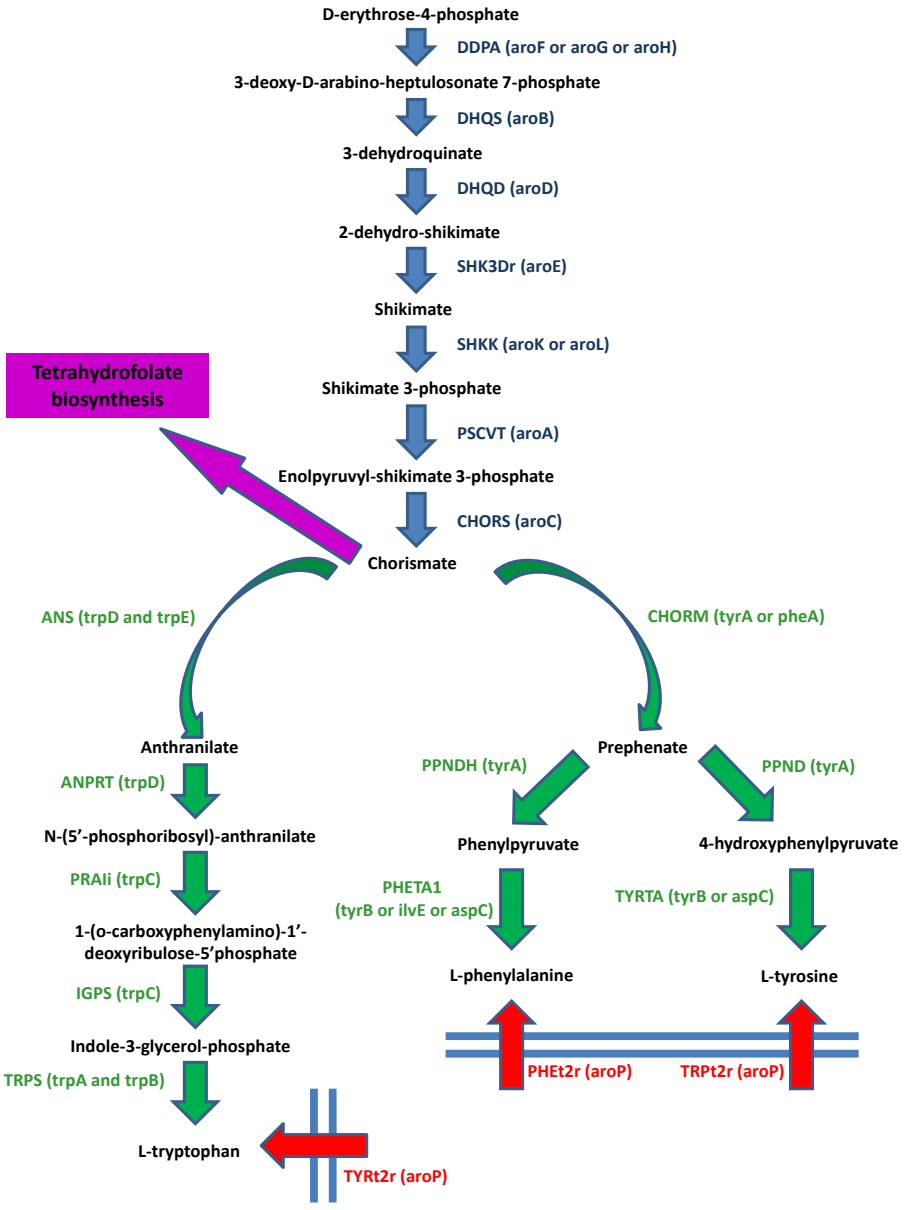


Figure 5.14: *Aromatic amino acids biosynthetic pathway from erythrose 4-phosphate. Blue arrows represent essential genes in all minimal networks in all conditions, whereas green arrows represent conditionally essential genes only under nutrient-limited conditions. Under nutrient-rich conditions, aromatic amino acids can be incorporated through transport by aromatic amino acid permease AroP. Chorismate is the precursor of the biomass constituent tetrahydrofolate, so it has to be produced in all FBA simulations.*

For reactions and genes non-essential or absent in all minimal networks, there is 109 genes and 172 reactions that are commonly absent in all minimal networks under original and nutrient-rich conditions (see supplementary Table 5.7). These includes the remaining steps of metabolic pathways from which some of their components appears pseudogenized or absent in the functional network of *S. glossinidius*, like the remaining steps of arginine biosynthesis pathway or thiamine biosynthesis pathway, both pseudogenized in *S. glossinidius*, the fructose 1-phosphate kinase *FruK* (SG0954) that phosphorylate fructose 1-phosphate to glycolytic intermediate fructose 1,6-biphosphate but whose substrate (fructose 1-phosphate) cannot be produced by *S. glossinidius* functional network due to the pseudogenization of fructose PTS system, or the genes *uxuB*, *uxuA* and *kdgK* (SG1837, SG1838 and SG0067 respectively) responsible of the conversion of fructuronate to 2-dehydro-3-deoxy-D-glucuronate 6-phosphate, the key intermediate in the Entner-Doudoroff pathway for hexuronides degradation (glucuronate and galacturonate) to glycolytic intermediates pyruvate and glyceraldehyde 3-phosphate (Peekhaus and Conway, 1998). In *S. glossinidius*, pseudogenization have affected to *eda* gene responsible of the cleavage of 2-dehydro-3-deoxy-D-glucuronate 6-phosphate to pyruvate and glyceraldehyde 3-phosphate (ps_SGL0711), whereas there is no signal for any transporter for glucuronate or galacturonate, so reactions catalyzed by *UxuB*, *UxuA* and *KdgK* will never be functional in *S. glossinidius* functional network under any external conditions. However, non-essential genes and reactions absent in all minimal networks include also complete biosynthetic pathways for metabolites that are not included in the biomass equation reaction and, as consequence, the network will never produce it in FBA simulations. These includes the complete pathways for Enterobacterial Common Antigen (ECA) extracellular polysaccharide, the complete pathway for GDP-L-fucose biosynthesis, one of the building blocks of Colanic acid outer membrane glycolipid, or the complete pathways for the biosynthesis of cofactors ubiquinone, pyridoxine 5-phosphate, biotin, and heme groups (See Supplementary table 5.8). Cofactor biosynthesis genes has been retained by tsetse primary endosymbiont *W. glossinidia* in order to provide the tsetse host with cofactors that cannot be assimilated from the vertebrate blood, and in fact all this cofactor biosynthesis pathways that appears absent in all minimal networks are present in *W. glossinidia* genome sequence (Akman et al., 2002; Zientz et al., 2004).

Orthologous genes between *S. glossinidius* and *W. glossinidia* were also identified in order to compare the sets of essential and non-essential genes defined *in-silico* in *S. glossinidius* minimal networks with the gene content of natural minimal genome evolved under similar external conditions in the context of the symbiotic association with the tsetse host. The results of this analysis are represented in Figure 5.15. Under original conditions with only arginine and glucose as external metabolites, 143 of the 255 essential genes present in all minimal networks were present in *W. glossinidia* , what represents a sensitivity in the

predictions of 56.08%, whereas 122 essential genes were absent in *W. glossinidia* (43.92% of essential genes). When similar analysis is carried out with the set of essential genes characterized in reductive evolution simulations with nutrient-rich conditions (139 genes), the sensitivity of the predictions increases to 69.06% (96 of 139 essential genes present in *W. glossinidia* genome). This is consequence of the absence of most amino acid biosynthesis genes in the set of essential genes under nutrient-rich conditions, that are absent also in *W. glossinidia* (Akman et al., 2002). However, there remain 43 essential genes of *S. glossinidius* minimal networks in nutrient-rich conditions still absent in *W. glossinidia* (30.94% of essential genes in nutrient-rich conditions).

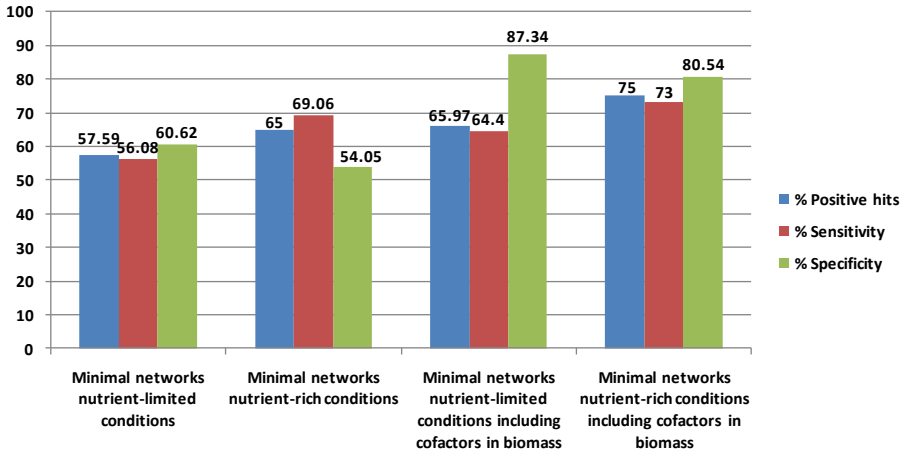


Figure 5.15: *Bar diagram of sensitivity and specificity of reductive evolution simulations with W. glossinidia gene content for essential and disposable genes in minimal networks. Sensitivity is defined as the fraction of essential genes in all minimal networks present in the genome of W. glossinidia (true positives), whereas specificity is defined as the fraction of dispensable genes absent in the genome of W. glossinidia (true negatives). The fraction of positive hits represents the sum of true positives and true negatives averaged over the total number of essential and disposable genes in all minimal networks*

These 43 genes correspond with essential enzymatic activities for the production of different biomass constituents in the context of FBA simulations even under nutrient-rich condition, like five of the seven enzymes involved in shikimate pathway for chorismate biosynthesis from the pentose phosphate pathway intermediate erythrose 4-phosphate (*aroD*, *aroA*, *aroC*, *aroE*, and *aroB* genes), essential in the context of tetrahydrofolate biosynthesis under nutrient-rich conditions together with *pabA*, *pabB* and *pabC* genes for the biosynthesis of the tetrahydrofolate essential precursor p-aminobenzoate from chorismate, that appears also absent in *W. glossinidia* despite the presence of the rest of genes for tetrahydrofolate biosynthesis, as was described in the previous chapter of this thesis.

There are also included all the genes needed for putrescine and spermidine biosynthesis from L-arginine, essential biomass constituents for the viability of *S. glossinidius* functional network that appears also absent in *W. glossinidia*, or the genes responsible of the biosynthesis of a complete bacterial lipopolysaccharide also included as biomass constituent but also absent in *W. glossinidia* (Zientz et al., 2004). The same analysis is carried out with the set of non-essential genes absent in minimal networks under original and nutrient-rich conditions (see Figure 5.15). In original conditions with only arginine and glucose as external metabolites, 127 genes appears absent in all minimal networks, of which 77 are also absent in *W. glossinidia*, what represents a specificity of the predictions of 60.62%, whereas in minimal networks under nutrient-rich conditions the specificity decreases to 54.05% (60 of the 111 non-essential genes absent in *W. glossinidia* genome). However, within the set of non-essential genes in both nutrient-limited and nutrient-rich conditions are included 48 genes involved in the biosynthesis of ubiquinone, pyridoxine 5-phosphate, biotin and heme groups, of which 40 are present in the genome of *W. glossinidia*. These genes appear as non-essential because the corresponding cofactors are not included in biomass equation, and as consequence the system will never produce it in steady state in FBA simulations. However, the inclusion of this cofactors in biomass equation will convert automatically this genes in essential genes under all conditions, because corresponds to unbranched linear pathways where all their reactions have to operate together in order to produce the final cofactor, like the 11 needed to synthesize heme-O group from L-glutamate. If we carry out the same analysis considering this cofactor biosynthesis genes as essential genes, the specificity of the predictions increases to more than 80%, increasing also the sensitivity at lower level (see Figure 5.15), what means that a significant proportion of non-essential genes lost in reductive evolution simulations over functional network of *S. glossinidius* are also absent in the highly streamlined genome of *W. glossinidia*.

5.3.5 Evolutionary analysis over essential and non-essential genes in reductive evolution simulations

In order to evaluate if the essential character of metabolic genes defined based on their importance for the functionality of the system in terms of biomass production in FBA simulations corresponds with a differential pattern of sequence evolution and gene expression, the codon adaptation index (CAI) for *S. glossinidius* genes and the number of non-synonymous (*dN*) and synonymous (*dS*) substitutions per site between *S. glossinidius* and *E. coli* K12 orthologs were estimated. The starting hypothesis was that essential genes for the functionality of metabolic networks are expected to show more restricted patterns of sequence evolution and a more optimized codon usage rather than non-essential genes whose removal does not affect the functionality of the system. The comparison between the average CAI for

essential and non-essential genes under original conditions with only arginine and glucose as external metabolites is represented in Figure 5.16.

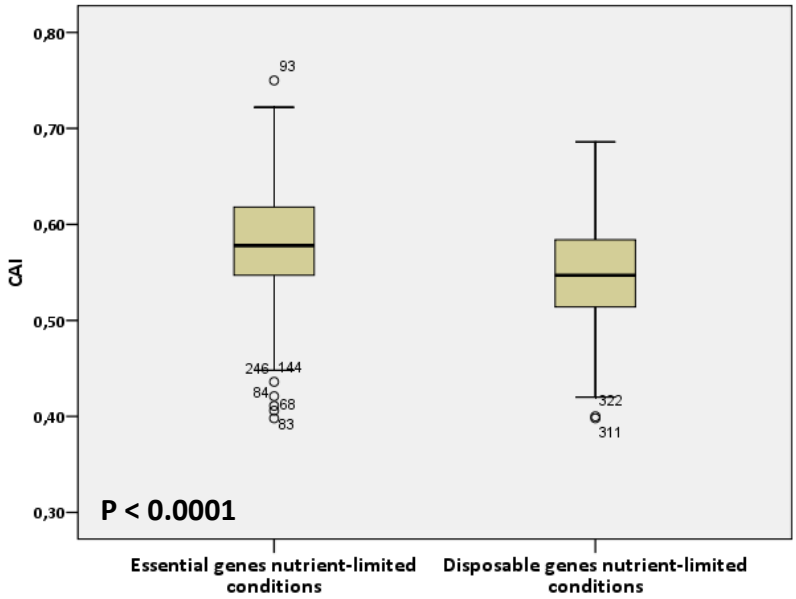


Figure 5.16: *CAI values for S. glossinidius essential and disposable genes in all minimal networks under nutrient-limited conditions. Global significance between means was tested by Student´s t-test based on the normal distribution and homocedasticity of CAI values in both gene groups.*

The 255 essential genes present in all minimal networks show a mean CAI of 0.581, whereas the 127 disposable genes present a mean CAI of 0.546. In order to test if these differences were statistically significant, a Student´s t-test to compare mean CAI values between the two groups was applied based on the normal distribution of CAI values in both groups ($p = 0.200$ in Kolgomorov-Smirnov test) and the equality of their variances ($p = 0.883$ in Levene´s test). These differences were statistically significant at 0.05 significance level ($p < 0.0001$ on Student´s t-test), which means that essential genes for the functionality of all minimal networks in terms of biomass production in all minimal networks have a more optimized codon usage than none-essential genes absent in all minimal networks whose removal does not affect the functionality of the system. In contrast, mean values for *dN* estimates between essential and disposable genes for *S. glossinidius* and *E. coli* K12 ortholog pairs show the inverse trend (see Figure 5.17A). We found that the 246 essential genes with orthology with *E. coli* K12 showed a mean *dN* of 0.176, whereas the 127 disposable genes, all of which have their corresponding ortholog in *E. coli* K12, had a mean *dN* of 0.254. In order to compare if this differences were statistically significant, a non-parametric Mann-Whitney U-test was applied due to

the absence of normal distribution in mean *dN* values in both groups ($p < 0.0001$ in Kolgomorov-Smirnov test) and the inequality of their variances ($p < 0.0001$ in Levene´s test). These differences were statistically significant at 0.05 significance level ($p < 0.0001$ on Mann-Whitney U-test), indicating that essential genes have a significantly more restricted patterns of sequence evolution at non-synonymous sites than non-essential genes. Similar results were obtained when *dS* values were compared (Figure 5.17B), with the 246 essential genes with functional ortholog in *E. coli* K12 genome having a mean *dS* of 1.916 whereas the 127 disposable genes having a mean *dS* of 2.205. A Mann-Whitney U-test was applied to compare mean *dS* values between both groups due to the absence of normal distribution ($p < 0.0001$ in Kolgomorov-Smirnov test), despite the homogeneity of their variances ($p = 0.208$ in Levene´s test), and the differences were statistically significant at 0.05 significance level ($p < 0.0001$ on Mann-Whitney U-test).

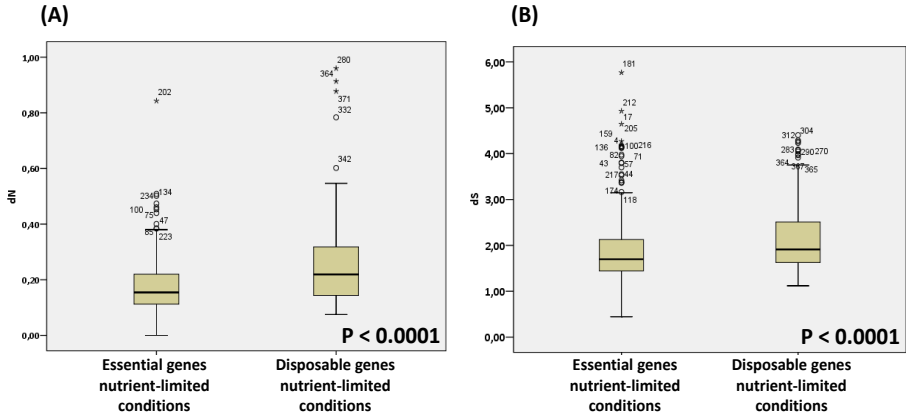


Figure 5.17: *Pairwise dN (A) and dS (B) values between S. glossinidius and E. coli K12 orthologs for essential and disposable genes in all minimal networks under nutrient-limited conditions. Global significance between medians was tested by Mann-Whitney tests due to the lack of normal distribution and homocedasticity of both variables.*

Similar results were obtained when the set of essential and disposable genes defined under nutrient-rich conditions were analyzed, with essential genes present in all minimal networks that have significantly higher CAI values than disposable genes absent in all minimal networks (mean CAI = 0.573 for essential genes, mean CAI =0.548 for disposable genes; $p < 0.0001$ in Student´s t-test), whereas disposable genes show significantly higher values of *dN* (mean *dN* = 0.197 for essential genes, mean *dN* = 0.260 for disposable genes; $p < 0.0001$ in Mann-Whitney U-tests) and *dS* (mean *dS* = 2.029 for essential genes, mean dS = 2.184 for disposable genes; $p = 0.026$ in Mann-Whitney U-test) than the set of essential genes (see Table 5.6).

		CAI (+- SE) (*2)	dN (+- SE) (*2)	dS (+- SE) (*2)
Minimal networks nutrient-limited conditions	Essential genes 1500 minimal networks	0.581 (+- 0.00353)	0.176 (+- 0.006)	1.916 (+- 0.05)
	Disposable genes 1500 minimal networks	0.546 (+- 0.005)	0.254 (+- 0.014)	2.205 (+- 0.07)
	Significance of the difference (*1)	P < 0.0001	P < 0.0001	P < 0.0001
Minimal networks nutrient-rich conditions	Essential genes 1500 minimal networks	0.573 (+-0.004)	0.197 (+- 0.009)	2.029 (+- 0.07)
	Disposable genes 1500 minimal networks	0.548 (+- 0.005)	0.260 (+- 0.02)	2.184 (+- 0.07)
	Significance of the difference (*1)	P < 0.0001	P < 0.0001	P = 0.026

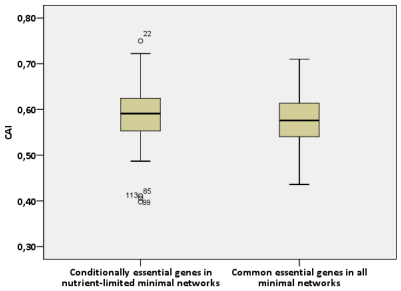
Table 5.6: *CAI, dN and dS values for essential and disposable genes in minimal networks under nutrient-limited and nutrient-rich conditions*
**1 Statistical significance of the differences between essential and disposable genes for CAI values determined by parametric Student´s t-test based on the normal distribution of the variable and the homocedasticity in their variances. Statistical significance of the differences between essential and disposables genes for dN and dS values determined using non-parametric Mann-Whitney U test based on the absence of normal distribution and homocedasticity.*
**2 The mean CAI, dN and dS for all genes within essential and disposable group are shown with the corresponding standard error (in parentheses).*

These results reveal a significant correlation between the essential character of metabolic genes defined in reductive evolution simulations and their corresponding patterns of sequence evolution, that is conserved despite changes in the external conditions; genes whose activity are essential to obtain a functional phenotype in terms of biomass production have a more optimized codon usage than genes whose removal does not affect the functionality of the metabolic system, and these non-essential genes, as a consequence, have significantly higher rates of sequence evolution at both synonymous and non-synonymous sites.

However, although the general trend is common under both external conditions (original and nutrient-rich), the difference in mean values of parameter estimates between essential and non-essential genes is lower under nutrient-rich conditions than under original conditions. If we compare the ratio of mean *dN* values between essential and non essential genes, this has a value of 0.693 in original condition simulations, whereas for nutrient-rich simulations this value increases to 0.758. Similarly, the ratio of mean *dS* values between essential and non-essential genes under original conditions is 0.869, whereas under nutrient-rich conditions is 0.929.

This indicates that despite essential and non-essential gene sets show significantly different patterns of sequence evolution in both original and nutrient-rich conditions, the differences between both gene sets are lower in minimal networks defined under nutrient-rich environment. As described previously, minimal networks under original and nutrient-rich conditions mainly differ in the different number of essential genes present in all minimal networks, with 117 genes that appear as essential under original conditions with only arginine and glucose as external metabolites but that are not essential in simulations under nutrient-rich environment (conditionally essential genes), whereas 138 of the 139 essential genes under nutrient-rich conditions were also essential under original conditions with only arginine and glucose as external metabolites. If we compare the patterns of sequence evolution between both gene sets (see Figure 5.18), we see that there is no statistically significant differences in their codon usage at 0.05 level (mean CAI = 0.589 for 117 conditionally essential genes, mean CAI =0.577 for 137 essential genes in all minimal networks; p = 0.067 in Student´s t-test). However, significant differences appear when mean values of *dN* and *dS* estimates were compared. Conditionally essential genes (115 of 117 genes with ortholog with *E. coli* K12) show a mean *dN* of 0.152 whereas common essential genes in all minimal networks (131 of 137 with ortholog with *E. coli* K12) show a mean *dN* of 0.197, being these differences statistically significant (p = 0.001 in Mann-Whitney U-test). Similar results are obtained when mean *dS* values were compared (mean *dS* = 1.793 for essential genes only in original conditions, mean *dS* = 2.023 for essential genes common to all minimal networks; p = 0.004 in Mann-Whitney U-test).

These results indicate that, despite having no significant differences in codon usage, genes that are only essential under original conditions (only glucose and arginine as external metabolites) show a more restricted pattern of sequence evolution at both synonymous and non-synonymous sites than the core of essential genes (present in all minimal networks in both conditions). These conditionally essential genes correspond to genes that *S. glossinidius* could lose in a nutrient-rich environment like that found within the tsetse flies in the context of the reductive evolution process. In fact, as described above, many of these conditionally essential genes are absent in the genome of *W. glossinidia* (like amino acid biosynthesis genes), the primary bacterial endosymbiont of tsetse flies that is in an advanced stage of the reductive evolution process due to their more ancestral symbiotic association (Chen et al., 1999; Akman et al., 2002; Zientz et al., 2004).



	CAI (+- SE) (*2)	dN (+- SE) (*2)	dS (+- SE) (*2)
Conditionally essential genes in nutrient-limited minimal networks	0,589 (+- 0,005)	0,152 (+- 0,007)	1,793 (+- 0,06)
Common essential genes in all minimal networks	0,577 (+- 0,005)	0,197 (+- 0,009)	2,023 (+- 0,07)
Significance of the difference (*1)	P = 0,067	P = 0,001	P = 0,004

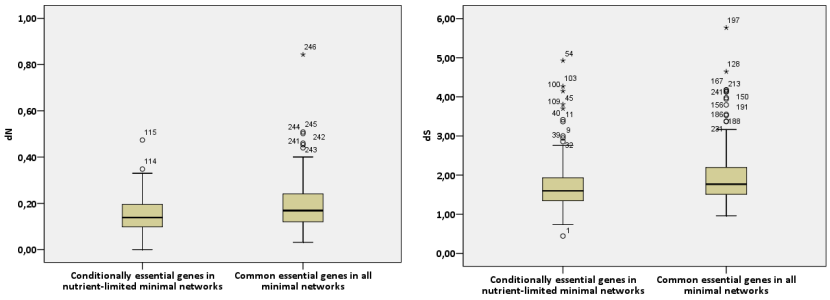


Figure 5.18: *CAI for S. glossinidius genes and pairwise dN and dS values between S. glossinidius and E. coli K12 orthologs in conditionally essential genes only under nutrient-limited conditions and common essential genes in all minimal networks in all conditions*

*(*1) Statistical significance of the differences between groups for CAI values determined by parametric Student´s t-test based on the normal distribution of the variable and the homocedasticity in their variances. Statistical significance of the differences between essential and disposables genes for dN and dS values determined using non-parametric Mann-Whitney U test based on the absence of normal distribution and homocedasticity*

*(*2) The mean CAI, dN and dS for all genes within each group is shown with the corresponding standard error (in parentheses).*

If *S. glossinidius* followed the same reductive evolution pattern than *W. glossinidia*, we would expect a decrease of the selective pressures over this set of conditionally essential genes, that can be lost in a nutrient-rich environment without affecting the functionality of the metabolic system, and, as a consequence, higher rates of sequence evolution at both synonymous and non-synonymous sites. However, the inverse situation is observed, with conditionally essential genes showing more restricted patterns of sequence evolution than core metabolic genes present in all minimal networks in all conditions. One possible explanation is that conditionally essential genes were needed for the survival of *S. glossinidius* within the tsetse host, and, as a consequence, their patterns of sequence evolution were restricted even further than core metabolic genes common to all minimal networks. The very recent transition to a host dependent lifestyle experienced by *S. glossinidius* revealed by their genome features (Toh et al., 2006), their wider tissue

tropism both intra and extracellularly within the tsetse host (Cheng and Aksoy, 1999) and their unique capability within bacterial endosymbionts to be cultured in-vitro (Welburn et al., 1987) can make that selective pressures that govern reductive evolution of *W. glossinidia* primary endosymbiont were not acting with the same strength over *S. glossinidius*. This is also in concordance with the presence of 48 cofactor biosynthesis genes within the set of disposable genes absent in all minimal networks under all conditions. These genes have been specifically retained by *W. glossinidia* despite its reduced genome due to their role in cofactor supply to the tsetse host (Akman et al., 2002). The presence of cofactor biosynthesis genes in the set of non-essential genes under all conditions can be also explained by the absence of these cofactors (ubiquinone, biotin and pyridoxine 5-phosphate) in the biomass equation; however, when the mean CAI, dN and dS values of these cofactor biosynthesis genes were compared with the rest of non-essential genes absent in all minimal networks under all conditions (61 genes) and with the set of 117 conditionally essential genes, these genes appear closer to the rest of disposable genes rather than to the conditionally essential genes in both codon usage and sequence evolution patterns. ANOVA tests over these three groups of genes for CAI, *dN* and *dS* show statistically significant differences between conditionally essential genes and the two other groups in the three variables, whereas differences between cofactor biosynthesis genes and the rest of disposable genes were not significant (see Table 5.7).

However, to validate this hypothesis, the analysis of the sequence evolution over the branch corresponding to the evolution of *S. glossinidius* since its divergence from the common ancestor with *E. coli* K12, is required. It should be noted that pairwise estimates of *dN* and *dS* calculates the number of synonymous and non-synonymous substitutions per site that took place in the whole evolutionary lineage between *S. glossinidius* and *E. coli* K12, so differences observed in mean *dN* and *dS* are reflecting not only *S. glossinidius* evolution but also *E. coli* K12 evolution. In order to test if the observed differences between gene sets in terms of *dN* and *dS* based on pairwise estimates are also present when the lineages of *S. glossinidius* and *E. coli* K12 since their divergence from their common ancestor are analyzed independently, specific *dN* and *dS* values for *S. glossinidius* and *E. coli* branches were calculated and the same statistical analysis were carried out over essential and non-essential genes under the different external conditions.

	CAI (+- SE) [*1]	dN (+- SE) [*1]	dS (+- SE) [*1]
Cofactor biosynthesis genes absent in all minimal networks	0.557 (+- 0.007)	0.249 (+- 0.022)	2.169 (+- 0.117)
Other disposable genes in all minimal networks	0.542 (+- 0.007)	0.270 (+- 0.022)	2.215 (+- 0.104)
Conditionally essential genes in nutrient-limited minimal networks	0.589 (+- 0.005)	0.152 (+- 0.007)	1.793 (+- 0.069)

		CAI P-values	log (dN) P-values	log (dS) P-values
Cofactor biosynthesis genes absent in all minimal networks	Other disposable genes in all minimal networks	P = 0.377	P = 0.857	P = 0.942
	Conditionally essential genes in nutrient-limited minimal networks	P = 0.004	P < 0.0001	P = 0.004
Other disposable genes in all minimal networks	Cofactor biosynthesis genes absent in all minimal networks	p = 0.377	P = 0.857	P = 0.942
	Conditionally essential genes in nutrient-limited minimal networks	P < 0.0001	P < 0.0001	P < 0.0001
Conditionally essential genes in nutrient-limited minimal networks	Cofactor biosynthesis genes absent in all minimal networks	P = 0.004	P < 0.0001	P = 0.004
	Other disposable genes in all minimal networks	P < 0.0001	P < 0.0001	P < 0.0001

Table 5.7: *CAI, dN and dS values for cofactor biosynthesis genes absent in all minimal networks, other disposable genes common to all minimal networks and the set of conditionally essential genes. ANOVA test was applied to evaluate statistically differences between the three groups of genes for the three variables. For dN and dS, natural logarithmic transformation is carried out to accomplish the principles of normality and homocedasticity. CAI values have normal distribution and accomplish homocedasticity principle*

*[*1] The mean CAI, dN and dS for all genes within each group is shown with the corresponding standard error (in parentheses).*

5.3.6 Analysis of specific synonymous and non-synonymous substitution rates in *S. glossinidius* and *E. coli* K12 lineages

dN and *dS* estimates for essential and non-essential genes in minimal networks were calculated by the approximate method of Yang and Nielsen (Yang and Nielsen, 2000) implemented in the program *yn00* included in the PAML software package (Yang, 2007) based on pairwise alignments of *S. glossinidius* and *E. coli* K12 orthologs. These values must be considered as the result of the number of synonymous (*dS*) and non-synonymous (*dN*) substitutions per site occurring in both lineages since their divergence from their common ancestor. However, specific *dN* and *dS* values for *S. glossinidius* and *E. coli* K12 branches since their divergence from their common ancestor can be inferred with the inclusion in the analysis of an

outgroup common to both species, in our case *V. cholerae* O1 (NC_002505) based on phylogenetic reconstructions obtained in the Chapter 3 of this thesis (see Figure 5.4 in Material and Methods section). Specific *dN* and *dS* values for the evolutionary branch leading to *S. glossinidius* and *E. coli* were calculated based on the equations described in Materials and Methods section in order to evaluate if the differences in sequence evolution patterns at both synonymous and non-synonymous sites between essential and non-essential genes in minimal networks are reproduced when lineage-specific *dN* and *dS* values are compared. In addition, comparisons between mean *dN* and *dS* estimates for *S. glossinidius* and *E. coli* K12 allow to test if both species are evolving at similar rates or there is acceleration in one of the two lineages.

When the set of essential and non-essential genes under original conditions were analyzed, 207 out of 255 *S. glossinidius* essential genes present in all minimal networks showed orthologs in *E. coli* K12 and *V. cholerae* O1, whereas for the 127 *S. glossinidius* genes absent in all minimal networks, 85 had orthologs in both species. The mean *dS* and *dN* values for essential and non-essential genes in each lineage are represented in Table 5.8.

		dN_sgl (+- SE) (*1)	dN_eco (+- SE) (*1)	dS_sgl (+- SE) (*1)	dS_eco (+- SE) (*1)
Minimal networks nutrient-limited conditions	Essential genes 1500 minimal networks	0.0911 (+- 0.003)	0.0788 (+- 0.004)	1.236 (+- 0.045)	0.669 (+- 0.045)
	Dispensable genes 1500 minimal networks	0.1136 (+- 0.008)	0.0943 (+- 0.005)	1.275 (+- 0.064)	0.787 (+- 0.059)
	Significance of the difference (*2)	P = 0.009	P = 0.004	P = 0.757	P = 0.059

Table 5.8: *Lineage-specific dN and dS values for essential and disposable genes in minimal networks under nutrient-limited conditions.*
*(*1)* *The mean dN and dS for all genes within each group is shown with the corresponding standard error (in parentheses).*
*(*2)* *Statistical significance of the difference between both groups for a given variable is determined using the Mann-Whitney U test*

Essential genes in both lineages show significantly higher *dN* values than non-essential genes absent in all minimal networks at 0.05 level (p = 0.009 for *S. glossinidius* estimates; p = 0.004 for *E. coli* K12 estimates in Mann-Whitney U-test), reflecting the same trend as observed when pairwise estimates between *S. glossinidius* and *E. coli* K12 were analyzed. In addition, the ratio between mean *dN* values between essential and non-essential genes is similar in both lineages (0.8021 for *S. glossinidius* lineage and 0.8355 for *E. coli* K12 lineage). However, for *dS* estimates, although the mean *dS* value of essential genes is lower than the mean *dS*

of non-essential genes in both lineages, the differences are not statistically significant at 0.05 level (p = 0.757 for *S. glossinidius* estimates; p = 0.059 for *E. coli* K12 estimates in Mann-Whitney u-test). In addition, mean *dN* and *dS* values are always higher in the lineage of *S. glossinidius* rather than in the *E. coli* K12 lineage (see Table 5.8).

Acceleration in substitution rates is a common trend in bacterial endosymbionts in comparison with free-living relatives at both synonymous and non-synonymous sites (Moran, 1996; Brynnel et al., 1998; Clark et al., 1999). In order to test if this acceleration is taking place in the lineage leading to *S. glossinidius*, the differences in *dN* and *dS* estimates between *S. glossinidius* and *E. coli* K12 lineage for each gene were compared through t-test for paired samples for the overall set of genes analyzed (207 genes present in all minimal networks and 85 genes absent in all minimal networks under original conditions). The result of this analysis shows that the average differences in *dN* and *dS* estimates for *S. glossinidius* and *E. coli* K12 lineages are significantly different at 0.05 level ($p < 0.0001$ for *dN_sgl-dN_eco*, $p < 0.0001$ for *dS_sgl-dS_eco* in Paired t-tests) indicating that *S. glossinidius* lineage are accumulating a significantly higher number of non-synonymous and synonymous substitutions per site than *E. coli* K12 lineage in the genes analyzed, being these differences higher at synonymous sites (Mean (*dS_sgl-dS_eco)* = 0.543, Mean (*dN_sgl-dN_eco)* =0.01432). As the divergence time between *S. glossinidius* and *E. coli* K12 since their divergence from their common ancestor is the same in both lineages, these results can be interpreted as a significant acceleration in substitution rates at both synonymous and non-synonymous sites in the lineage of *S. glossinidius*, being this acceleration higher at synonymous sites. However, no reliable substitution rates can be inferred because the exact divergence time between *S. glossinidius* and *E. coli* K12 is not known. The acceleration of substitution rates, especially at synonymous sites, can affect the patterns of codon usage in bacterial genomes. A sign of purifying selection at synonymous sites associated with adaptative codon usage is the presence of negative correlation between CAI and synonymous substitutions per synonymous site (Sharp and Li, 1987b; Sharp, 1991). In order to test if the acceleration in synonymous and non-synonymous substitution rates experienced by *S. glossinidius* is affecting the patterns of codon usage, correlation between specific *dN* and *dS* values for each lineage with their corresponding CAI values were analyzed by Spearman correlation test due to the absence of normal distribution for *dN* and *dS* variables. The results of this analysis are represented in Figure 5.19. There is a significant negative correlation between CAI and *dN* values for both *S. glossinidius* and *E. coli* K12 lineages (rho Spearman = -0.513 for *S. glossinidius*, $p < 0.0001$; rho Spearman = -0.615 for *E. coli* K12, $p < 0.0001$), indicative of a higher conservation of amino acid sequences of highly expressed genes in both lineages (see Figure 5.19-A). However, whereas significant negative correlation between CAI and *dS* values is observed in *E. coli* K12 lineage (rho Spearman = -0.213, $p < 0.0001$), no significant correlation between CAI and *dS*

values is found in *S. glossinidius* lineage (rho Spearman = 0.016, p = 0.788), which indicates that acceleration at synonymous sites in *S. glossinidius* lineage are reducing their adaptative codon usage in comparison with *E. coli* K12 lineage (see Figure 5.19-B).

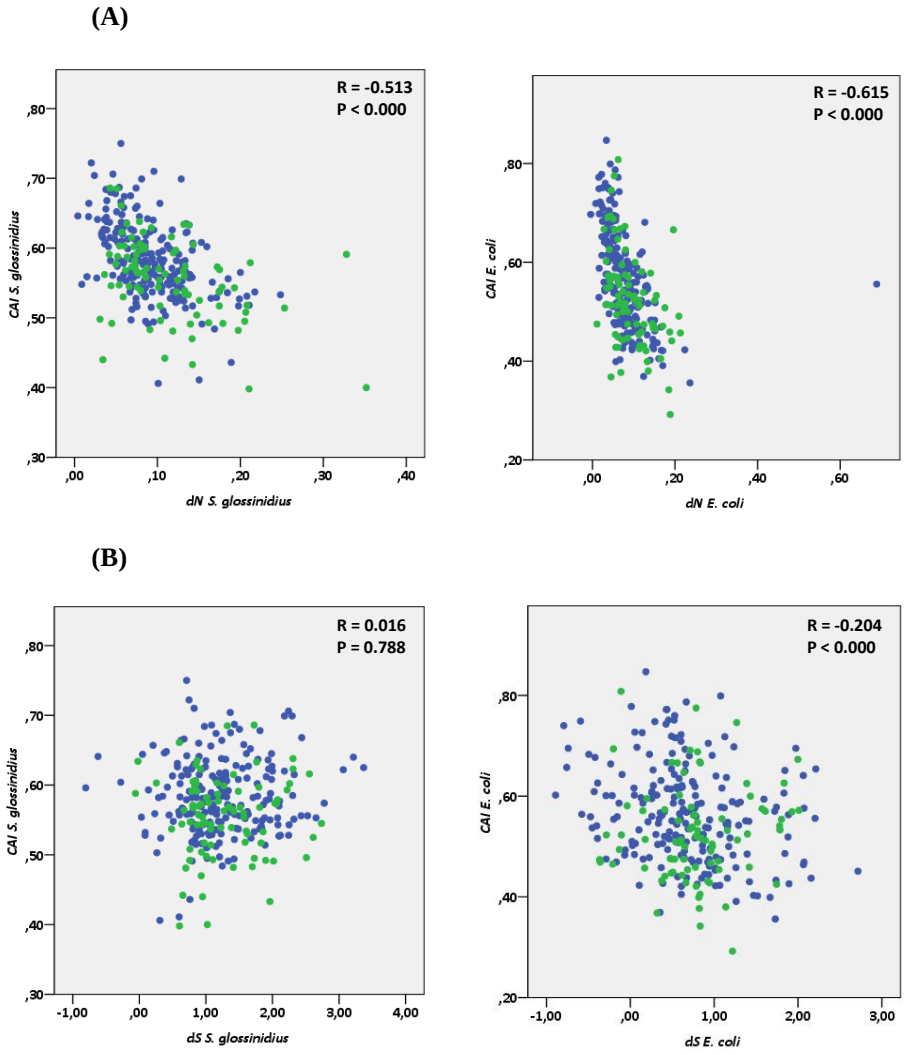


Figure 5.19: *Relationship between lineage-specific CAI and dN (A) and CAI and dS (B) values in S. glossindius and E. coli K12 lineages for essential (blue) and disposable (green) genes in minimal networks under nutrient-limited conditions. Non-parametric test of Spearman was carried out in order to evaluate statistically the correlation between CAI-dN and CAI-dS for each lineage.*

Similar analyses were carried out with the sets of essential and non-essential genes defined under nutrient-rich conditions. As was previously described when pairwise *dN* and *dS* estimates between *S. glossinidius* and *E. coli* K12 were analyzed, essential genes under nutrient-rich conditions showed significantly lower mean *dN* and *dS* values than non-essential genes absent in all minimal networks, but the differences between both gene sets were minor than differences observed when essential and non-essential genes under nutrient-limited conditions were analyzed. Similar to what happens in the analysis of essential and non-essential genes under original conditions, the number of genes in each set for which specific *dN* and *dS* values for each lineage can be inferred is reduced as consequence of *V. cholerae* O1 ortholog identification (114 out of 139 essential genes and 73 out of 111 non-essential genes with ortholog in *V. cholerae* O1). However, when both gene sets were compared in each lineage, no statistically significant differences in mean *dN* and *dS* values between essential and non-essential gene sets were found, in contrast with the results obtained when pairwise estimates of *dN* and *dS* between *S. glossinidius* and *E. coli* were analyzed (See Table 5.9).

		dN_sgl (+- SE) (*1)	dN_eco (+- SE) (*1)	dS_sgl (+- SE) (*1)	dS_eco (+- SE) (*1)
Minimal networks nutrient-rich conditions	Essential genes 1500 minimal networks	0.0993 (+-0.004)	0.0908 (+- 0.008)	1.157 (+- 0.06)	0.849 (+- 0.06)
	Dispensable genes 1500 minimal networks	0.116 (+- 0.007)	0.0993 (+- 0.006)	1.311 (+- 0.07)	0.732 (+- 0.08)
	Significance of the difference (*2)	P = 0.077	P = 0.070	P = 0.167	P = 0.297

Table 5.9: *Lineage-specific dN and dS values for essential and disposable genes in minimal networks under nutrient-rich conditions.*
*(*1) The mean dN and dS for all genes within each group is shown with the corresponding standard error (in parentheses).*
*(*2) Statistical significance of the difference between both groups for a given variable is determined using the Mann-Whitney U test*

This can be consequence of the more conserved patterns of sequence evolution of conditionally essential genes that are not present in all minimal networks when nutrient-rich conditions were analyzed. As was previously observed, this set of conditionally essential genes shows significantly higher levels of sequence conservation at both synonymous and non-synonymous sites than the set of essential genes present in all minimal networks under all conditions. However, when individual lineages are analyzed, differences in the patterns of sequence evolution of

these conditionally essential genes in *S. glossinidius* and *E. coli* K12 lineage were observed (See Table 5.10).

	dN_sgl (+- SE) (*1)	dN_eco (+- SE) (*1)	dS_sgl (+- SE) (*1)	dS_eco (+- SE) (*1)
Conditionally essential genes in nutrient-limited conditions	0.0811 (+-0.004)	0.0642 (+- 0.003)	1.333 (+- 0.07)	0.448 (+- 0.06)
Common essential genes in all minimal networks	0.0993 (+- 0.004)	0.0908 (+- 0.007)	1.157 (+- 0.06)	0.849 (+- 0.06)
Significance of the difference (*2)	P = 0.006	P = 0.001	P = 0.055	P = 0.000

Table 5.10: *Lineage-specific dN and dS values for conditionally essential genes only under nutrient-limited conditions and common essential genes in all minimal networks in all conditions.*
*[*1] The mean dN and dS for all genes within each group is shown with the corresponding standard error (in parentheses).*
*[*2] Statistical significance of the difference between both groups for a given variable is determined using the Mann-Whitney U test*

In *E. coli* K12 lineage, conditionally essential genes show significantly lower mean *dN* and *dS* values than the core of essential genes in all minimal networks under all conditions, in concordance with the results observed when pairwise estimates were analyzed, that demonstrates the essential character of this conditionally essential genes for a free-living bacterium that must synthesize all biomass constituents by itself. However, in *S. glossinidius* lineage, whereas conditionally essential genes show significantly lower mean *dN* values than essential genes present in all minimal networks in all conditions (mean *dN_sgl* = 0.0811 for conditionally essential genes, mean *dN_sgl* = 0.0993 for common essential genes in all minimal networks; p = 0.006 in Mann-Whitney u-test), the inverse situation is found when *dS* values are analyzed. Conditionally essential genes show higher *dS* values than essential genes in all minimal networks under both conditions (mean *dS_sgl* = 1.333 for conditionally essential genes, mean *dS_sgl* = 1.157 for common essential genes in all minimal networks), although these differences are not statistically significant at 0.05 level in Mann-Whitney u-test (p = 0.055). This indicates that the accelerated rates of sequence evolution at both synonymous and non-synonymous sites are affecting the set of conditionally essential genes in *S. glossinidius* in a different manner as in *E. coli* K12 lineage. Whereas *E. coli* K12

shows significantly more restricted pattern of sequence evolution in these conditionally essential genes at both synonymous and non-synonymous sites, probably due to the selective importance of this genes in a free-living environment, *S. glossinidius* shows a more relaxed pattern of sequence evolution in this gene set. In fact, despite conditionally essential genes have significantly lower mean *dN* values than core essential genes in all minimal networks, the ratio between mean *dN* values of both groups in *S. glossinidius* lineage is higher (0.8165) than the ratio in *E. coli* K12 lineage (0.7072). These differences are more evident when the same ratios for mean *dS* values in both lineages are compared, with *E. coli* K12 lineage that have a ratio of 0.5275 between mean *dS* values of conditionally essential genes and the rest of essential genes in all minimal networks whereas in *S. glossinidius* lineage this ratio is equal to 1.15. These results point out to a major role of conditionally essential genes, specifically in the *E. coli* K12 lineage, that limitis their rates of sequence evolution at both synonymous and non-synonymous sites. On the other hand, in *S. glossinidius* their relevance would be much less due to their nutrient-rich environment and, as consequence, their rates of sequence evolution appear increased, specially at synonymous sites, consequence of the acceleration in substitution rates in this lineage. In fact, when CAI values between both gene sets are compared in both lineages, conditionally essential genes in *E. coli* K12 have significantly high values of mean CAI than the core of essential genes present in all minimal networks under all conditions (mean CAI = 0.5870 for conditionally essential genes, mean CAI = 0.5486 for core metabolic genes in all minimal networks; p = 0.004 on Student´s t-test), indicative of significantly higher expression level and in concordance with their more relevant functional role, whereas in *S. glossinidius* this differences are not statistically significant (mean CAI = 0.5920 for conditionally essential genes, mean CAI = 0.5774 for core metabolic genes in all minimal networks; p = 0.057 on Student´s t-test).

Finally, the patterns of sequence evolution of cofactor biosynthesis genes absent in all minimal networks in both conditions in *S. glossinidius* and *E. coli* K12 lineages were also analyzed to determine if the closer proximity to the rest of non-essential genes in all minimal networks observed based on pairwise estimates of *dN* and *dS* are consequence of *S. glossinidius* evolution within their association with the tsetse host or is a common feature also present in *E. coli* K12 lineage. *dN* and *dS* values for cofactor biosynthesis genes absent in all minimal networks (42 out of 48 with orthology in *V. cholerae* O1), the rest of genes absent in all minimal networks under all conditions (30 out of 61 genes with orthology in *V. cholerae* O1) and the set of conditionally essential genes (93 out of 117 genes with orthology in *V. cholerae* O1) were compared in each lineage through ANOVA test followed by Scheffe correction for *post-hoc* comparisons between groups. The results are represented in Table 5.11.

	dN_sgl (+- SE) (*1)	dN_eco (+- SE) (*1)	dS_sgl (+- SE) (*1)	dS_eco (+- SE) (*1)
Cofactor biosynthesis genes absents in all minimal networks	0.119 (+- 0.007)	0.099 (+- 0.008)	1.407 (+- 0.095)	0.668 (+- 0.090)
Other disposable genes in all minimal networks	0.111 (+- 0.012)	0.099 (+- 0.009)	1.151 (+- 0.108)	0.862 (+- 0.094)
Conditionally essential genes in nutrient-limited minimal networks	0.081 (+- 0.004)	0.064 (+- 0.003)	1.333 (+- 0.070)	0.448 (+- 0.056)

		log(dN_sgl) P-values	log(dN_eco) P-values	log(dS_sgl) P-values	log(dS_eco) P-values
Cofactor biosynthesis genes absents in all minimal networks	Other disposable genes in all minimal networks	P = 0.427	P = 0.918	P = 0.275	P = 0.906
	Conditionally essential genes in nutrient-limited minimal networks	P < 0.0001	P = 0.001	P = 0.490	P = 0.016
Other disposable genes in all minimal networks	Cofactor biosynthesis genes absents in all minimal networks	p = 0.427	P = 0.918	P = 0.275	P = 0.906
	Conditionally essential genes in nutrient-limited minimal networks	P = 0.069	P = 0.001	P = 0.737	P = 0.007
Conditionally essential genes in nutrient-limited minimal networks	Cofactor biosynthesis genes absents in all minimal networks	P < 0.0001	P = 0.001	P = 0.490	P = 0.016
	Other disposable genes in all minimal networks	P = 0.069	P = 0.001	P = 0.737	P = 0.007

Table 5.11: *Lineage-specific dN and dS values for cofactor biosynthesis genes absent in all minimal networks, other disposable genes common to all minimal networks and the set of conditionally essential genes under nutrient-limited conditions. ANOVA test was applied to evaluate statistically significant differences between the three groups of genes for each variable. Natural logarithmic transformation is carried out to accomplish the principles of normality and homocedasticity*

*(*1) The mean dN and dS for all genes within each group is shown with the corresponding standard error (in parentheses).*

E. coli K12 lineage shows similar evolutionary patters as observed when pairwise *dN* and *dS* estimates were analyzed. Statistically significant differences at 0.05 level were detected only for the set of conditionally essential genes and the other two groups, with cofactor biosynthesis genes and the rest of non-essential genes showing significantly higher values of *dN* and *dS* than conditionally essential genes. This indicates that the patterns of sequence evolution of cofactor biosynthesis genes absent in all minimal networks in *E. coli* K12 lineage are closer to the rest of non-essential genes for the viability of the system, and significantly higher at both synonymous and non-synonymous sites than the set of conditionally essential genes, that appears as the more conserved group as was previously described. In concordance with this relaxed pattern of sequence evolution of this cofactor biosynthesis genes in *E. coli* K12 and the acceleration in substitution rates in *S. glossinidius* lineage, specially at synonymous sites, there is no significant

differences between the three groups of genes in *dS* estimates for *S. glossinidius*, whereas for *dN* there is significant differences only between cofactor biosynthesis genes and the set of conditionally essential genes, with cofactor biosynthesis genes having significantly high average *dN* values than the set of conditionally essential genes at 0.05 level.

5.4 DISCUSSION

Minimal genomes of bacterial endosymbionts of insects represent final stages of the general evolutionary process of genome reduction from larger genomes of free-living ancestor. The availability of completely sequenced genomes of different bacterial endosymbionts has allowed the study of the genome reduction process from different points of view. Comparative genomics has allowed to trace the process of gene loss at different branches of the phylogeny in *Buchnera aphidicola* (Gomez-Valero et al., 2004a), *Mycobacterium leprae* (Gomez-Valero et al., 2007), or the *Ricketsia* genus (Boussau et al., 2004; Blanc et al., 2007), and to compare the metabolic profiles of different bacterial genomes corresponding to different symbiotic associations (Zientz et al., 2004), as well as identify minimal gene sets needed to maintain a minimal metabolic system (Gil et al., 2004). In the present study, an approach to the reductive evolution process based on systems biology is carried out, consisting in the reconstruction of the genome-scale metabolic networks of *S. glossinidius* at different moments of its evolution from its free-living ancestor. This study has two main differentiated purposes. First, the determination of the most relevant changes linked to the transition from a hypothetical free-living ancestor to the actual stage associated with the tsetse host, and second, the prediction of the possible outcomes of the reductive evolution process from the present genome. This is carried out based on the capability of steady-state analysis of genome-scale metabolic networks to evaluate quantitatively the functional capabilities of the whole metabolic system through FBA using as objective function biomass production, which has been demonstrated as a reliable approximation to the growth phenotype of the organism under study (Covert et al., 2001; Edwards et al., 2001; Price et al., 2004). In this context, *S. glossinidius* represents an ideal model system to carry out this type of study, because the metabolic network of its hypothetical ancestor can be reliable inferred from the complete set of genes and pseudogenes described in the previous chapter of this thesis, whereas the transition to the functional network can be simply accomplished by considering only the functional genes in the network reconstruction. In addition, the availability of *E. coli* JR904 metabolic network helps to the network reconstruction process by means of orthologous identification between both genomes from which an initial backbone of *S. glossinidius* ancestral network is retrieved that is further refined by manual curation and addition of *S. glossinidius* specific metabolic and transport activities, allowing also to compare functional inferences with a reference free-living organism

from which extensive research on its metabolic network structure and function has been carried out (Edwards and Palsson, 2000a; Reed et al., 2003; Feist et al., 2007).

The results of the reconstruction process renders an ancestral metabolic network of *S. glossinidius* comprised by 741 reactions and 668 gene products that includes 479 genes, 148 pseudogenes and 41 genes specifically deleted in the lineage leading to *S. glossinidius* after the divergence from the common ancestor with *E. coli* K12. Most of these genes and reactions are present in *E. coli* JR904 metabolic network, although *S. glossinidius* specific metabolic activities are also represented by the inclusion of 27 pseudogenes and 13 genes without orthology with *E. coli* K12 genome reflecting differences in the composition of bacterial lipopolysaccharide, the capability of ancestral *S. glossinidius* to growth with glycerol as carbon source, or the capability to produce methionine from both acetyl and succinyl coenzyme A. From the ancestral network, *S. glossinidius* functional network is obtained by removing all reactions catalyzed by pseudogenes or deleted genes, rendering a functional metabolic network comprised by 458 genes and 560 reactions. In order to evaluate the functional phenotype of both metabolic networks, FBA with biomass production as objective function has been employed using the biomass equation of *E. coli* JR904 network (Reed et al., 2003). The biomass reaction represents a weighted ratio of the components forming the dry weight of a cell together with the energy demands for cellular growth and maintenance in the form of ATP hydrolysis, and their utilization as objective function in FBA simulations allow to test the network capability to support growth (Feist et al., 2009). Specific biomass composition can be inferred from the detailed cellular composition of cell cultures of the organism under study (Edwards and Palsson, 2000a; Forster et al., 2003a; Lee et al., 2009). In absence of this information for *S. glossinidius*, the biomass equation of *E. coli* JR904 is considered. This equation has been employed in different metabolic reconstructions and is considered as a valid approach if the real biomass composition of the modeled organism is not known (Puchalka et al., 2008; Oberhardt et al., 2008; Zhang et al., 2009). In fact, comparison of the biomass production rates by FBA with *E. coli* JR904 as objective function in *Pseudomonas putida* metabolic network with experimental measures of their cellular growth on cell cultures reveals high accuracy of FBA predictions with cellular growth observations, whereas variations in the composition of biomass constituents on the outcomes of FBA simulations have nearly negligible effect on the final biomass production rate (Puchalka et al., 2008). In the case of *S. glossinidius*, its close evolutionary relationship with *E. coli* K12 and the absence of detailed information about its cellular composition justifies the utilization of biomass equation of *E. coli* JR904 to test the functionality of *S. glossinidius* metabolic networks. In addition, the utilization of the same biomass equation allows to compare the results of FBA simulations over *S. glossinidius* and *E. coli* JR904 metabolic networks, in order to determine the relative degree of metabolic independency of *S. glossinidius* network reconstructions and the metabolites that the external environment must supply in

order to achieve a viable phenotype in terms of biomass production, equivalent to cellular growth. The results of FBA simulations in a minimal aerobic environment with only glucose as external metabolites reveals a complete functionality of *S. glossinidius* ancestral network at biomass production rates very similar to that of *E. coli* JR904 network, a feature that is in concordance with the very recent transition of *S. glossinidius* to a host-dependent lifestyle, because its ancestral network is completely functional in a minimal environment with only glucose as external carbon source from which to produce all biomass constituents. This supposes an additional evidence of the recent transition to a host-dependent lifestyle in addition to the lack of coevolution between phylogenies of *S. glossinidius* and *Glossina* genus. (Chen et al., 1999), of its unique capability among bacterial endosymbionts of insects to be cultured in-vitro (Welburn et al., 1987), and of its particular genome features, including a large genome size and the presence of a massive number of pseudogenes described in the original genome paper (Toh et al., 2006) and analyzed in the previous chapter of this thesis. By contrast, the inactivation of glycogen and specially the genes involved in arginine biosynthesis and the gene *ppc* encoding the central anaplerotic enzyme phosphoenolpyruvate carboxylase makes the functional network of *S. glossinidius* lethal under minimal environment with only glucose as external carbon source. The removal of glycogen from the biomass equation and the addition of an external source of L-arginine produces a viable phenotype in terms of biomass production. The inactivation of *ppc* gene has particular lethal consequences even under *S. glossinidius* ancestral metabolic network. Phosphoenolpyruvate carboxylase (*Ppc*) is a central anaplerotic enzyme that replenish the TCA cycle intermediate oxaloacetate from phosphoenolpyruvate, being essential as carbon supply for the biosynthesis of 10 amino acids (aspartate, asparagine, methionine, threonine, isoleucine and lysine directly derived from oxalacetate and glutamate, arginine, proline and glutamine derived from α-ketoglutarate) and other cellular building blocks derived from TCA cycle intermediates (March et al., 2002). Experimental measures of *Ppc* activity on *E. coli* based on ^{13}C-labelling experiments followed by measurements of the isotopomer distribution by gas chromatography-mass spectrometry (GC-MS) reveals that 50.7% of the carbon flux in wild type strains is channeled through this enzyme (Peng et al., 2004), and that knockout mutants of *ppc* gene activates the glyoxylate bypass of TCA cycle, an alternative anaplerotic pathway that replenish oxalacetate in response to the blockage through *Ppc* (Peng et al., 2004; Peng and Shimizu, 2004). This experimental results are reproduced in our *in-silico* FBA simulations over *E. coli* JR904 and *S. glossinidius* ancestral network, where under glucose as unique carbon source the flux proceeds through *Ppc* reaction connecting glycolytic pathway with TCA cycle (See Figure 5.8). In absence of *Ppc*, *E. coli* JR904 network activates enzymes of the glyoxylate bypass to replenish oxalacetate consumed for biosynthetic purposes, but in *S. glossinidius* ancestral network, that lacks this anaplerotic pathway, the single deletion of *Ppc* reaction renders a lethal phenotype in terms of biomass production.

The addition of an external source of arginine constitutes an essential requirement for the functionality of the functional network of *S. glossinidius* due not only to their own role as biomass constituent and as a precursor of spermidine and putrescine (also biomass constituents), but also as energy source through putrescine degradation to TCA cycle intermediate succinate that supplies the anaplerotic role of pseudogenized *Ppc* (see Figure 5.8). This indicates that a single inactivation of a few essential enzymatic activities (specially arginine biosynthesis genes and phosphoenolpyruvate carboxylase *ppc* gene) produces a drastic change in the metabolic capabilities of *S. glossinidius* metabolic network from an ancestor completely functional as a free-living bacteria to their actual state, that depends on external supply of arginine to produce a viable phenotype on FBA simulations.

Comparison of metabolic network behavior under different carbon sources with experimental data is also a common practice in order to evaluate the predictive capability of metabolic models (Edwards et al., 2001; Puchalka et al., 2008; Oberhardt et al., 2008). Similar analysis with *S. glossinidius* ancestral and functional networks to evaluate the capability to metabolize 27 different carbon sources reveals coincident results between network simulations and experimental data for 19 out of 27 carbon sources tested (70.37% of accuracy) (Dale and Maudlin, 1999), whereas 8 false positives were detected in the sense of metabolites for which the functional network renders a viable phenotype whereas no growth is reported experimentally. In two of these cases (ethanol and pyruvate), their positive growth appears associated with free diffusion of these metabolites to the cytoplasmic compartment in *E. coli* JR904 network that has been assumed in *S. glossinidius* ancestral and functional network. However, in other cases, the observed phenotype in terms of experimental cellular growth contradicts the expected behavior based on genome sequences, like in the case of the metabolization of glucose but not fructose and mannose reported by Dale and Maudlin (Dale and Maudlin, 1999) despite the presence of functional PTS system for mannose and the pseudogenization of the PTS systems and additional transporters for glucose and fructose. The broad substrate-specificity of mannose PTS system has been previously described (Curtis and Epstein, 1975; Postma et al., 1993; Kornberg, 2001), whereas there is also reports on mutational changes on PTS system components that change substrate specificity in response to specific external conditions (Oh et al., 1999; Notley-McRobb and Ferenci, 2000), so the observed behavior of *S. glossinidius* in cell cultures under glucose and mannose as carbon sources points out to a possible change in substrate specificity of mannose PTS system for growth with glucose.

The robustness of *S. glossinidius* metabolic networks to gene deletion events and to changes in the metabolic fluxes across particular reactions has been also analyzed. Different studies over the structure of natural metabolic systems have concluded that the power-law distribution that governs the connectivity of natural networks makes these systems robust to random removal of nodes but sensitive to direct attacks,

being this a common organizational properties of large-scale network systems (Jeong et al., 2000; Albert et al., 2000; Podani et al., 2001; Barabasi and Oltvai, 2004). Most of these studies are focused on the topological properties of metabolic networks, analyzing how random removal of network nodes affects to topological parameters like connectivity, clustering coefficient, average path length, or network diameter, but this is also observed when collections of single knockouts for a particular genome are generated. For example, only 119 out of 3888 single knockout mutants of E. coli renders a lethal phenotype under glycerol-minimal medium, with 91% of these predictions being predicted also on FBA simulations over *E. coli* JR904 network (Joyce et al., 2006). However, in our simulations, the robustness of the networks to gene deletion events decreases significantly as consequence of the gene inactivation process, with the fraction of lethal knockouts producing more than 99% of decrease in the biomass production rate on FBA simulations increasing since 23.59% of network genes in *E. coli* JR904 metabolic network to 55.02% of the genes in *S. glossinidius* functional network. In fact, this fraction increases to 84% of network genes in the metabolic network of *B. aphidicola* from the pea aphid *Acyrtosiphon pisum*, the unique metabolic model available from an endosymbiotic bacterium up to date (Thomas et al., 2009), reflecting a significant reduction of the robustness of metabolic systems to gene deletion events associated with different stages of the reductive evolution process. A similar trend was observed by Gabaldon and collaborators when the robustness of a minimal network inferred from an hypothetical minimal gene set of 206 genes proposed by Gil and collaborators were analyzed (Gil et al., 2004), where most mutations had limited effect on the overall topology of the system in terms of average path length and network diameter. However, when the deleterious effect of the removal of a single enzyme is analyzed, the deletion of most enzymes of the minimal network (76%) prevents the biosynthesis of at least one metabolite, that would be equivalent to a lethal phenotype in FBA simulations with biomass as objective function, in contrast with results with *E. coli* K12 metabolic network, where most mutations produces no network damage (Gabaldon et al., 2007). In addition, gene inactivation also affects to the range of metabolic fluxes that network reactions can admit while producing a viable phenotype in terms of biomass production. This analysis has been carried out only with reactions involved in the glycolysis, but a decrease in the range of reaction fluxes over which the system is able to produce a functional phenotype is observed from *E. coli* JR904 to *S. glossinidius* networks (See Figure 5.12). These results indicate that gene inactivation process reduces the robustness of the system not only to gene deletion events but also to variations in enzymatic activities in comparison with free-living bacteria. In the context of the association with eukaryotic hosts, this reduction of system robustness can be associated to their strict location in a more stable and nutrient-rich environment like that found within the insect host, where environmental fluctuations are less pronounced than in free-living environment (Moran and Plague, 2004; Silva et al., 2007), indicating a reduction in the adaptability of the system to changes in their gene content and enzymatic activity

consequence of the gene inactivation process. This effect will be probably enhanced in advanced stages of the genome reduction process due to the loss of many of the genes involved in DNA repair pathways (Sharples, 2009), that can be associated with the even tight association of long-term bacterial endosymbionts in the more stable intracellular environment of bacteriocyte cells, in comparison with the wider tissue tropism both intra and extracellularly associated with the more recent symbiotic association of *S. glossinidius* with the tsetse host.

In addition to study the transition to the host-dependent lifestyle from a hypothetical free-living ancestor, an approach to the possible future evolution of *S. glossinidius* in the context of the reductive evolution process has been carried out based on reductive evolution simulations over functional network of *S. glossinidius* that produces a set of minimal networks capable to produce functional phenotype in terms of biomass production under different external conditions. The concept of the minimal genomes and minimal gene sets is a common issue in comparative genomics based on the assumption that genes conserved across distantly related evolutionary groups are good candidates to be considered as essential genes. Based on this idea, minimal gene sets has been defined based on the identification of orthologous gene sets across different genomes (Mushegian and Koonin, 1996; Koonin, 2000; Klasson and Andersson, 2004). These computational approaches has been complemented with experimental identification of essential genes in different bacterial genomes though transposon mutagenesis (Hutchison et al., 1999; Akerley et al., 2002; Jacobs et al., 2003; Gerdes et al., 2003), translation inhibition by antisense RNA (Forsyth et al., 2002) or vector insertion (Kobayashi et al., 2003). Most of these computational and experimental minimal gene sets are enriched with genes involved in genetic information-processing systems like genes responsible of replication, transcription and translation, but in order to produce a viable phenotype in terms of cell growth, reliable minimal gene sets must include metabolic genes necessary to maintain metabolic homeostasis (Szathmary, 2005; Pereto, 2005). Based on a combination of comparative genomics and experimental approaches, a minimal gene set of 206 genes that includes a minimal metabolic machinery necessary to sustain life has been proposed (Gil et al., 2004). In this context, stochiometric analysis of metabolic networks by FBA allows to evaluate until what point a particular metabolic system can be reduced while maintains a viable phenotype in terms of biomass production, equivalent to cellular growth, allowing to quantify the functionality of minimal metabolic systems. The predictive power of these approaches has been demonstrated through reductive evolution simulations over *E. coli* JR904 metabolic networks, which under external conditions that mimics the environments that *Buchnera aphidicola* and *W. glossinidia* found within their insect host, arrives to minimal gene sets that reproduces with high accuracy their respective gene content (Pal et al., 2006b). In our simulations, gene and reaction content of minimal networks generated from *S. glossinidius* metabolic network is strongly dependent of the external metabolites available for uptake. Whereas under

minimal conditions with only arginine and glucose as external metabolites all minimal networks are highly similar, sharing 84.6% of the reactions and 87.4% of their genes, more variability is found when the simulations are carried out under a nutrient-rich environment with 41 different metabolites available for uptake, with 60.86% of the reactions and 58.69% of the genes shared by all minimal networks. This suggests that environmental conditions play a decisive role in the potential reductive patterns of *S. glossinidius*, and that minimal gene sets or minimal genomes must be defined in the context of the environment associated with the corresponding organism. In the case of *S. glossinidius*, the high levels of identity between minimal networks, especially in a nutrient-limited environment, is also consequence of their streamlined functional network consequence of the gene inactivation process. Minimal networks under both conditions differ mainly in the number of genes that appears as essential under nutrient-limited environment due to their role in the biosynthesis of essential biomass constituents but that can be lost under nutrient-rich environment if these biomass constituents are available from the outside (137 conditionally essential genes). This does not mean that these genes are non-essential under nutrient-rich conditions, but that their essentiality becomes defined by the random order of the gene inactivation process. A comparison of the minimal networks generated under both conditions with the gene content of *W. glossinidia* reveals that simulations under nutrient-rich environment predicts with higher accuracy the gene content of a real minimal genome evolved under the same conditions as *S. glossinidius*, with 69% of the essential genes in all minimal networks under nutrient-rich conditions being present in the genome of *W. glossinidia*. These results are similar to those obtained when similar reductive evolution simulations were carried out with *E. coli* JR904 network under similar nutrient-rich environment (Pal et al., 2006b), revealing the potential of network analysis in the prediction of gene content evolution. Differences between minimal networks and *W. glossinidia* gene content are also observed mainly due to the metabolites included in the biomass reaction formula that is used as objective function in FBA simulations. This is the case of 48 genes involved in cofactor biosynthesis present in *W. glossinidia* (for ubiquinone, biotin, heme groups and pyridoxine 5-phosphate biosynthesis) that are absent in all minimal networks under all conditions because these cofactors are not included in the biomass reaction, and alternatively 43 genes present in all minimal networks in all conditions, needed to produce some of the biomass constituents like bacterial LPS, tetrahydrofolate, or the polyamines purine and spermidine, but that are absent in the genome of *W. glossinidia*. Higher accuracy in the predictions would be possible with precise adjustments of biomass equation composition in order to define the precise biosynthetic capabilities of *W. glossinidia*, but this biomass composition should not include tetrahydrofolate or Coenzyme A biosynthesis *de-novo* unless an external source of p-aminobenzoate or β-alanine would be available, that contrasts with the assumed biosynthetic capability de-novo for this two cofactors by *W. glossinidia* (Akman et al., 2002; Zientz et al., 2004).

Finally, the patterns of sequence evolution of essential and non-essential genes defined based on reductive evolution simulations were analyzed in order to evaluate if the essential character of metabolic genes corresponds with restricted patterns of sequence evolution. This hypothesis was first proposed by Alan Wilson and colleagues in 1977 as the so-called "knockout-rate" prediction, that postulates that two proteins subject to the same level of functional constraints but differing in their dispensability will evolve at different rates (Wilson et al., 1977; Hurst and Smith, 1999). This prediction has been confirmed for three bacterial lineages (*E. coli*, *Helicobacter pylori* and *Neisseria meningitidis* lineages) based on experimentally defined knockouts in *E. coli* genome, with essential genes showing significantly higher values of *dN* and *dS* than non-essential genes (Jordan et al., 2002), although related analysis in eukaryotic lineages have not found such correlation at significant levels (Hurst and Smith, 1999). When pairwise estimates of *dN* and *dS* were analyzed between essential and non essential genes defined in reductive evolution simulations, essential genes shows significantly more conservation at both synonymous and non-synonymous sites than non-essential genes in minimal networks under both nutrient-limited and nutrient-rich environment, in concordance with the results of Jordan and collaborators (Jordan et al., 2002), and that can be explained by the "knockout-rate" prediction. In addition, similar comparative analysis with the CAI values of *S. glossinidius* essential and non-essential genes reveals a significant higher CAI values in essential genes in comparison with non-essential genes in all minimal networks under both nutrient-rich and nutrient-limited conditions. These results were also found when CAI values of experimentally defined essential genes in *E. coli* and *Bacillus subtilis* were analyzed, where essential genes tend to have higher CAI and sequence conservation than non-essential genes (Fang et al., 2005). In the context of the genome reduction process, it has been demonstrated that genes with low CAI and higher dN values corresponds to genes less selectively constrained that has been predominantly lost in ancient endosymbiotic lineages since their divergence from a free-living ancestor (Delmotte et al., 2006), whereas putatively highly expressed genes in these endosymbiotic lineages appear also more evolutionary constrained at both synonymous and non-synonymous sites despite their increased mutational rates that produce the lack of adaptative codon usage, that is mainly shaped by mutational bias towards AT rich codons (Herbeck et al., 2003; Schaber et al., 2005). This indicates that the evolutionary patterns of non-essential genes defined in the context of reductive evolution simulations correspond to genes that could be lost in the future evolution of *S. glossinidius* associated with the process of genome reduction. However, whereas CAI estimates are specific of *S. glossinidius* genome, pairwise estimates of *dN* and *dS* between *S. glossinidius* and *E. coli* K12 are the result of the substitutions taking place on both lineages after their divergence from their common ancestor. Accelerated rates of sequence evolution in bacterial endosymbionts compared with free-living bacteria is a common feature in most endosymbiotic lineages, with extremely large increase in the non-synonymous and synonymous substitution rates

during their evolution since transition to host-dependent lifestyle in comparison with free-living bacteria (Moran, 1996; Brynnel et al., 1998; Clark et al., 1999). Although the presence of adaptative codon usage and the lack of compositional bias towards AT suggest that acceleration in substitution rates are not affecting *S. glossinidius* to the same strength as long-term bacterial endosymbionts, the comparison of *dN* and *dS* estimates for *S. glossinidius* and *E. coli* K12 lineages since the divergence from their common ancestor reveals significant higher *dN* and *dS* values in the evolutionary branch leading to *S. glossinidius*. Because they occur in the same evolutionary time in both lineages, these results can be interpreted as a significant acceleration in the substitution rates at both synonymous and non-synonymous sites in the lineage leading to *S. glossinidius*, although exact estimates of substitution rates cannot be inferred because divergence time between both lineages is unknown. This acceleration in substitution rates, especially at synonymous sites, is expected to affect the patterns of codon usage in *S. glossinidius* in comparison with *E. coli* K12. Whereas negative correlation between CAI and *dN* is indicative of increased conservation of highly expressed genes (Herbeck et al., 2003; Rocha and Danchin, 2004), negative correlation between CAI and *dS* is consequence of purifying selection acting on synonymous sites in highly expressed genes consequence of adaptative codon usage (Sharp and Li, 1987b; Sharp, 1991). In fact, in ancient bacterial endosymbionts like *B. aphidicola*, homogeneous distribution of synonymous substitutions among loci is indicative of the lack of adaptative codon usage due to the increased substitution rates and the compositional bias towards AT (Clark et al., 1999). When a similar analysis is carried out with *S. glossinidius* and *E. coli* K12 lineages, although significant negative correlation is found between CAI and *dN* values in both lineages, no significant correlation between CAI and *dS* is observed for *S. glossinidius* lineage, consequence of acceleration in substitution rates at synonymous sites (See Figure 5.19). This result points out to a decrease in the intensity of adaptative codon usage in comparison with free-living bacteria like *E. coli* K12 rather than as a lack of adaptative codon usage in *S. glossinidius*. In fact, negative correlation between CAI and *dS* values for *E. coli* K12 lineage, although significant, is lower (-0.213) than correlation observed when the analysis is carried out in the lineage between *E. coli* K12 and *Salmonella typhimurium* (-0.68) (Sharp and Li, 1987b; Sharp, 1991), a feature that can be explained by the large evolutionary time over which our estimates of *dN* and *dS* values are inferred. In the case of *S. glossinidius*, codon usage bias has been demonstrated by the significant clustering of putatively high-expressed genes (ribosomal protein genes) in correspondence analysis of the Relative Synonymous Codon Usage (RCSU) of all genes of the genome (Puigbo et al., 2008b) (See Figure 5.20), so the lack of correlation between CAI and *dS* can be attributed to the large divergence time over which the estimates of *dS* have been made and the acceleration in synonymous substitution rates in the lineage of *S. glossinidius*, although this acceleration can reduce the strength of adaptative codon usage in comparison with the lineage of *E. coli* K12.

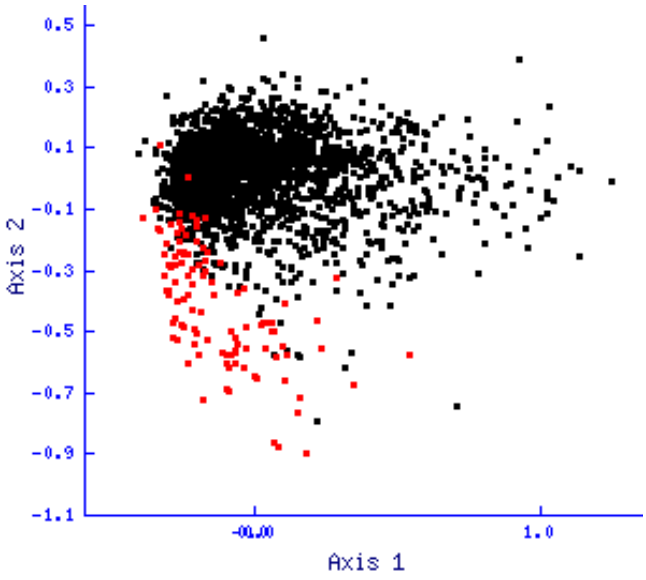


Figure 5.20: *Correspondence analysis plot of the Relative Synonymous Codon Usage (RSCU) of all S. glossinidius genes extracted from http://genomes.urv.cat/HEG-DB / . Axes correspond to the correspondence factors. Red squares correspond to the predicted group of highly expressed genes and black squares correspond to the rest of genes of the genome.*

Acceleration in substitution rates in *S. glossinidius* lineage is further evident when the set of conditionally essential genes were analyzed. These genes appear as the most conserved at both synonymous and non-synonymous sites when pairwise estimates of *dN* and *dS* were analyzed. These conditionally essential genes can be considered as genes susceptible to be lost in a nutrient-rich environment similar to that *S. glossinidius* found within the tsetse host, and in fact it contains most of the genes involved in amino acid biosynthesis that are absent in the genome of the more ancient tsetse endosymbiont *W. glossinidia*. Lineage-specific analysis of these conditionally essential genes reveals a significant conservation at both synonymous and non-synonymous sites of these genes in *E. coli* K12 lineage, in concordance with the essential role that these genes must fulfill for a free-living bacterium with more restricted environmental conditions. In contrast, in *S. glossinidius* lineage, although significant conservation of conditionally essential genes is observed for non-synonymous substitutions in comparison with the rest of essential genes for all minimal networks, differences between both groups are lower than in *E. coli* K12 lineage, whereas for synonymous sites the inverse trend is observed, with conditionally essential genes accumulating higher number of synonymous substitutions than the rest of essential genes, although these differences were not statistically significant. These results points out to a relaxation in the selective pressures over this conditionally essential genes in a nutrient-rich environment in

comparison with their more essential role, and as consequence sequence conservation, in *E. coli* K12 lineage. A similar situation explains the evolutionary patterns of cofactor biosynthesis genes that appears as absent in all minimal networks under all conditions, where patterns of sequence evolution at both synonymous and non-synonymous sites appears more closer to the rest of non-essential genes both in pairwise estimates and in lineage-specific estimates of *dN* and *dS*. It must be also considered that estimates of *dN* and *dS* for *S. glossinidius* lineage are calculated for the whole evolutionary branch after its divergence from a free-living ancestor common to *E. coli* K12. This evolutionary branch must be divided in the initial evolution of *S. glossinidius* as a free-living bacteria until its transition to the host-dependent lifestyle, where evolutionary rates would be similar to that of a free-living bacterium like *E. coli* K12, and a second evolutionary period corresponding to its evolution as a bacterial endosymbiont, where acceleration in evolutionary rates must take place, so acceleration in rates of sequence evolution observed for the whole evolutionary branch of *S. glossinidius* since their divergence from a free-living ancestor would be higher during the last evolutionary period as a bacterial endosymbiont.

5.5 CONCLUSIONS

Metabolic network analysis represents a novel approach to study the functionality of biological systems that complements functional inferences based on genome annotation, allowing a more comprehensive approach to the overall functionality of a particular organism The reconstruction and functional analysis of the metabolic network of *S. glossinidius* at different stages of the genome reduction process based on the complete set of genes and pseudogenes present in this genome has confirmed the recent transition to a host-dependent lifestyle experienced by this bacterium, revealed by the capability of its ancestral metabolic network to produce a viable phenotype in terms of biomass production under a minimal environment with only glucose as external carbon source, and the important role that the inactivation of a few enzymatic activities have over the functionality of the whole metabolic system. In addition, network analysis confirms the inability of *S. glossinidius* functional network to synthesize arginine and not alanine, as was proposed in the previous chapter of this thesis. In addition, a gradual decrease in the network robustness to gene deletion events and to changes in enzymatic activities has been also reported, that can be associated with the strict host-dependent lifestyle of these bacterial endosymbionts of insects, that in *S. glossinidius* are in the very initial stages in comparison with ancient associations like that of *Buchnera aphidicola* (Thomas et al., 2009). Minimal networks obtained from *S. glossinidius* functional network through reductive evolution simulations reveal that the gene content of minimal network is strongly dependent on the environment under which the system is evolving, and that under external conditions that simulate the tsetse internal environment it is possible the evolution of the metabolic system towards outcomes

rather similar to a real minimal genome like that of *W. glossinidia*, with increasing accuracy that would be possible with precise adjustment of the optimization function to the functional capabilities of *W. glossinidia*.

Finally, computational simulations of reductive evolution over *S. glossinidius* functional network have also revealed a significant correlation between the essential character of the genes in the context of system viability and their patterns of sequence evolution. Thus, essential genes in all minimal networks show more restricted patterns of sequence evolution and more optimized codon usage that genes whose removal has no effect over the functionality of the system, in concordance with similar studies carried out with experimentally defined essential and non-essential genes (Jordan et al., 2002). The fact that these genes with more relaxed patterns of sequence evolution are predominantly lost in ancient bacterial endosymbionts (Delmotte et al., 2006) points out to the capability of network analysis to predict gene content evolution. However, differences in gene content between minimal networks defined under different external conditions indicate a significant change in the essential character of genes after the transition to a host-dependent lifestyle, with the loss of essentiality in a nutrient-rich environment that is associated with a more relaxed patterns of sequence evolution in comparison with free-living bacteria as evidenced with the analysis of the set of conditionally essential genes under nutrient-limited environment. In addition, lineage-specific analysis of the patterns of sequence evolution shows that the acceleration in substitution rates that has affected to long-term bacterial endosymbionts is also affecting to *S. glossinidius* at both non-synonymous and specially at synonymous sites. This can be affecting the level of adaptative codon usage in comparison with free-living bacteria, revealed by the lack of significant correlation between the number of synonymous substitutions and the codon adaptation index for *S. glossinidius* essential and disposable genes.

Overall, these results have allowed study the complete reductive evolution process experienced by *S. glossinidius*, and have revealed that although patterns of sequence evolution in essential and non-essential genes are similar to those in free-living bacteria, common evolutionary features like the acceleration in the rates of sequence evolution and the loss of adaptative codon usage are starting to affect *S. glossinidius*, leading to a point of no return in the context of the reductive evolution process that is also correlated with the decrease of system robustness to external perturbations in terms of gene content and enzymatic activities.

6 General discussion

General discussion

The availability of hundreds of completely sequenced genomes from different bacterial lineages at different levels of evolutionary relationships has supposed a revolution in the study of the dynamics of bacterial evolution.. Genomic data have allowed study adaptative evolution of bacterial lineages from a systemic and a genomic perspective, integrating how point mutations and changes in the content and organization of bacterial genomes affect the overall structure and functionality of the whole cellular system. Comparative genomic analyses in different bacterial lineages with different degrees of evolutionary proximity has allowed exploring genome dynamics under different ecological contexts, providing novel insights on the basis of prokaryotic evolution. Whereas the study of individual bacterial genomes that are being progressively sequenced is still revealing the enormous diversity in terms of functional capabilities and gene content existing within the prokaryotic kingdom, the availability of multiple genomes from closely related bacteria, from different strains of the same species to different lineages of a particular bacterial genus, has allowed to study the mechanisms of genome evolution and adaptation in different bacterial lineages, and have revealed that the diversity in terms of gene content and genome organization, even at very short evolutionary distances, is surprising large (Koonin and Wolf, 2008). As described in the introduction of this thesis, the gene repertoire of bacterial genomes changes quickly as a consequence of horizontal gene transfer, gene duplication and gene loss events, allowing rapid responses to environmental fluctuations, but the relative importance of these processes varies between bacterial lineages based on their ecological context and evolutionary pressures. The largest bacterial genomes correspond to bacteria that live in complex and changing environments like soil or rhizosphere. Under these changing conditions, many genes have been acquired and maintained in the genomes despite being only occasionally advantageous (Bentley and Parkhill, 2004). In contrast, the highly reduced genomes of obligatory intracellular pathogenic and symbiotic bacteria are consequence of a massive process of gene loss associated to the transition to a more stable ecological niche corresponding to the intracellular environment of their eukaryotic hosts, that generates minimal genomes with highly streamlined functional capabilities that are strongly dependent of the type of association established with the host (Silva and Latorre, 2008). In both processes, mobile genetic elements proliferation plays a decisive role both in terms of gene content evolution by allowing the acquisition of novel metabolic capabilities by lateral gene transfer of transposons and prophage elements (Ohnishi et al., 2001; Thomson et al., 2004; Brussow et al., 2004; Canchaya et al., 2004) and as a consequence of their impact over the structure and organization of the genome as causative agents of genome rearrangements and gene inactivation events by their repeated insertion in different genome locations (Moran and Plague, 2004; Siguier et al., 2006a; Galas and Chandler, 2009).

Variability in the dynamics of genome evolution is present not only between bacterial lineages, but also within a particular evolutionary lineage at different stages

of its evolution. This is the case of the different stages within the reductive evolution process associated with intracellular symbiotic and pathogenic bacteria since their transition from a free-living ancestor. In this process, initial transitions to the host-dependent lifestyle appears associated with a massive process of gene inactivation and mobile genetic elements proliferation consequence of the relaxation in the selective pressure over a large number of genes that become redundant and non-essential in the more stable environment associated with the eukaryotic host, a strict vertical transmission of these intracellular bacteria from mothers to descendents that prevents the influx of genetic material through lateral gene transfer, and the drastic reduction of effective population sizes due to population bottlenecks consequence of the vertical transmission that allows the fixation of slightly deleterious mutations by genetic drift in a process known as Muller´s ratchet (Moran, 1996; Wernegreen, 2005).The proliferation of mobile elements in this initial transition to strict host dependence is coupled with an increase of genome rearrangements and deletions by recombination events at these repeated sequences. In contrast, advanced stages of the genome reduction process are characterized by highly streamlined genomes with very few if any mobile genetic elements and pseudogenes and, as consequence, a very stable genomic organization consequence of the absence mobile genetic elements and the loss of most genes involved in recombination processes (Wernegreen, 2002). In addition, the loss of DNA repair genes appears also associated with an acceleration of evolutionary rates in comparison with free-living bacteria, with a increasing bias towards A+T content and loss of effective codon usage (Clark et al., 1999; Rispe et al., 2004).

In the present thesis, both comparative genomic approaches and individual genomic analyses have been combined. First, to study the dynamics of gene order evolution in the lineage of γ-proteobacteria, which includes different bacterial endosymbiont lineages at different stages of the genome reduction process, and second to study the reductive evolution process in the particular lineage of *S. glossinidius*, a facultative endosymbiont of tsetse flies in initial stages of the transition to host-dependent lifestyle, allowing to evaluate the relative impact of gene inactivation and mobile genetic element proliferation over the structure and functionality of the whole cellular system. In addition, the whole reductive evolution process has been modeled by metabolic network analysis to evaluate the most relevant features of the transition from free-living to host dependent lifestyle, how the process of gene inactivation affects the robustness of the whole metabolic system to different mutational processes, and how the system could evolve in the context of the reductive evolution process under different environmental conditions. Consequently, this discussion will start with some consideration about the evolution of genome organization in γ-proteobacterial genomes and how variations in the dynamics of genome rearrangements correlates with differential evolutionary pressures acting on different bacterial lineages. Then, the most relevant conclusions about the relative impact of gene inactivation and mobile genetic element

proliferation over the genome structure and function of *S. glossinidius* will be evaluated and finally, the potential of network analysis to trace the complete reductive evolution of this genome will be also revisited.

6.1 Evolution of genome organization in prokaryotic genomes: The special case of γ-proteobacteria

The availability of completely sequenced genomes has allowed the study of different features concerning both chromosome organization and the generation of genomic variability in bacterial populations. The generation of genetic variability, an essential property of living cells in order to ensure their maintenance across evolution depends on mutation rates at sequence level, horizontal gene transfer, duplication and gene loss events affecting the gene content of bacterial genomes, and processes of intra-genomic recombination that affect the overall genome structure and organization much more rapidly than point mutations. However, those mutational events, specially recombination ones, tend to produce rearrangements in bacterial chromosomes that alter chromosome organization, which suppose a trade-off between genome stability in terms of gene organization and the generation of rapid variation through genome rearrangements by homologous recombination at repeated sequences (Rocha 2004). A classic example of chromosomal organization is the organization of functionally related genes in operons conserved across closely related genomes, as well as the conservation of supra-operonic organization of translation related genes in uber-operons despite the presence of genome rearrangements (Lathe, III et al., 2000; Lawrence, 2003). In addition, a preferential distribution of genes in the leading strand has been also detected in bacterial chromosomes. This was initially associated with a more efficient replication and expression of highly-expressed genes, but posterior studies revealed that this preferential gene distribution is governed by gene essentiality, that even constrains the number of possible genome rearrangements (Rocha and Danchin, 2003b; Rocha and Danchin, 2003a). In the opposite side, the presence of repeated elements across the genome in the form of duplicated genes or mobile genetic elements like IS sequences of prophages generates different types of genome rearrangement events by homologous or illegitimate recombination (Rocha, 2003a; Achaz et al., 2003), resulting in a trade-off between the selective advantages of chromosomal stability in terms of gene organization and the benefits of genome plasticity through genome rearrangements whose results will depend of the ecological context of the organism under study. In bacterial species under periodical stresses, like those imposed by host immune system in bacterial pathogens, adaptation strategies associated with rapid changes in genome structure and organization by genome rearrangements through recombination are positively selected, like in facultative pathogens such as *Helicobacter pylori* (Aras et al., 2003) or different *Streptococcus* strains (Nakagawa et al., 2003). However, many of these rearrangements have detrimental effects over organismal fitness by disrupting selective features of chromosomal organization like

gene dosage, chromosome symmetry or gene strand bias, and as a consequence they tend to be eliminated by natural selection. In fact, the most prevalent observed inversions between ribosomal genes or IS elements are symmetrical relative to the origin of replication because minimizes the disruption of chromosome organization (Hughes, 2000; Tillier and Collins, 2000). In this context, rearrangement rates are correlated with the presence of DNA repeats and recombination mechanisms in bacterial chromosomes, with a negative correlation between the number of repeats and the conservation of gene order that has been demonstrated for the γ-proteobacteria lineage (Rocha, 2003b). This is also evident from the results obtained in chapter 3 of this thesis, where strict collinearity is observed between different strains of *B. aphidicola* that are explained by the lack of repetitive sequences and recombination genes in their highly streamlined genome, whereas in other lineages like in *S. glossinidius*, a higher rate of genome rearrangement is observed and can be explained by the presence of repetitive sequences in the form of IS elements and specially prophage elements together with their completely functional genes involved in DNA recombination as is described in chapter 4 of this thesis.

Despite the selective constraints associated with chromosomal organization in bacterial chromosomes, a decrease in the levels of genome stability is observed with increased divergence times between the compared genomes. This analysis has been carried out by comparing measures of gene order conservation based on the fraction of orthologous genes between two genomes that are adjacent in both genomes with sequence divergence at 16S level or divergence between protein coding genes, but have revealed an exponential drop of the relative order conservation of core genes between genome pairs with divergence time (Huynen and Bork, 1998; Tamames, 2001; Rocha, 2006). These results are also observed when the genome rearrangement distances and amino acid substitution distances are compared in chapter 3, with a central tendency corresponding with an exponential increase of genome rearrangement distances (equivalent to a decrease in gene order conservation) with increasing amino-acid substitution distances. However, in addition to this central tendency, there is a marked heterogeneity in the patterns of genome rearrangements at two different levels. First, heterogeneity is observed between different bacterial lineages that shows large variations in genome rearrangement distances with their closer relatives. This is exemplified by the strict collinearity in the genome sequences of different *B. aphidicola* strains, reflecting a genome stasis common to other ancient endosymbiont lineages that can be explained by the lack of genes involved in recombination processes and the absence of repeated sequences across their highly streamlined genomes (Silva et al., 2003), in contrast with the increased rates of genome rearrangements observed in the *Pasteurellaceae* lineage in comparison with the rest of genomes analyzed. This heterogeneity in genome rearrangement distances between lineages is also reflected when the branch lengths in gene order and sequence based phylogenies are compared. For example, the high number of rearrangements observed between

Shigella and *E. coli* strains in comparison with the single inversion observed between the more distant lineages of *E. coli* K12 and *S. thyphimurium* breaks the monophyletic grouping of the *Shigella-E. coli* clade observed in sequence based phylogenies and puts *Shigella* strains as basal group of enterics in gene order phylogenies. This is also the situation observed when the two strains of *Yersinia pestis* are compared, with 14 rearrangements between them despite their very recent divergence (Achtman et al., 1999). Although this can be explained by the high frequency of mobile genetic elements in *Yersinia* and *Shigella* genomes (Parkhill et al., 2001; Deng et al., 2002; Jin et al., 2002; Wei et al., 2003), it must be taken into account that many of the rearrangements observed between closely related genomes can be consequence of transitional rearrangements that will be subsequently purged by selection, leading to a relative excess of genome rearrangements when closely related genomes are compared. In fact, detailed computational analysis of rearrangement scenarios in *Yersinia* lineage have revealed that although *Yersinina* genomes have accumulated many rearrangements, their genomic organization is in general less deleterious than expected by chance, pointing out to the adaptative value of genome organization despite transient increases of genome rearrangements consequence of proliferation of mobile genetic elements across the genome (Darling et al., 2008). Second, heterogeneity within a particular bacterial lineage at different periods of their evolution has been also detected, reflected by the acceleration in the rates of genome rearrangements in endosymbiont lineages in initial stages of the transition to a host-dependent lifestyle followed by a complete genome stability that corresponds with the divergence between different endosymbiont strains. In ancient endosymbiotic lineages like *B. aphidicola* in aphids or *Blochmannia* in carpenter ants, the genome stasis observed in advanced stages of the association can be explained by the loss of repeated sequences and genes involved in recombination during the genome reduction process, despite that their small effective population sizes could lead to increases in genome instability by inefficient selection against chromosomal rearrangements. However, in initial stages of the association, a combination of a functional profile closer to a free-living ancestor, that includes many of the genes involved in recombination, with the relaxation of the selective pressure over large segments of the genome that becomes non-essential in the context of the host association, favors the proliferation of different types of mobile genetic elements and the fixation of multiple genome rearrangements by recombination processes at these repeated sequences that will not be eliminated by natural selection. The most clear example of this situation corresponds to *S. glossinidius*, where the proliferation of mobile genetic elements combined with the presence of functional genes of the recombination machinery explains the acceleration in the rates of genome rearrangements observed both when relative inversion rates are compared since its divergence from free-living enterobacterial ancestors and in the comparison between genome rearrangement distances and amino acid substitution distances, where *S. glossinidius* comparisons appears above the central tendency observed for most genome comparisons. In addition, this

acceleration in rearrangement rates can be also inferred for the lineage of *B. aphidicola* after its divergence from free-living enterobacterial ancestor, where acceleration in rearrangement rates is also observed although to a lesser extent that in *S. glossinidius* because is averaged for the complete evolution since their free-living ancestor, and that coincides with the large number of chromosomal rearrangements and deletions observed in comparison with free-living enteric relatives (Moran and Mira, 2001). Thus, relaxed purifying selection combined with the proliferation of repeated sequences across the genome lead initially to higher rates of genome rearrangement that originates rapid changes in genome structure, whereas in advanced stages of the association, the loss of elements responsible of this changes such as repeated sequences and genes responsible of the recombination machinery combined with the absence of lateral gene transfer due to their strict isolation in the intracellular environment of their eukaryotic host leads to the observed higher genome stability of the ancient endosymbionts.

Phylogenetic reconstructions obtained from genome rearrangement data will be affected by the heterogeneity in the genome rearrangement rates in a similar manner as variations in substitution rates affect phylogenetic reconstructions based on sequence data. This constitutes one of the main limitations of phylogenetic analysis based on whole-genome features, both gene order and gene content, in comparison with the more developed methodological background associated with phylogenetic reconstruction based on sequence data. An argument in favor of phylogenetic reconstruction methods based on whole genome features is that changes in gene content and gene order within genomes results in characters with billions of possible states in comparison with the only four possible states in phylogenetic reconstruction methods based on nucleotide sequence data, so in principle are less prone to homoplasy by convergence or reversal and as consequence might represent good phylogenetic markers (Gribaldo and Philippe, 2002). However, a major drawback of gene content and gene order phylogenies is the lack of evolutionary models capable to deal with the different types of genome rearrangement events affecting genome evolution and with the heterogeneity in genome rearrangement rates present among bacterial lineages, that combined with the rapid loss of gene content and gene order similarity observed between bacterial lineages in comparison with sequence similarity precludes their utilization as reliable phylogenetic marker to closer related species (Snel et al., 2005; Delsuc et al., 2005). Taking into account this consideration, comparative analysis of phylogenetic reconstructions based on sequence and gene order data provides insights into the differential evolution of both characters in different species. This is the case of the basal location of the *Pasteurellaceae* lineage in gene order phylogenies in comparison with their more central location as outgroup of enteric bacteria in sequence based phylogenies observed in chapter 3, that can be explained by the observed acceleration in the rates of genome rearrangements in this particular lineage in comparison with the rest of γ-proteobacterial genomes analyzed. In a similar manner, the acceleration in the rates

of genome rearrangements in initial stages of the transition to a host dependent lifestyle in bacterial endosymbionts of insects explains the paraphyletic grouping of ancient endosymbionts in gene order phylogenies, as well as the long branches of these endosymbiotic lineages in comparison with sequence based phylogenies, where ancient endosymbionts appear grouped in a monophyletic group. This contrast between acceleration in rearrangement rates in comparison with sequence evolution is especially evident for *S. glossinidius*, where an extremely large branch in gene order phylogeny is observed in comparison with the sequence based phylogeny.

6.2 Dynamics of genome evolution in bacterial endosymbionts of insects: The special case of *Sodalis glossinidius*

The particular situation of *S. glossinidius* within the bacterial endosymbionts of insects, with the accelerated rates of genome rearrangements observed in the results of chapter 3, leads us to study more deeply the structure and functional profile of this bacterial endosymbiont that is one of the few complete genomes of a bacterium at the initial stages of the adaptation to the host-dependent lifestyle. A common feature of bacterial species recently evolved towards a host-dependent lifestyle, either pathogenic or symbiotic, is the proliferation of insertion sequences in comparison with closer relatives that retain the free-living stage. This is the case of pathogenic lineages of enterobacteria like different strains of *Shigella flexneri*, which despite their very recent divergence from *E. coli* strains (35,000-270,000 years ago), have experienced a massive proliferation of IS elements that have lead to rapid loss of gene order conservation in comparison with non-pathogenic *E. coli* strains consequence of recombination between copies of IS elements (Jin et al., 2002; Wei et al., 2003). Another example is the proliferation of IS elements in the lineage of *Yersinia pestis* CO92, where genome rearrangements by recombination over IS elements are commonly observed even in *in-vitro* cell cultures, leading to a highly fluid genome organization that has been associated with their rapid evolution from a clone of *Y. pseudotuberculosis* widely found in the environment to a blood-borne pathogenic form with limited capability of survival outside their hosts (Parkhill et al., 2001). Although in both examples (*Shigella* and *Yersina*) horizontal gene transfer also plays a decisive role in the acquisition of virulence determinants responsible of the pathogenic lifestyle of each bacterium, IS proliferation appears associated with a restriction in the range of hosts that the bacterium is able to colonize. In this context, the characterization of the total number of IS elements in the genome sequence of *S. glossinidius* carried out in chapter 4 of this thesis yields a total number of 130 IS elements scattered across the genome that supposes an IS load similar to other recent host-dependent bacteria without free-living stage like *Yersinia pestis* or the secondary endosymbiont of aphids *Hamiltonella defensa*, and that contrasts with their absence in ancient host-associated lineages like that of most primary endosymbionts of aphids with the exception of parasitic *Wolbachia*

pipientis wMel (See Figure 6.1). In the case of *Wolbachia pipientis wMel*, despite having genome features typical of ancient host-restricted bacteria (small genome size, rapid sequence evolution, A+T bias), have a genome with a large fraction of mobile genetic elements that has been associated with the capability of different *Wolbachia* strains to co-infect a particular insect host, where they can undergo recombination and potentially expansion of mobile genetic elements between different strains (Werren and Bartos, 2001; Jiggins, 2002; Wu et al., 2004). However, the case of *Wolbachia* does not represent the most extreme case of IS proliferation, that corresponds to the primary endosymbionts of rice and maize weevils *Sitophilus oryzae* (SOPE) and *Sitophilus zeamays* (SZPE), where quantitative PCR assays yields estimates of IS loads of 5000 copies for a particular IS element (Plague et al., 2008). However, this is probably an overestimation because it would imply that IS elements represents a minimum of 5 megabases in SOPE and SZPE genomes, although the fact that IS elements represents at least 20% of SOPE genome sequence indicates that it could be more than 1000 IS copies in the genome (Gil et al., 2008). SOPE and SZPE represent the closest bacterial relatives to *S. glossinidius* in phylogenetic reconstructions, with a divergence time between both endosymbiotic lineages roughly estimated in 100 Myr based on similar values of non-synonymous substitutions in protein coding genes between *S. glossinidius*-SZPE and *Salmonella enterica-E. coli* (Charles et al., 1997; Dale et al., 2002), and can be considered as a further step in the integration of a bacterial lineage into the host-dependent lifestyle.

Whereas *S. glossinidius* is a facultative secondary endosymbiont with wider tissue tropism that can reside both intra- and extracellularly in different host tissues, SOPE and SZPE are obligatory mutualistic primary endosymbionts that resides exclusively in bacteriocyte cells of the rice and maize weevils respectively (*Coleoptera, Curculionidae*) supplying the weevil with several vitamins (pantothenic acid, biotin and riboflavin) that increase mitochondrial enzymatic activity, host fertility and fly ability of weevil host (Wicker, 1983; Heddi et al., 1993; Grenier et al., 1994). However, both SOPE and SZPE have established very recent association with their weevil host, a feature reflected by the lack of compositional bias towards A+T (Heddi et al., 1998), their large genome size, estimated around 3 megabases (Charles et al., 1997), and the differential expression of Type III secretion system genes closely related to those found in *S. glossinidius* during weevil metamorphosis associated with the transmission to bacteriome cells (Dale et al., 2002), similar to the behavior of *S. glossinidius*, where differential expression of genes of two of the three type III secretion systems present in the genome has been detected associated with host cell invasion and intracellular proliferation (Dale et al., 2005).

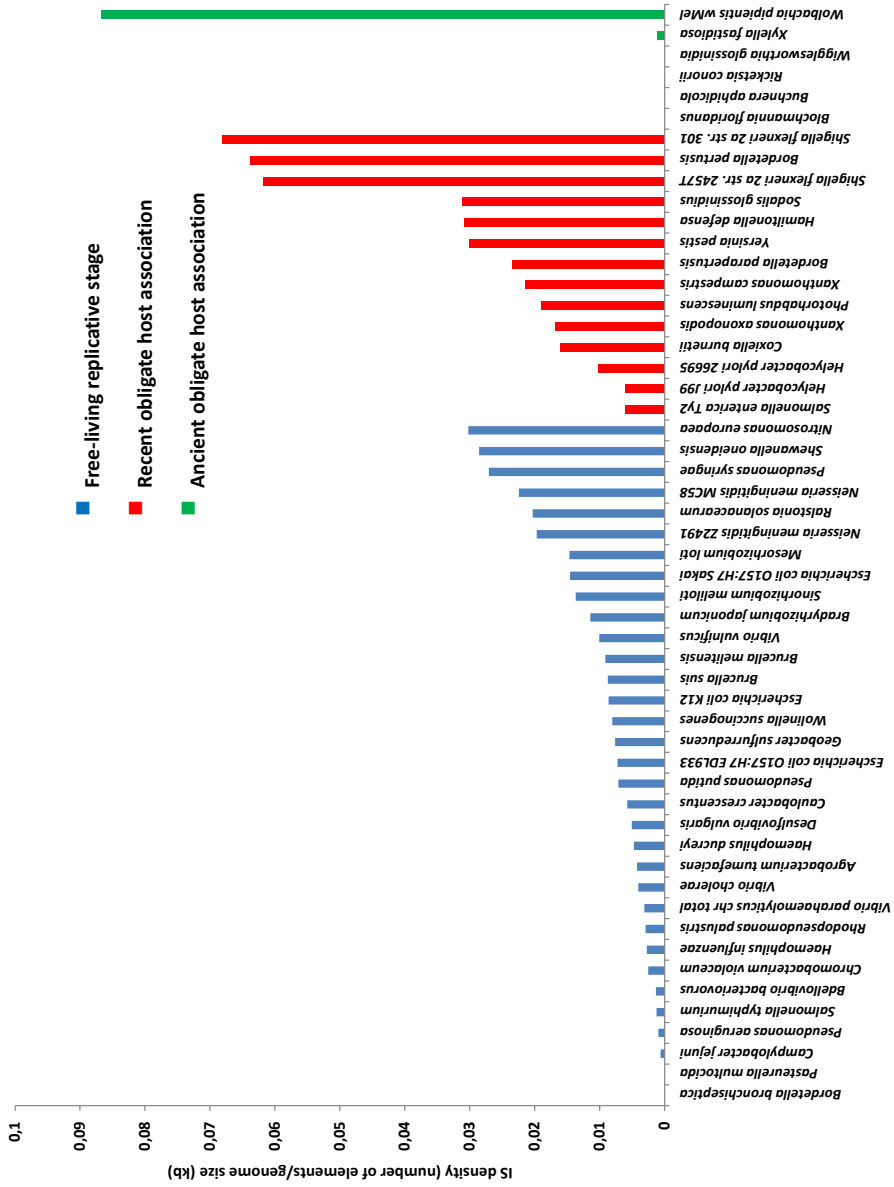


Figure 6.1: *Density of insertion sequences (number of IS elements per kilobase) in different proteobacterial genomes. The species are divided in three main groups based on their lifestyle and the age of the clade as specified in the corresponding genome paper. Adapted from (Moran and Plague, 2004)*

The differences in the nature of interactions between *S. glossinidius* and weevil endosymbionts could be responsible of the massive proliferation of IS elements in both SOPE and SZPE in comparison with *S. glossinidius*, in the sense that the more stable intracellular environment associated with bacteriocyte cells will produce a further decrease in the selective pressures for gene function maintenance in weevil endosymbionts that will favor the massive proliferation of IS elements that will not be counterbalanced by selection. In fact, IS transposition has been characterized even within 23S genes in SZPE lineage (Dale et al., 2003). Comparison of the different families of IS elements characterized in *S. glossinidius* in chapter 4 with IS elements from SOPE and SZPE reveals also a common origin of ISSgl1 from *S. glossinidius* and ISsope*1* of SOPE and SZPE, with no similarity for the rest of IS elements characterized for *S. glossinidius* and the rest of IS elements characterized in SOPE and SZPE (Plague et al., 2008; Gil et al., 2008), representing lineage-specific acquisitions during the evolution of *S. glossinidius* (ISSgl2-5 and single copy IS elements) and weevil endosymbionts.

Another relevant characteristic of *S. glossinidius* genome is the presence of a high number of genes and pseudogenes of phage origin (698 CDSs), which represent the 17.8% of the total number of CDS of the genome. Bacteriophages represent the most abundant organisms on earth, with around ten phages for every microbial cell in almost all investigated ecosystems , and they have been postulated as major contributors to the diversity of prokaryotic populations (Rodriguez-Valera et al., 2009). Comparative genomics has revealed that they are responsible of the majority of strain-specific DNA in closely related genomes, like different strains of enterohaemorrhagic *E. coli*, where prophages are responsible of the differences in gene content with non-enterohaemorrhagic *E. coli* strains through independent acquisition of different prophage elements in each enterohaemorrhagic lineage, which confer their pathogenic character (Ogura et al., 2009). In a similar manner as happens with IS elements, prophage elements are nearly absent in the genomes of ancient host associated endosymbionts due to their long term degenerative evolution since their initial transition to a host-dependent lifestyle, but the genome sequences of more recently host-associated bacteria like *Wolbachia pipientis Wmel* or facultative endosymbionts of aphids *Hamiltonella defensa* and *Regiella insecticola* are known to contain complete prophage genomes and substantial numbers of prophage remnants that harbor toxin-coding genes with important activities in the context of the interaction with the eukaryotic host (Wu et al., 2004; Moran et al., 2005a; Degnan et al., 2009b; Degnan et al., 2009a). In addition, bacteriophages have been postulated as the source of most of strain-specific DNA without significant sequence similarity with other genomes (ORFan genes) based on their common compositional features in terms of increased AT content and similar codon biases and their significant clustering in genome regions adjacent to well known phage integration sites like tRNA genes, pointing out to a large hidden (unsequenced) reservoir of integrative elements like prophages or plasmid genes that is responsible

of the variable component of a particular genome with respect to their close relatives (Daubin and Ochman, 2004; Lerat et al., 2005; Cortez et al., 2009). Similar conclusions can be inferred from the distribution of phage genes and pseudogenes in the genome of *S. glossinidius*, where most of prophage CDSs appears clustered across the genome associated with genome regions with differential compositional features in terms of oligonucleotide composition and lower %GC than the rest of the genome, and that appears close to phage integration sites like tRNA genes (See Figure 6.2). However, prophages appear also as the functional class most affected by pseudogenization with 353 pseudogenes (50.6% of phage CDS), pointing out to the inactive profile of most of these prophage elements.

Parallel to this increase in IS and prophage elements, an increase in the number of pseudogenes is also observed in this bacterial lineage with recent transition to host-dependent lifestyle. The identification of pseudogenes in prokaryotic genomes has been favored by the availability of large number of complete bacterial genomes for comparative analysis. Previously to the massive availability of complete genome sequences, it was thought that bacteria would contain few pseudogenes because their genome sequences are small, with high coding density and very little non-coding DNA in comparison with eukaryotic genomes (Lawrence et al., 2001). However, with the availability of complete genome sequences it has been revealed that almost all bacterial genomes contains variable numbers of pseudogenes, and that the number of pseudogenes is specially higher in the genomes of pathogens and symbionts with host-dependent lifestyle. The most relevant example is the massive proliferation of pseudogenes in the obligatory intracellular pathogen *Mycobacterium leprae*, with more than 1116 pseudogenes originally annotated in their genome sequence (Cole et al., 2001). This is also the case of different sequenced genomes of *Y. pestis*, that constitutes a recent pathogen of mammals that is estimated that has evolved from a clone of the environmental enteropathogen *Y. pseudouberculosis* less than 20,000 years ago (Achtman et al., 1999) and where hundreds of pseudogenes have been identified in original annotations of completely sequenced genomes of different *Y. pestis* strains (Parkhill et al., 2001; Deng et al., 2002; Chain et al., 2006), or the even higher number of pseudogenes originally annotated in different strains of the human pathogen *Shigella flexneri* since their divergence from *E. coli* lineage (Jin et al., 2002; Wei et al., 2003).

Similarly, human-restricted *S. enterica* serovars *Thypi* and *Paratyphi* have accumulated a large number of pseudogenes in comparison with host-generic relatives such as *S. enterica* serovar *Typhimurium*, being able to trace the process of gene inactivation at different levels of the evolution of this human-restricted serovars (Holt et al., 2009). Because the mutational process in bacteria is biased toward deletions, pseudogenes ultimately diverge and erode beyond recognition and are eventually eliminated from the genome (Mira et al., 2001)

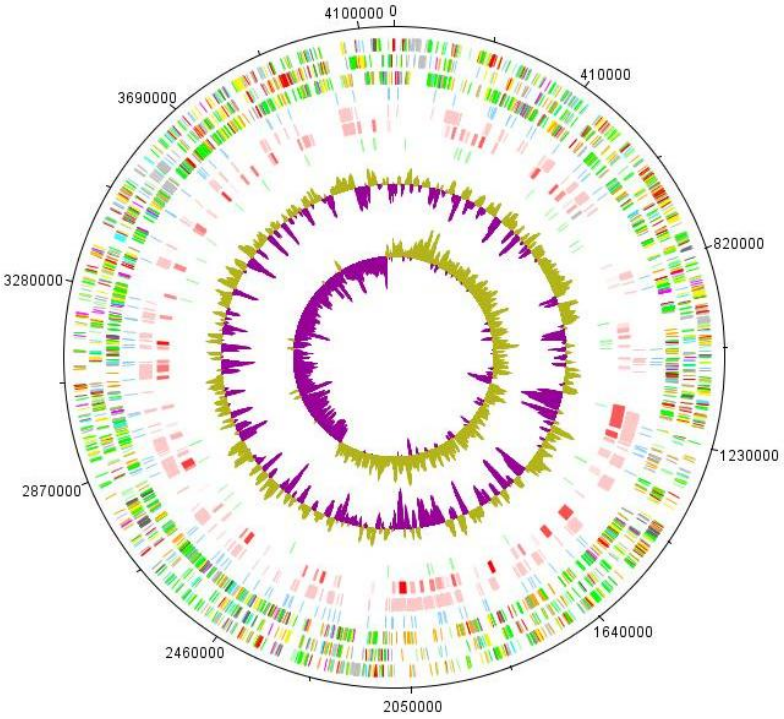


Figure 6.2: *S. glossinidius circular genome representation with the results of the re-annotation carried out in chapter 4 of this thesis. The graph is generated with DNAplotter software (Carver et al., 2009). The tracks, from the outside to the inside, represent: (1) Forward genes (without phage genes); (2) Reverse genes (without phage genes); (3) Pseudogenes (without phage pseudogenes); (4) IS elements; (5) Phage CDS (genes and pseudogenes); (6) Genome regions with deviant compositional features respect the average genome in terms of K-mers of different size predicted by Alien Hunter software (Vernikos and Parkhill, 2006). 92 putative horizontally transferred regions were predicted by this method, with color intensity reflecting the score of each predicted region; (7) tRNA genes; (8) % GC plot; (9) GC skew.*

However, although pseudogene proliferation is predominant in initial stages of the transition to host-dependent lifestyle, traces of gene inactivation processes are also evident in more ancient intracellular bacteria, like in the intracellular parasite *Rickettsia prowazekii*, where 12 pseudogenes were originally annotated together with a large fraction of non-coding DNA (only 76% of their genome represented by functional genes) that represents remnants of ancient genes in the process of being eliminated from the genome (Andersson et al., 1998). Posterior comparative analysis of this pseudogene sequences in additional *Ricketsia* genomes have revealed

different stages of gene disintegration in different lineages, with a progressive process of gene disintegration since the initial inactivating mutation characterized by increases in nucleotide substitution rates that leads to rapid loss of sequence similarity with functional counterparts in other genomes, together with deletional events predominantly of small size that leads to the elimination of non-functional DNA from the genome sequence (Andersson and Andersson, 1999a; Andersson and Andersson, 2001). Similar dynamics of pseudogene evolution were characterized during reductive evolution of different *B. aphidicola* lineages, where analysis of gene losses over different branches of the phylogeny of *BAp*, *BSg* and *BBp* leads to estimates of average half-life of pseudogenes of 23.9 Myr, with most of pseudogenes inactivated in initial stages of the association that has been completely lost in actual genomes but with recent inactivation events still identifiable by sequence similarity searches with functional counterparts (Gomez-Valero et al., 2004a). The gradual loss of sequence similarity with functional counterparts along the evolution of pseudogenes since the initial inactivation event represents also one of the main difficulties in their identification, and makes that pseudogenes annotated in original genome papers were restricted to recent inactivation events where the ancestral gene retains enough sequence similarity to be predicted by *ab-initio* gene prediction methods. In fact, systematic comparisons of genome sequences with functional genes of closer relative genomes by means of BLAST yields large number of pseudogenes not identified in the original annotations (Homma et al., 2002; Lerat and Ochman, 2004; Liu et al., 2004; Lerat and Ochman, 2005). In the original annotation of *S. glossinidius*, 972 pseudogenes were described although they remain absent in all sequence files available in public databases, which leads us to carry out a characterization and functional annotation of 1501 pseudogenes by surveying *S. glossinidius* intergenic regions in order to fulfill the objectives of chapter 4 and specially chapter 5 of this thesis, which supposes a significant increase in comparison with original annotation. Most of these pseudogenes are consequence of point mutations in the form of single base substitutions that lead to premature stop codons or single base insertion or deletion that lead to frameshift mutations, with the influence of IS transposition on the gene inactivation process being minimal (only 18 pseudogenes originated by IS insertion). This has been also observed for pseudogenes characterized by similar comparative genomic approach in different bacterial lineages, where most of pseudogenes are originated by frameshifts or premature stop codons (Lerat and Ochman, 2004; Lerat et al., 2005). Only in *S. flexneri 2a str. 301* a significant number of pseudogenes are originated by insertion of IS elements (79 out of 422 pseudogenes), associated with the large number of IS elements identified in this genome (314 complete or partial IS elements), although most of the pseudogenes (155 out of 422) are originated by premature stop codons (Jin et al., 2002; Lerat and Ochman, 2004).

Similar dynamics of gene inactivation were described also for *Mycobacterium leprae*, where the reconstruction of the ancestral genome of *M. leprae* and *M.*

tuberculosis reveals that most gene losses in the lineage of *M. leprae* are consequence of single point mutations yielding pseudogenes still identifiable in *M. leprae* genome sequence, consequence of a massive event of gene inactivation 20 m.y.a based on the inferred age of the pseudogenes (Gomez-Valero et al., 2007). The large number of pseudogenes identified in *S. glossinidius*, most of them with high sequence similarity with closer relatives, points to similar massive gene inactivation process in the very recent evolutionary history of *S. glossinidius*, although different stages of gene inactivation can be also detected. For example, very recent inactivation events are exemplified by the 142 originally annotated genes shorter than their corresponding homologs in sequence similarity searches and for which adjacent pseudogenes are characterized during the re-annotation process, like the genes *galK* and *galT* described in chapter 4 involved in UDP-glucose and UDP-galactose biosynthesis within the pathway for galactose catabolism, where single frameshift mutation (in *galK*) and premature stop codon (in *galT*) disrupts the ancestral gene, although a complete open reading frame is still recognizable by gene prediction methods. More advanced stages of gene inactivation are exemplified with *araB* pseudogene, that encodes an L-ribulokinase involved in arabinose catabolism (ps_SGL0828c), where multiple frameshift mutations are observed (See Figure 6.2).

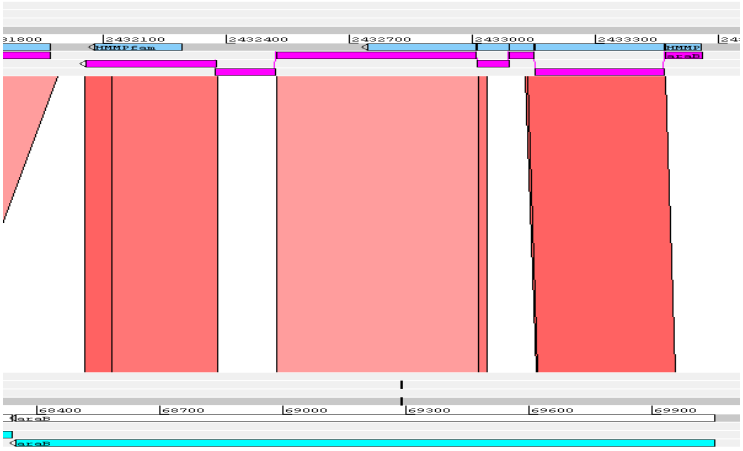


Figure 6.2: *ACT comparison between S. glossinidius (top) and E. coli K12 (bottom) for araB pseudogene.*

Finally, the complete coding density of *S. glossinidius* genome including both genes and pseudogenes is 75.53 %, similar to that reported for the obligatory pathogen *Rickettsia prowazekii* (76 %), although lower than gene density observed for free-living bacteria like *E. coli* K12 (85 %). This high fraction of non-coding DNA in the genome of *Rickettsia prowazekii* has been associated with remnants of ancient genes that are in the process of being eliminated from the genome but for which no sequence similarity is conserved (Andersson et al., 1998). In *S.*

glossinidius, the high fraction of non-coding DNA could be remnants of ancient pseudogenes present even before the transition to host dependent lifestyle for which no significant sequence similarity is conserved with functional counterparts.

However, is important to consider that the situation observed in *S. glossinidius* represents only a snapshot in the evolution of this particular linage, and the absence of genome sequences of other strains from different tsetse host difficult to understand the dynamics of pseudogene evolution in a similar manner as has been possible for other intracellular bacterial lineages for which multiple genomes are available like for *Mycobacterium* or *Ricketsia* lineages (Gomez-Valero et al., 2007; Fuxelius et al., 2008). There is no overlap in the set of pseudogenes harbored by closely related genomes like *E. coli*, *Yersinia*, or *Shigella* (Lerat and Ochman, 2004; Lerat and Ochman, 2005), and similar variable pool of pseudogenes is observed across different *Ricketssia* lineages (Fuxelius et al., 2008), indicating different profiles of recent gene inactivation events in closer relative lineages. Similarly, IS elements have experienced frequent episodes of expansion and extinction that leads to variations in IS loads between closely related bacterial lineages (Wagner, 2006), and large variations in IS density from particular families has been also observed even in SZPE from different *Sitophilus zeamays* populations (Dougherty and Plague, 2008).In the case of *S. glossinidius*, the availability of additional genomes from different tsetse species or even for the same tsetse host (*Glossina morsitans morsitans*) would allow to determine common trends in the process of gene inactivation between lineages and to study the dynamics of prophage and IS element proliferation in order to better understand the long term evolution of this particular bacteria.

6.3 Tracing the reductive evolution process of *S. glossinidius* through metabolic network analysis

One of the objectives of this thesis is the study of the effects of pseudogenization over the functional capabilities of *S. glossinidius* based on the complete functional re-annotation of all genes and newly characterized pseudogenes carried out in chapter 4. This objective can be addressed by using information of gene annotation and different bioinformatic tools that produce a more or less descriptive picture of the global metabolic map of a particular organism. This is what is carried out in chapter 4 of this thesis, where the combined analysis of information from gene and pseudogene annotation with the results of KAAS and Blast2GO programs, that maps gene and pseudogene functions on well-characterized biochemical pathways based on sequence similarity searches, allows to describe the most relevant aspects of cellular metabolism of *S. glossinidius*, how pseudogenization has affected to particular gene functions, and how these pseudogenization events affect the functionality of particular metabolic pathways. This analysis have revealed a functional profile of *S. glossinidius* as an aerobic heterotroph bacterium highly

similar to most free-living enterics, with complete functionality of all pathways of the central metabolism (Glycolysis, TCA cycle, Pentose Phosphate pathway and oxidative phosphorylation), being able to produce energy from different carbon sources although pseudogenization has affected some components of PTS systems. Based on its gene complement, *S. glossinidius* is able to synthesize all essential amino acids except L-arginine and all non-essential amino acids, together with most essential vitamins except thiamine, in which a possible case of metabolic complementation is reported between *S. glossinidius* and primary tsetse endosymbiont *W. glossinidia*, because none of both endosymbionts is capable to synthesize de novo the functional coenzyme thiamine diphosphate. The analysis of the functional profile of *W. glossinidia* also reveals their inability for de-novo biosynthesis of coenzyme A and folate, both completely functional in *S. glossinidius*, pointing out to another possible metabolic complementation between both endosymbionts but this time with *S. glossinidius* as the possible source of β-alanine and ketoisovalerate precursors for coenzyme A biosynthesis and p-aminobenzoate for folate biosynthesis in *W. glossinidia*. In addition, *S. glossinidius* retains all functional genes for the biosynthesis of lipids, cell envelop structures, as well as completely functional machinery for DNA recombination and repair. This constitutes the traditional reductionist approach to study biological functions based on the characterization of individual components, that in the case of genomic analysis correspond to the identification and functional annotation of the complete set of genes and pseudogenes present in a particular genome, and based on this information the functionality of the whole cellular system is described, as is carried out in the majority of genome papers that are currently published.

The genomic revolution experienced in the last decade together with the development of high-throughput experimental technologies (proteomics, metabolomics, and transcriptomics) that allow obtaining massive information about the structure, function, and expression of almost all components of cellular systems has provided an integration of traditional molecular biology into the realm of systems biology. Systems biology is the integration of experimental and computational approaches to achieve the overall objective of explaining and predicting complex cellular behaviors of biological systems (Borodina and Nielsen, 2005). Its objective is to study the interactions between the system components rather than the individual properties of the individual components themselves, and this is achieved through different network reconstructions that represent such interactions between cellular components. One of the most revealing examples of the potential of integrative systems biology in the understanding of the overall cellular responses of a particular biological system is represented in three recently published papers where the small genome of the human pathogen *Mycoplasma pneumoniae* is used as model system (Guell et al., 2009; Yus et al., 2009; Kuhner et al., 2009). *M. pneumoniae* have a small genome of only 816 Kb encoding proportionately fewer genes and regulatory elements that other common bacterial

model systems like *E. coli*, but has the particularity that can be cultivated in vitro, allowing to experimentally determine parameters on its physiology (Himmelreich et al., 1996; Dandekar et al., 2000). This is used to reconstruct the metabolic network of *M. pneumoniae* by integrating biochemical information from their genome sequence with experimental measures of enzymatic activities and biomass constituents, and a minimal medium with 19 essential nutrients is designed and experimentally tested through more than 1300 growth curves that allow the delimitation of the maximum uptake rates for different nutrients (Yus et al., 2009). Metabolic network analysis is complemented with an analysis of the patterns of gene expression of the whole genome through transcriptome sequencing (Guell et al., 2009) and genome-scale analysis of protein complexes through tandem affinity purification-mass spectrometry (Kuhner et al., 2009), revealing a metabolic profile less redundant in terms of enzyme paralogy than larger bacteria although with a higher fraction of multifunctional enzymes and an unexpected complex transcriptional landscape with coordinated changes in gene expression in response to external perturbations similar to more complex bacteria and with large number of antisense transcripts with regulatory functions in a similar manner as happens in eukaryotes.

In the context of genomic analysis, the complete genome sequence of a particular organism can be considered as the system to study, and metabolic network reconstructions allow the integration of all particular gene functions related with metabolism in a mathematical model that has the ability to simulate the growth of living cells through Flux Balance Analysis (FBA) maximizing biomass production as objective function (Covert et al., 2001; Price et al., 2003). This approach has been carried out in chapter 5 of this thesis in order to evaluate the complete functional profile of *S. glossinidius* at different stages of the genome reduction process under different external conditions. The results reveal the complete functionality of the ancestral metabolic network comprised by genes, pseudogenes and deleted genes in a minimal aerobic medium with only glucose as external carbon source at similar biomass production rates as model free-living bacteria like *E. coli* K12, in concordance with their very recent transition to a host dependent lifestyle, and the drastic consequences that a few inactivation events, in this case the inactivation of the pathways for glycogen and arginine biosynthesis and specially the inactivation of the gene *ppc* encoding the anaplerotic phosphoenolpyruvate carboxykinase, have over the functionality of the metabolic system under minimal conditions and that may have marked the transition to a host-dependent lifestyle. This evolutionary transitions, not only from free-living to host-dependent lifestyle but also between pathogenic and symbiotic bacteria, have been studied in many other bacterial lineages, and have been demonstrated that gene acquisition and deletion are the major events underlying the emergence and evolution of symbiotic or pathogenic bacteria from very closely related bacterial lineages (Ochman and Moran, 2001; Tamas et al., 2001). This is the case of a recent work where experimental evolution

of a pathogenic *Ralstonia solanacearum* carrying a symbiotic plasmid of the Rhizobium *Cupriavidus taiwanensis* yields a mutualistic symbiont capable of fixing atmospheric nitrogen that loss pathogenic character by inactivation of two genes involved in virulence (Marchetti et al., 2010), or the recent identification of a obligatory mutualistic *Wolbachia* in the bedbug *Cimex lectularius* that, in contrast with the parasitic association and wider tissue tropism of *Wolbachia* in most infected insects, have a strict intracellular location in bacteriocyte cells and have an essential role as nutrient mutualist that putatively supplies B-vitamins to the bedbugs (Hosokawa et al., 2010). In *S. glossinidius*, the narrower limits between pathogenicity and symbiosis are reflected by the dependence of type III secretion systems for the invasion and proliferation of tsetse host cells (Dale et al., 2005) and by the pathogenic character that *S. glossinidius* acquires when is transformed with *E. coli ompA* gene encoding outer membrane protein A that is consequence of variations in the exposed residues of OmpA protein between *S. glossinidius* and *E. coli* orthologs (Weiss et al., 2008). In the context of the metabolic network analysis carried out in chapter 5 of this thesis, the above mentioned inactivation events make essential an external supply of arginine for the functionality of *S. glossinidius* functional metabolic network together with the removal of glycogen from the biomass equation, pointing out to the role that these inactivation events (arginine biosynthesis, *ppc* gene and glycogen biosynthesis) may have had in the transition to a host dependent lifestyle. In addition, alanine biosynthesis is completely viable in functional metabolic network without any external supply in contrast with the results of *S. glossinidius* genome paper (Toh et al., 2006), validating the conclusions of chapter 4 based on the results of the functional re-annotation of the whole genome. In addition, the transition to host-dependent lifestyle appears also associated with a significant decrease in network robustness both to gene deletion events and to changes in enzymatic activities when single gene deletions and flux variability analysis is carried out in *E. coli* K12 JR904 and *S. glossinidius* metabolic networks. In the case of gene deletion events, robustness is further reduced in minimal metabolic networks of ancient bacterial endosymbionts like that of *B. aphidicola*. This can be associated with the next evolutionary transition from a facultative lifestyle with broad tissue tropism within the insect host exemplified in the case of *S. glossinidius* to a more intimate association of the bacterial endosymbiont within specialized bacteriocyte host cells with more stable environment exemplified in the case of *B. aphidicola*. As is pointed out in discussion of chapter 5, the decrease in network robustness when the functionality of the metabolic network is analyzed in terms of biomass production contrast with the topological robustness of many biological networks, including metabolic networks, consequence of the power-law distribution of connectivity that favors robustness of the most highly connected nodes (Jeong et al., 2000; Barabasi and Oltvai, 2004). This is also observed when the patterns of network evolution by gene gain and loss in different prokaryotic lineages are analyzed, showing that below an apparent topological similarity, different biological networks evolve in diverse ways depending on the biological

objects included, the type of interactions and the parameters analyzed (Rocha, 2008a). For example, comparison of protein-protein interaction (PPI) networks in different bacterial lineages including host-dependent bacteria at different stages of the association reflects that PPI networks in bacterial lineages under reductive evolution evolves by preferential loss of proteins that interact with only one protein partner, retaining the most connected nodes, whereas regulatory networks evolves in the opposite way, by removing the most connected nodes that corresponds to the regulators while conserving the least connected nodes corresponding to the regulated elements (Ochman et al., 2007). In addition, whereas PPI networks evolves by preferential acquisition and loss of nodes that interacts directly with network hubs, in metabolic networks components tend to attach and detach from the periphery of the system in response to changing environments (Pal et al., 2005a).

Finally, minimal networks generated from *S. glossinidius* functional network under different external conditions have revealed differential evolutionary patterns between essential and non-essential genes defined based on their presence or absence in all minimal networks under each condition, with essential genes showing a more restricted pattern of sequence evolution at both synonymous and non-synonymous sites and more optimized codon usage than non-essential genes absents in all minimal networks, although there is variability between *S. glossindius* and *E. coli* K12 lineages when are analyzed independently consequence of the acceleration in the rates of sequence evolution in the lineage of *S. glossinidius* and the differences in the selective pressures over genes that becomes non-essential under a nutrient-rich environment similar to that found by *S. glossinidius* within the tsetse host. As described in the discussion of chapter 5, this differential patterns of sequence evolution between essential and non-essential genes were also observed by Jordan and collaborators in different bacterial lineages with experimentally defined essential genes based on knockout experiments (Jordan et al., 2002), and demonstrates the good correlation of FBA over genome-scale metabolic models with experimental data in predicting the essentiality of the genes for growth of the microorganism observed also in different studies (Schilling et al., 2002; Joyce et al., 2006). In addition, the observed correlation between gene dispensability and evolutionary rates fits well with predictions of the neutral theory of molecular evolution, which proposes that important proteins evolve more slowly than disposable ones (Kimura and Ota, 1974). However, there are discrepancies about if essentiality is the main determinant of protein evolution. Whereas global comparisons between essential and non-essential genes like that of Jordan and collaborators and similar studies with eukaryotic lineages like *S. cerevisiae* (Zhang and He, 2005) supports this view, studies analyzing the correlation between protein dispensability and evolutionary rate based on fitness measures of individual gene knockouts provide equivocal results. They alternate from a low, although significant, negative correlation between essentiality and rates of protein evolution in comparisons between *S. cerevisiae* and *C. elegans* (Hirsh and Fraser, 2001) and

between *S. cerevisiae* lineages (Zhang and He, 2005) to a non-significant correlation between both variables when genes evolving under positive selection were excluded in comparisons between mouse and rat (Hurst and Smith, 1999). In addition, statistical analyses considering both dispensability and gene expression levels together with rates of sequence evolution show that dispensability explains a relatively low fraction of the rate variation, with gene expression levels appearing as the most important determinant of the rates of protein evolution both in bacteria (Rocha and Danchin, 2004) and in eukaryotes (Drummond et al., 2005; Drummond et al., 2006). However this can be also consequence of the differences between the measures of fitness of experimental knockouts under laboratory conditions and the real importance of a particular protein in their natural environment, together with the fact that this measures of dispensability in extant species might not be representatives for the past evolution of the protein (Pal et al., 2006a). In fact, this negative correlation between dispensability and rate of evolution is stronger for closely related species (Zhang and He, 2005). The results of chapter 5 shows a significant negative correlation between CAI and *dN* values in both *S. glossinidius* and *E. coli* K12 lineages indicative of lower rates of sequence evolution in highly expressed genes (Pal et al., 2001; Rocha and Danchin, 2004), although is not possible to determine if is essentiality or expressiveness what determines the patterns of sequence evolution in *S. glossinidius*, but it is clear that on average, essential genes appear more evolutionarily conserved and with putatively higher expression levels than non-essential genes, in concordance with results obtained with experimentally measures of essentiality (Jordan et al., 2002; Fang et al., 2005).

7 General conclussions

- Gene order distances inferred from a conserved core set of genes present in multiple genomes of γ-proteobacteria are suitable for phylogenetic reconstruction, especially for closely related taxa. For more distant lineages the resolution power is reduced as a consequence of the faster loss of similarity in gene order than in sequence.

- Comparison of gene order and sequence evolution distances reveals that inversions are the predominant genome rearrangement in the evolution of γ-proteobacteria in light of the almost perfect correlation between breakpoint and inversion distances. In contrast, although positive correlation is observed between gene order and sequence evolution distances, marked heterogeneity is present in different lineages, finding groups like the *Pasteurellaceae* with extremely high number of rearrangements and other groups like different genomes of *B. aphidicola* with an almost complete conservation of gene order.

- Heterogeneity in the rates of gene order evolution is also revealed at different stages of the genome reduction process. The study of the relative rates of gene order evolution have revealed a significant acceleration in the rates of genome rearrangement in initial stages of the transition from a free-living ancestor that are exemplified in the evolution of *S. glossinidius*, the secondary endosymbiont of tsetse flies, where the rates of inversions are two times higher than in free-living enteric relatives. In contrast, in advanced stages of the genome reduction a complete stability is observed in terms of gene order evolution, exemplified by the almost strict collinearity observed between the three genomes of *B. aphidicola* included in the study.

- Heterogeneity in the rates of gene order evolution is also reflected in the large number of rearrangements whithin *Shigella* and *Yersinia* strains despite their recent divergence in contrast with the conservation of gene order between the more divergent lineages of *E. coli* and *Salmonella*. This acceleration is probably a consequence of the presence and proliferation of mobile genetic elements, specially IS, in *Shigella* and *Yersinia* genomes.

- The discrepancy about the relations of the primary endosymbionts of insects between phylogenies based on sequence data or gene, can be consequence of long branch attraction due to the compositional biases towards AT and the accelerated rates of sequence evolution in the former.

- A complete re-analysis of the genome sequence of *S. glossinidius* reveals novel insights in the initial stages of the transition from free-living to a host dependent lifestyle in this particular bacterial lineage. A detailed analysis of their intergenic regions has revealed a total of 1501 pseudogenes,

substantially higher than the 972 pseudogenes reported but unnanotated in the original genome paper.

- A functional re-annotation of the whole genome sequence including genes and newly characterized pseudogenes reveals a massive presence of CDSs related with mobile genetic elements, especially of phage origin, being also the functional class mostly affected by pseudogenization.

- 122 IS elements from five different families have proliferated across *S. glossinidius* genome sequence that, together with 8 single copy IS elements represents 2.71 % of the genome. IS proliferation has produced the inactivation of 18 pseudogenes, indicating that it has not been a major source of gene inactivation.

- The reconstruction of the complete metabolism of *S. glossinidius* based on the functional re-annotation of genes and pseudogenes reveals a metabolic profile highly similar to aerobic heterotrophic free-living enterics, with complete functionality of almost all biosynthetic pathways with the exception, not reported in the original annotation, of the pathway for L-arginine biosynthesis. In addition, L-alanine biosynthesis pathway is completely functional in contrast to the postulated inactivation in the original genome paper.

- Comparison of the metabolic profile of *S. glossinidius* with *W. glossinidia*, the primary endosymbiont of tsetse flies, reveals that almost all metabolic pathways functional in *W. glossinidia* are also functional in *S. glossinidius* with the exception of thiamin biosynthesis pathway. Detailed analysis of thiamin biosynthesis in *W. glossinidia* reveals a previously undescribed pseudogenization event at *thiF* gene and a possible metabolic complementation of both tsetse endosymbionts needed to produce the active form of the coenzyme thiamine diphosphate.

- Coenzyme A and folate biosynthesis pathways appear also severely impaired in *W. glossinidia*, that retains final stages of the pathways but are not able to synthesize the essential precursors p-aminobenzoate (for folate biosynthesis) , ketoisovalerate and β-alanine (for Coenzyme A biosynthesis). These pathways are completely functional in *S. glossinidius* that could supply these essential precursors to *W. glossinidia*.

- The reconstruction and stochiometric analysis of genome-scale metabolic networks of *S. glossinidius* at different stages of the genome reduction process confirms the very recent transition of *S. glossinidius* to host-dependent lifestyle. This is revealed by the complete functionality of *S. glossinidius* ancestral metabolic network under minimal aerobic conditions

with only glucose as external carbon source in FBA simulations with biomass production as objective function, with similar biomass production rates as free-living enteric like *E. coli*.

- The pseudogenization of genes involved in glycogen biosynthesis, L-arginine biosynthesis and the gene *ppc* encoding the key anaplerotic enzyme phosphoenolpyruvate carboxykinase can explain the transition to host-dependent lifestyle. This is reflected by the absolute requirement of external L-arginine supply and glycogen removal from biomass equation in order to obtain a functional phenotype in terms of biomass production once pseudogenes and deleted genes have been eliminated from ancestral metabolic network.

- A significant decrease in the robustness of metabolic networks to gene deletion events and to changes in enzymatic activities is observed as consequence of the process of gene inactivation, starting from a highly robust metabolic system like that of a free living bacterium like *E. coli*, following with a bacterium at initial stages of the transition to host-dependent lifestyle like *S. glossinidius*, where the fraction of essential genes overcomes that of non-essential genes, and finishing with the highly streamlined metabolic system of an ancient bacterial endosymbiont like *B. aphidicola*, where most genes appears as essential for the functionality of the metabolic system..

- Reductive evolution simulations over *S. glossinidius* functional network under different external condition reveals that gene and reaction content of minimal networks is strongly dependent of the environmental conditions under which the system are evolving.

- Analysis of the patterns of codon usage and sequence evolution at synonymous and non-synonymous sites over essential and non-essential genes in minimal metabolic networks reveals that essential genes common to all minimal networks have more restricted patterns of sequence evolution and more optimized patterns of codon usage than non-essential genes absent in all minimal networks.

- Analysis of the patterns of sequence evolution in *S. glossinidius* and *E. coli* lineages after divergence from their common ancestor reveals a significant acceleration in the rates of sequence evolution at both non-synonymous and synonymous sites in the lineage of *S. glossinidius* in a similar manner as in more ancient primary endosymbionts. The acceleration in synonymous substitution rates are affecting to the strength of adaptative codon usage in

comparison with *E. coli* lineage, reflected by the absence of significant correlation between CAI and *dS* in *S. glossinidius* lineage.

- Differential patterns of sequence evolution between *S. glossinidius* and *E. coli* lineages are observed in conditionally essential genes that can be lost under nutrient-rich environment, and that corresponds with many genes that have been lost in the more ancient tsetse primary endosymbiont *W. glossinidia*. Whereas in *E. coli* these genes show significant restricted patterns of sequence evolution at both synonymous and non-synonymous sites in comparison with core essential genes in all minimal networks under all conditions, in *S. glossinidius* they show more relaxed patterns of sequence evolution, specially evident at synonymous sites. This can be explained by the decrease in the selective pressures for conservation of gene function over these genes that can be lost under a nutrient-rich environment similar to that found by *S. glossinidius* in the context of its association with the tsetse host.

8 Bibliography

Bibliography

Abby,S. and Daubin,V. (2007). Comparative genomics and the evolution of prokaryotes. Trends Microbiol. *15*, 135-141.

Achaz,G., Coissac,E., Netter,P., and Rocha,E.P. (2003). Associations between inverted repeats and the structural evolution of bacterial genomes. Genetics *164*, 1279-1289.

Achaz,G., Rocha,E.P., Netter,P., and Coissac,E. (2002). Origin and fate of repeats in bacteria. Nucleic Acids Res. *30*, 2987-2994.

Achtman,M., Zurth,K., Morelli,G., Torrea,G., Guiyoule,A., and Carniel,E. (1999). Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proc. Natl. Acad. Sci U. S. A *96*, 14043-14048.

Akerley,B.J., Rubin,E.J., Novick,V.L., Amaya,K., Judson,N., and Mekalanos,J.J. (2002). A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. Proc. Natl. Acad. Sci. U. S. A *99*, 966-971.

Akman,L., Yamashita,A., Watanabe,H., Oshima,K., Shiba,T., Hattori,M., and Aksoy,S. (2002). Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia. Nat. Genet. *32*, 402-407.

Aksoy,S. (2000). Tsetse--A haven for microorganisms. Parasitol. Today *16*, 114-118.

Aksoy,S. (1995). Wigglesworthia gen. nov. and Wigglesworthia glossinidia sp. nov., taxa consisting of the mycetocyte-associated, primary endosymbionts of tsetse flies. Int. J. Syst. Bacteriol. *45*, 848-851.

Aksoy,S., Berriman,M., Hall,N., Hattori,M., Hide,W., and Lehane,M.J. (2005). A case for a Glossina genome project. Trends Parasitol. *21*, 107-111.

Aksoy,S., Chen,X., and Hypsa,V. (1997). Phylogeny and potential transmission routes of midgut-associated endosymbionts of tsetse (Diptera:Glossinidae). Insect Mol. Biol. *6*, 183-190.

Aksoy,S., Gibson,W.C., and Lehane,M.J. (2003). Interactions between tsetse and trypanosomes with implications for the control of trypanosomiasis. Adv. Parasitol. *53*, 1-83.

Aksoy,S. and Rio,R.V. (2005). Interactions among multiple genomes: tsetse, its symbionts and trypanosomes. Insect Biochem. Mol. Biol. *35*, 691-698.

Albert,R. (2005). Scale-free networks in cell biology. J. Cell Sci. *118*, 4947-4957.

Albert,R., Jeong,H., and Barabasi,A.L. (2000). Error and attack tolerance of complex networks. Nature *406*, 378-382.

Allison,G.E., Angeles,D., Tran-Dinh,N., and Verma,N.K. (2002). Complete genomic sequence of SfV, a serotype-converting temperate bacteriophage of Shigella flexneri. J. Bacteriol. *184*, 1974-1987.

Almaas,E. (2007). Biological impacts and context of network theory. J. Exp. Biol. *210*, 1548-1558.

Almaas,E., Kovacs,B., Vicsek,T., Oltvai,Z.N., and Barabasi,A.L. (2004). Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature *427*, 839-843.

Almaas,E., Oltvai,Z.N., and Barabasi,A.L. (2005). The activity reaction core and plasticity of metabolic networks. PLoS Comput. Biol. *1*, e68.

Alper,H., Jin,Y.S., Moxley,J.F., and Stephanopoulos,G. (2005a). Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli. Metab Eng *7*, 155-164.

Alper,H., Miyaoku,K., and Stephanopoulos,G. (2005b). Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets. Nat. Biotechnol. *23*, 612-616.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., and Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389-3402.

Anderson,R.P. and Roth,J.R. (1977). Tandem genetic duplications in phage and bacteria. Annu. Rev. Microbiol. *31*, 473-505.

Andersson,D.I. and Hughes,D. (1996). Muller's ratchet decreases fitness of a DNA-based microbe. Proc. Natl. Acad. Sci. U. S. A *93*, 906-907.

Andersson,J.O. and Andersson,S.G. (1999a). Genome degradation is an ongoing process in Rickettsia. Mol. Biol. Evol. *16*, 1178-1191.

Andersson,J.O. and Andersson,S.G. (1999b). Insights into the evolutionary process of genome degradation. Curr. Opin. Genet. Dev. *9*, 664-671.

Andersson,J.O. and Andersson,S.G. (2001). Pseudogenes, junk DNA, and the dynamics of Rickettsia genomes. Mol. Biol. Evol. *18*, 829-839.

Andersson,S.G., Alsmark,C., Canback,B., Davids,W., Frank,C., Karlberg,O., Klasson,L., ntoine-Legault,B., Mira,A., and Tamas,I. (2002). Comparative genomics of microbial pathogens and symbionts. Bioinformatics. *18 Suppl 2*, S17.

Andersson,S.G. and Kurland,C.G. (1998). Reductive evolution of resident genomes. Trends Microbiol. *6*, 263-268.

Bibliography

Andersson,S.G., Zomorodipour,A., Andersson,J.O., Sicheritz-Ponten,T., Alsmark,U.C., Podowski,R.M., Naslund,A.K., Eriksson,A.S., Winkler,H.H., and Kurland,C.G. (1998). The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature *396*, 133-140.

Aras,R.A., Kang,J., Tschumi,A.I., Harasaki,Y., and Blaser,M.J. (2003). Extensive repetitive DNA facilitates prokaryotic genome plasticity. Proc. Natl. Acad. Sci. U. S. A *100*, 13579-13584.

Aravind,L., Tatusov,R.L., Wolf,Y.I., Walker,D.R., and Koonin,E.V. (1998). Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet *14*, 442-444.

Arifuzzaman,M., Maeda,M., Itoh,A., Nishikata,K., Takita,C., Saito,R., Ara,T., Nakahigashi,K., Huang,H.C., Hirai,A., Tsuzuki,K., Nakamura,S., taf-Ul-Amin,M., Oshima,T., Baba,T., Yamamoto,N., Kawamura,T., Ioka-Nakamichi,T., Kitagawa,M., Tomita,M., Kanaya,S., Wada,C., and Mori,H. (2006). Large-scale identification of protein-protein interaction of Escherichia coli K-12. Genome Res. *16*, 686-691.

Arita,M. (2000). Metabolic reconstruction using shortest paths. Simulat. Pract. Theory *8*, 109-125.

Ashelford,K.E., Day,M.J., and Fry,J.C. (2003). Elevated abundance of bacteriophage infecting bacteria in soil. Appl. Environ. Microbiol. *69*, 285-289.

Attardo,G.M., Lohs,C., Heddi,A., Alam,U.H., Yildirim,S., and Aksoy,S. (2008). Analysis of milk gland structure and function in Glossina morsitans: milk protein production, symbiont populations and fecundity. J. Insect Physiol *54*, 1236-1242.

Awano,N., Wada,M., Kohdoh,A., Oikawa,T., Takagi,H., and Nakamori,S. (2003). Effect of cysteine desulfhydrase gene disruption on L-cysteine overproduction in Escherichia coli. Appl. Microbiol. Biotechnol. *62*, 239-243.

Babinski,K.J., Kanjilal,S.J., and Raetz,C.R. (2002a). Accumulation of the lipid A precursor UDP-2,3-diacylglucosamine in an Escherichia coli mutant lacking the lpxH gene. J. Biol. Chem. *277*, 25947-25956.

Babinski,K.J., Ribeiro,A.A., and Raetz,C.R. (2002b). The Escherichia coli gene encoding the UDP-2,3-diacylglucosamine pyrophosphatase of lipid A biosynthesis. J. Biol. Chem. *277*, 25937-25946.

Babu,M.M., Luscombe,N.M., Aravind,L., Gerstein,M., and Teichmann,S.A. (2004). Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. *14*, 283-291.

Bach,S., de,A.A., and Carniel,E. (2000). The Yersinia high-pathogenicity island is present in different members of the family Enterobacteriaceae. FEMS Microbiol. Lett. *183*, 289-294.

Bailey,J.E. (2001). Complex biology with no parameters. Nat. Biotechnol. *19*, 503-504.

Bakkali,M., Chen,T.Y., Lee,H.C., and Redfield,R.J. (2004). Evolutionary stability of DNA uptake signal sequences in the Pasteurellaceae. Proc. Natl. Acad. Sci. U. S. A *101*, 4513-4518.

Banks,D.J., Lei,B., and Musser,J.M. (2003). Prophage induction and expression of prophage-encoded virulence factors in group A Streptococcus serotype M3 strain MGAS315. Infect. Immun. *71*, 7079-7086.

Bapteste,E., Boucher,Y., Leigh,J., and Doolittle,W.F. (2004). Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol. *12*, 406-411.

Barabasi,A.L. and Albert,R. (1999). Emergence of scaling in random networks. Science *286*, 509-512.

Barabasi,A.L. and Oltvai,Z.N. (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet *5*, 101-113.

Batey,R.T., Rambo,R.P., Lucast,L., Rha,B., and Doudna,J.A. (2000). Crystal structure of the ribonucleoprotein core of the signal recognition particle. Science *287*, 1232-1239.

Baumann,P. (2005). Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. Annu. Rev. Microbiol. *59*, 155-189.

Baumann,P., Baumann,L., Lai,C.Y., Rouhbakhsh,D., Moran,N.A., and Clark,M.A. (1995). Genetics, physiology, and evolutionary relationships of the genus Buchnera: intracellular symbionts of aphids. Annu. Rev. Microbiol. *49*, 55-94.

Beale,S.I. (1996). Biosynthesis of hemes. In Escherichia coli and Salmonella cellular and molecular biology, F.C.Neidhardt, ed. (Washington D.C.: American Society for Microbiology), pp. 731-748.

Beard,C.B., O'Neill,S.L., Mason,P., Mandelco,L., Woese,C.R., Tesh,R.B., Richards,F.F., and Aksoy,S. (1993). Genetic transformation and phylogeny of bacterial symbionts from tsetse. Insect Mol. Biol. *1*, 123-131.

Becker,S.A., Feist,A.M., Mo,M.L., Hannum,G., Palsson,B.O., and Herrgard,M.J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat. Protoc. *2*, 727-738.

Bibliography

Becker,S.A. and Palsson,B.O. (2005). Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. BMC. Microbiol. *5*, 8.

Begley,T.P., Downs,D.M., Ealick,S.E., McLafferty,F.W., Van Loon,A.P., Taylor,S., Campobasso,N., Chiu,H.J., Kinsland,C., Reddick,J.J., and Xi,J. (1999). Thiamin biosynthesis in prokaryotes. Arch. Microbiol. *171*, 293-300.

Beiko,R.G., Harlow,T.J., and Ragan,M.A. (2005). Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. U. S. A *102*, 14332-14337.

Bentley,S.D. and Parkhill,J. (2004). Comparative genomic structure of prokaryotes. Annu. Rev. Genet *38*, 771-792.

Beres,S.B., Sylva,G.L., Barbian,K.D., Lei,B., Hoff,J.S., Mammarella,N.D., Liu,M.Y., Smoot,J.C., Porcella,S.F., Parkins,L.D., Campbell,D.S., Smith,T.M., McCormick,J.K., Leung,D.Y., Schlievert,P.M., and Musser,J.M. (2002). Genome sequence of a serotype M3 strain of group A Streptococcus: phage-encoded toxins, the high-virulence phenotype, and clone emergence. Proc. Natl. Acad. Sci. U. S. A *99*, 10078-10083.

Berg,C.M. and Rossi,J.J. (1974). Proline excretion and indirect suppression in Escherichia coli and Salmonella typhimurium. J. Bacteriol. *118*, 928-934.

Berg,D.E. (1977). Insertion and excision of the transposable kanamycin resistance determinant Tn5. In DNA Insertion Elements, Plasmids and Episomes, A.I.Bukhari, J.S.Shapiro, and S.L.Adhya, eds. (New York: Cold Spring Harbor), pp. 205-212.

Bergthorsson,U. and Ochman,H. (1998). Distribution of chromosome length variation in natural isolates of Escherichia coli. Mol. Biol. Evol. *15*, 6-16.

Bininda-Emonds,O.R. (2004). Trees versus characters and the supertree/supermatrix "paradox". Syst. Biol. *53*, 356-359.

Binnewies,T.T., Motro,Y., Hallin,P.F., Lund,O., Dunn,D., La,T., Hampson,D.J., Bellgard,M., Wassenaar,T.M., and Ussery,D.W. (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. Funct. Integr. Genomics *6*, 165-185.

Birney,E., Clamp,M., and Durbin,R. (2004). GeneWise and Genomewise. Genome Res. *14*, 988-995.

Blanc,G., Ogata,H., Robert,C., Audic,S., Suhre,K., Vestris,G., Claverie,J.M., and Raoult,D. (2007). Reductive genome evolution from the mother of Rickettsia. PLoS Genet *3*, e14.

Blanchette,M., Kunisawa,T., and Sankoff,D. (1999). Gene order breakpoint evidence in animal mitochondrial phylogeny. J. Mol. Evol. *49*, 193-203.

Blank,L.M., Kuepfer,L., and Sauer,U. (2005). Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. Genome Biol. *6*, R49.

Blot,M. (1994). Transposable elements and adaptation of host bacteria. Genetica *93*, 5-12.

Bonarius,H.P.J., Schmid,G., and Tramper,J. (1997). Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. Trends in Biotechnology *15*, 308-314.

Bordenstein,S.R., Marshall,M.L., Fry,A.J., Kim,U., and Wernegreen,J.J. (2006). The tripartite associations between bacteriophage, Wolbachia, and arthropods. PLoS. Pathog. *2*, e43.

Bordenstein,S.R. and Reznikoff,W.S. (2005). Mobile DNA in obligate intracellular bacteria. Nat. Rev. Microbiol. *3*, 688-699.

Bordenstein,S.R. and Wernegreen,J.J. (2004). Bacteriophage flux in endosymbionts (Wolbachia): infection frequency, lateral transfer, and recombination rates. Mol. Biol. Evol. *21*, 1981-1991.

Borodina,I., Krabben,P., and Nielsen,J. (2005). Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. Genome Res. *15*, 820-829.

Borodina,I. and Nielsen,J. (2005). From genomes to in silico cells via metabolic networks. Curr. Opin. Biotechnol. *16*, 350-355.

Boschi-Muller,S., Azza,S., Pollastro,D., Corbier,C., and Branlant,G. (1997). Comparative enzymatic properties of GapB-encoded erythrose-4-phosphate dehydrogenase of Escherichia coli and phosphorylating glyceraldehyde-3-phosphate dehydrogenase. J. Biol. Chem. *272*, 15106-15112.

Botstein,D. (1980). A theory of modular evolution for bacteriophages. Ann. N. Y. Acad. Sci. *354*, 484-490.

Boucher,Y., Douady,C.J., Papke,R.T., Walsh,D.A., Boudreau,M.E., Nesbo,C.L., Case,R.J., and Doolittle,W.F. (2003). Lateral gene transfer and the origins of prokaryotic groups. Annu. Rev. Genet. *37*, 283-328.

Bourque,G. and Pevzner,P.A. (2002). Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Res. *12*, 26-36.

Boussau,B., Karlberg,E.O., Frank,A.C., Legault,B.A., and Andersson,S.G. (2004). Computational inference of scenarios for alpha-proteobacterial genome evolution. Proc. Natl. Acad. Sci. U. S. A *101*, 9722-9727.

Bibliography

Braun,V. (1995). Energy-coupled transport and signal transduction through the gram-negative outer membrane via TonB-ExbB-ExbD-dependent receptor proteins. FEMS Microbiol. Rev. *16*, 295-307.

Breed,R.S., Murray,E.G.D., and Smith,N.R. (1957). Bergey´s manual of determinative bacteriology. (Baltimore: Williams and Wilkins).

Breitbart,M., Salamon,P., Andresen,B., Mahaffy,J.M., Segall,A.M., Mead,D., Azam,F., and Rohwer,F. (2002). Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U. S. A *99*, 14250-14255.

Brenner,D.J., Staley,J.T., and Krieg,N.R. (2005). Classification of Procaryotic Organisms and the Concept of Bacterial Speciation. In Bergey´s manual of systematic bacteriology, D.J.Brenner, J.T.Staley, N.R.Krieg, and G.Garrity, eds. (New York: SPRINGER), pp. 27-32.

Brenner,S.E., Hubbard,T., Murzin,A., and Chothia,C. (1995). Gene duplications in H. influenzae. Nature *378*, 140.

Brochier,C., Philippe,H., and Moreira,D. (2000). The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Genet *16*, 529-533.

Brosch,R., Pym,A.S., Gordon,S.V., and Cole,S.T. (2001). The evolution of mycobacterial pathogenicity: clues from comparative genomics. Trends Microbiol. *9*, 452-458.

Brown,J.R. (2003). Ancient horizontal gene transfer. Nat. Rev. Genet *4*, 121-132.

Brown,J.R., Douady,C.J., Italia,M.J., Marshall,W.E., and Stanhope,M.J. (2001). Universal trees based on large combined protein sequence data sets. Nat. Genet *28*, 281-285.

Brown,J.R. and Volker,C. (2004). Phylogeny of gamma-proteobacteria: resolution of one branch of the universal tree? Bioessays *26*, 463-468.

Brownlie,J.C. and O'Neill,S.L. (2005). Wolbachia genomes: insights into an intracellular lifestyle. Curr. Biol. *15*, R507-R509.

Brussow,H., Canchaya,C., and Hardt,W.D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol. Mol. Biol. Rev. *68*, 560-602.

Brussow,H. and Hendrix,R.W. (2002). Phage genomics: small is beautiful. Cell *108*, 13-16.

Brynnel,E.U., Kurland,C.G., Moran,N.A., and Andersson,S.G. (1998). Evolutionary rates for tuf genes in endosymbionts of aphids. Mol. Biol. Evol. *15*, 574-582.

Buchrieser,C., Brosch,R., Bach,S., Guiyoule,A., and Carniel,E. (1998). The high-pathogenicity island of Yersinia pseudotuberculosis can be inserted into any of the three chromosomal asn tRNA genes. Mol. Microbiol. *30*, 965-978.

Burgard,A.P. and Maranas,C.D. (2003). Optimization-based framework for inferring and testing hypothesized metabolic objective functions. Biotechnol. Bioeng. *82*, 670-677.

Burgard,A.P., Pharkya,P., and Maranas,C.D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol. Bioeng. *84*, 647-657.

Butland,G., Peregrin-Alvarez,J.M., Li,J., Yang,W., Yang,X., Canadien,V., Starostine,A., Richards,D., Beattie,B., Krogan,N., Davey,M., Parkinson,J., Greenblatt,J., and Emili,A. (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature *433*, 531-537.

Campbell,A. (1981). Evolutionary significance of accessory DNA elements in bacteria. Annu. Rev. Microbiol. *35*, 55-83.

Campbell,A., Berg,D.E., Botstein,D., Lederberg,E.M., Novick,R.P., Starlinger,P., and Szybalski,W. (1977). Nomenclature of transposable elements in prokaryotes. In DNA Insertion Elements, Plasmids and Episomes, A.I.Bukhari, J.S.Shapiro, and S.L.Adhya, eds. (New York: Cold Spring Harbor), pp. 15-22.

Campbell,A., Berg,D.E., Botstein,D., Lederberg,E.M., Novick,R.P., Starlinger,P., and Szybalski,W. (1979a). Nomenclature of transposable elements in prokaryotes. Gene *5*, 197-206.

Campbell,A., Starlinger,P., Berg,D.E., Botstein,D., Lederberg,E.M., Novick,R.P., and Szybalski,W. (1979b). Nomenclature of transposable elements in prokaryotes. Plasmid *2*, 466-473.

Campo,N., Dias,M.J., veran-Mingot,M.L., Ritzenthaler,P., and Le,B.P. (2004). Chromosomal constraints in Gram-positive bacteria revealed by artificial inversions. Mol. Microbiol. *51*, 511-522.

Canback,B., Tamas,I., and Andersson,S.G. (2004). A phylogenomic study of endosymbiotic bacteria. Mol. Biol. Evol. *21*, 1110-1122.

Canchaya,C., Fournous,G., and Brussow,H. (2004). The impact of prophages on bacterial chromosomes. Mol. Microbiol. *53*, 9-18.

Canchaya,C., Proux,C., Fournous,G., Bruttin,A., and Brussow,H. (2003). Prophage genomics. Microbiol. Mol. Biol. Rev. *67*, 238-76, table.

Bibliography

Carniel,E., Guilvout,I., and Prentice,M. (1996). Characterization of a large chromosomal "high-pathogenicity island" in biotype 1B Yersinia enterocolitica. J. Bacteriol. *178*, 6743-6751.

Carver,T., Berriman,M., Tivey,A., Patel,C., Bohme,U., Barrell,B.G., Parkhill,J., and Rajandream,M.A. (2008). Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics. *24*, 2672-2676.

Carver,T., Thomson,N., Bleasby,A., Berriman,M., and Parkhill,J. (2009). DNAPlotter: circular and linear interactive genome visualization. Bioinformatics. *25*, 119-120.

Carver,T.J., Rutherford,K.M., Berriman,M., Rajandream,M.A., Barrell,B.G., and Parkhill,J. (2005). ACT: the Artemis Comparison Tool. Bioinformatics. *21*, 3422-3423.

Casjens,S. (2003). Prophages and bacterial genomics: what have we learned so far? Mol. Microbiol. *49*, 277-300.

Casjens,S. (1998). The diverse and dynamic structure of bacterial genomes. Annu. Rev. Genet *32*, 339-377.

Casjens,S.R. (2005). Comparative genomics and evolution of the tailed-bacteriophages. Curr. Opin. Microbiol. *8*, 451-458.

Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F., Kaipa,P., Karthikeyan,A.S., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A., Paley,S., Popescu,L., Pujar,A., Shearer,A.G., Zhang,P., and Karp,P.D. (2009). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res.

Castresana,J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17*, 540-552.

Cavalier-Smith,T. (2005). Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann. Bot. (Lond) *95*, 147-175.

Chain,P.S., Carniel,E., Larimer,F.W., Lamerdin,J., Stoutland,P.O., Regala,W.M., Georgescu,A.M., Vergez,L.M., Land,M.L., Motin,V.L., Brubaker,R.R., Fowler,J., Hinnebusch,J., Marceau,M., Medigue,C., Simonet,M., Chenal-Francisque,V., Souza,B., Dacheux,D., Elliott,J.M., Derbise,A., Hauser,L.J., and Garcia,E. (2004). Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. Proc. Natl. Acad. Sci. U. S. A *101*, 13826-13831.

Chain,P.S., Hu,P., Malfatti,S.A., Radnedge,L., Larimer,F., Vergez,L.M., Worsham,P., Chu,M.C., and Andersen,G.L. (2006). Complete genome sequence of

Yersinia pestis strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen. J. Bacteriol. *188*, 4453-4463.

Chalmers,R. and Blot,M. (1999). Insertion Sequences and Transposons. In Organization of the Prokaryotic Genome, R.L.Charlebois, ed. (Washington: American Society for Microbiology), pp. 151-171.

Charlebois,R.L. and Doolittle,W.F. (2004). Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res. *14*, 2469-2477.

Charles,H., Heddi,A., Guillaud,J., Nardon,C., and Nardon,P. (1997). A molecular aspect of symbiotic interactions between the weevil Sitophilus oryzae and its endosymbiotic bacteria: over-expression of a chaperonin. Biochem. Biophys. Res. Commun. *239*, 769-774.

Charles,H., Heddi,A., and Rahbe,Y. (2001). A putative insect intracellular endosymbiont stem clade, within the Enterobacteriaceae, infered from phylogenetic analysis based on a heterogeneous model of DNA evolution. C. R. Acad. Sci. III *324*, 489-494.

Chen,X., Li,S., and Aksoy,S. (1999). Concordant evolution of a symbiont with its host insect species: molecular phylogeny of genus Glossina and its bacteriome-associated endosymbiont, Wigglesworthia glossinidia. J. Mol. Evol. *48*, 49-58.

Chen,Y., Yu,P., Luo,J., and Jiang,Y. (2003). Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. Mamm. Genome *14*, 859-865.

Cheng,Q. and Aksoy,S. (1999). Tissue tropism, transmission and expression of foreign genes in vivo in midgut symbionts of tsetse flies. Insect Mol. Biol. *8*, 125-132.

Christensen,B. and Nielsen,J. (2000a). Metabolic network analysis of Penicillium chrysogenum using (13)C-labeled glucose. Biotechnol. Bioeng. *68*, 652-659.

Christensen,B. and Nielsen,J. (2000b). Metabolic network analysis. A powerful tool in metabolic engineering. Adv. Biochem. Eng Biotechnol. *66*, 209-231.

Christensen,H., Kuhnert,P., Olsen,J.E., and Bisgaard,M. (2004). Comparative phylogenies of the housekeeping genes atpD, infB and rpoB and the 16S rRNA gene within the Pasteurellaceae. Int. J. Syst. Evol. Microbiol. *54*, 1601-1609.

Clark,A.J., Inwood,W., Cloutier,T., and Dhillon,T.S. (2001). Nucleotide sequence of coliphage HK620 and the evolution of lambdoid phages. J. Mol. Biol. *311*, 657-679.

Clark,A.J., Pontes,M., Jones,T., and Dale,C. (2007). A possible heterodimeric prophage-like element in the genome of the insect endosymbiont Sodalis glossinidius. J. Bacteriol. *189*, 2949-2951.

Bibliography

Clark,M.A., Moran,N.A., and Baumann,P. (1999). Sequence evolution in bacterial endosymbionts having extreme base compositions. Mol. Biol. Evol. *16*, 1586-1598.

Cohan,F.M. (2001). Bacterial species and speciation. Syst. Biol. *50*, 513-524.

Cole,S.T. (1998). Comparative mycobacterial genomics. Curr. Opin. Microbiol. *1*, 567-571.

Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R., Honore,N., Garnier,T., Churcher,C., Harris,D., Mungall,K., Basham,D., Brown,D., Chillingworth,T., Connor,R., Davies,R.M., Devlin,K., Duthoy,S., Feltwell,T., Fraser,A., Hamlin,N., Holroyd,S., Hornsby,T., Jagels,K., Lacroix,C., Maclean,J., Moule,S., Murphy,L., Oliver,K., Quail,M.A., Rajandream,M.A., Rutherford,K.M., Rutter,S., Seeger,K., Simon,S., Simmonds,M., Skelton,J., Squares,R., Squares,S., Stevens,K., Taylor,K., Whitehead,S., Woodward,J.R., and Barrell,B.G. (2001). Massive gene decay in the leprosy bacillus. Nature *409*, 1007-1011.

Comas,I., Moya,A., Azad,R.K., Lawrence,J.G., and Gonzalez-Candelas,F. (2006). The evolutionary origin of Xanthomonadales genomes and the nature of the horizontal gene transfer process. Mol. Biol. Evol. *23*, 2049-2057.

Comas,I., Moya,A., and Gonzalez-Candelas,F. (2007). From phylogenetics to phylogenomics: the evolutionary relationships of insect endosymbiotic gamma-Proteobacteria as a test case. Syst. Biol. *56*, 1-16.

Comeau,A.M., Hatfull,G.F., Krisch,H.M., Lindell,D., Mann,N.H., and Prangishvili,D. (2008). Exploring the prokaryotic virosphere. Res. Microbiol. *159*, 306-313.

Cooper,J.E. and Feil,E.J. (2004). Multilocus sequence typing--what is resolved? Trends Microbiol. *12*, 373-377.

Cooper,S. (1991). Synthesis of the cell surface during the division cycle of rod-shaped, gram-negative bacteria. Microbiol. Rev. *55*, 649-674.

Cortez,D., Forterre,P., and Gribaldo,S. (2009). A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. Genome Biol. *10*, R65.

Covert,M.W. and Palsson,B.O. (2002). Transcriptional regulation in constraints-based metabolic models of Escherichia coli. J. Biol. Chem. *277*, 28058-28064.

Covert,M.W. and Palsson,B.O. (2003). Constraints-based models: regulation of gene expression reduces the steady-state solution space. J. Theor. Biol. *221*, 309-325.

Covert,M.W., Schilling,C.H., Famili,I., Edwards,J.S., Goryanin,I.I., Selkov,E., and Palsson,B.O. (2001). Metabolic modeling of microbial strains in silico. Trends Biochem. Sci. *26*, 179-186.

Cox,R.J. and Wang,P.S. (2001). Is N-acetylornithine aminotransferase the real N-succinyl-ll-diaminopimelate aminotransferase in E. coli and M. smegmatis? J. Chem. Soc. Perkin Trans. *1*, 2006-2008.

Cronan,J.E., Jr., Littel,K.J., and Jackowski,S. (1982). Genetic and biochemical analyses of pantothenate biosynthesis in Escherichia coli and Salmonella typhimurium. J. Bacteriol. *149*, 916-922.

Cunin,R., Glansdorff,N., Pierard,A., and Stalon,V. (1986). Biosynthesis and metabolism of arginine in bacteria. Microbiol. Rev. *50*, 314-352.

Curtis,S.J. and Epstein,W. (1975). Phosphorylation of D-glucose in Escherichia coli mutants defective in glucosephosphotransferase, mannosephosphotransferase, and glucokinase. J. Bacteriol. *122*, 1189-1199.

da Silva,A.C., Ferro,J.A., Reinach,F.C., Farah,C.S., Furlan,L.R., Quaggio,R.B., Monteiro-Vitorello,C.B., Van Sluys,M.A., Almeida,N.F., Alves,L.M., do Amaral,A.M., Bertolini,M.C., Camargo,L.E., Camarotte,G., Cannavan,F., Cardozo,J., Chambergo,F., Ciapina,L.P., Cicarelli,R.M., Coutinho,L.L., Cursino-Santos,J.R., El-Dorry,H., Faria,J.B., Ferreira,A.J., Ferreira,R.C., Ferro,M.I., Formighieri,E.F., Franco,M.C., Greggio,C.C., Gruber,A., Katsuyama,A.M., Kishi,L.T., Leite,R.P., Lemos,E.G., Lemos,M.V., Locali,E.C., Machado,M.A., Madeira,A.M., Martinez-Rossi,N.M., Martins,E.C., Meidanis,J., Menck,C.F., Miyaki,C.Y., Moon,D.H., Moreira,L.M., Novo,M.T., Okura,V.K., Oliveira,M.C., Oliveira,V.R., Pereira,H.A., Rossi,A., Sena,J.A., Silva,C., de Souza,R.F., Spinola,L.A., Takita,M.A., Tamura,R.E., Teixeira,E.C., Tezza,R.I., Trindade dos,S.M., Truffi,D., Tsai,S.M., White,F.F., Setubal,J.C., and Kitajima,J.P. (2002). Comparison of the genomes of two Xanthomonas pathogens with differing host specificities. Nature *417*, 459-463.

Dagan,T. and Martin,W. (2006). The tree of one percent. Genome Biol. *7*, 118.

Dale,C., Jones,T., and Pontes,M. (2005). Degenerative evolution and functional diversification of type-III secretion systems in the insect endosymbiont Sodalis glossinidius. Mol. Biol. Evol. *22*, 758-766.

Dale,C. and Maudlin,I. (1999). Sodalis gen. nov. and Sodalis glossinidius sp. nov., a microaerophilic secondary endosymbiont of the tsetse fly Glossina morsitans morsitans. Int. J. Syst. Bacteriol. *49 Pt 1*, 267-275.

Dale,C. and Moran,N.A. (2006). Molecular interactions between bacterial symbionts and their hosts. Cell *126*, 453-465.

Dale,C., Plague,G.R., Wang,B., Ochman,H., and Moran,N.A. (2002). Type III secretion systems and the evolution of mutualistic endosymbiosis. Proc. Natl. Acad. Sci. U. S. A *99*, 12397-12402.

Bibliography

Dale,C., Wang,B., Moran,N., and Ochman,H. (2003). Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. Mol. Biol. Evol. *20*, 1188-1194.

Dale,C. and Welburn,S.C. (2001). The endosymbionts of tsetse flies: manipulating host-parasite interactions. Int. J. Parasitol. *31*, 628-631.

Dale,C., Young,S.A., Haydon,D.T., and Welburn,S.C. (2001). The insect endosymbiont Sodalis glossinidius utilizes a type III secretion system for cell invasion. Proc. Natl. Acad. Sci. U. S. A *98*, 1883-1888.

Dandekar,T., Huynen,M., Regula,J.T., Ueberle,B., Zimmermann,C.U., Andrade,M.A., Doerks,T., Sanchez-Pulido,L., Snel,B., Suyama,M., Yuan,Y.P., Herrmann,R., and Bork,P. (2000). Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames. Nucleic Acids Res. *28*, 3278-3288.

Darby,A.C., Cho,N.H., Fuxelius,H.H., Westberg,J., and Andersson,S.G. (2007). Intracellular pathogens go extreme: genome evolution in the Rickettsiales. Trends Genet. *23*, 511-520.

Darby,A.C., Lagnel,J., Matthew,C.Z., Bourtzis,K., Maudlin,I., and Welburn,S.C. (2005). Extrachromosomal DNA of the symbiont Sodalis glossinidius. J. Bacteriol. *187*, 5003-5007.

Darling,A.E., Miklos,I., and Ragan,M.A. (2008). Dynamics of genome rearrangement in bacterial populations. PLoS. Genet. *4*, e1000128.

Daubin,V., Gouy,M., and Perriere,G. (2002). A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res. *12*, 1080-1090.

Daubin,V. and Moran,N.A. (2004). Comment on "The origins of genome complexity". Science *306*, 978.

Daubin,V., Moran,N.A., and Ochman,H. (2003). Phylogenetics and the cohesion of bacterial genomes. Science *301*, 829-832.

Daubin,V. and Ochman,H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res. *14*, 1036-1042.

de Kievit,T.R. and Iglewski,B.H. (2000). Bacterial quorum sensing in pathogenic relationships. Infect. Immun. *68*, 4839-4849.

Degnan,P.H., Lazarus,A.B., Brock,C.D., and Wernegreen,J.J. (2004). Host-symbiont stability and fast evolutionary rates in an ant-bacterium association: cospeciation of camponotus species and their endosymbionts, candidatus blochmannia. Syst. Biol. *53*, 95-110.

Degnan,P.H., Lazarus,A.B., and Wernegreen,J.J. (2005). Genome sequence of Blochmannia pennsylvanicus indicates parallel evolutionary trends among bacterial mutualists of insects. Genome Res. *15*, 1023-1033.

Degnan,P.H., Leonardo,T.E., Cass,B.N., Hurwitz,B., Stern,D., Gibbs,R.A., Richards,S., and Moran,N.A. (2009a). Dynamics of genome evolution in facultative symbionts of aphids. Environ. Microbiol.

Degnan,P.H. and Moran,N.A. (2008b). Diverse phage-encoded toxins in a protective insect endosymbiont. Appl. Environ. Microbiol. *74*, 6782-6791.

Degnan,P.H. and Moran,N.A. (2008a). Evolutionary genetics of a defensive facultative symbiont of insects: exchange of toxin-encoding bacteriophage. Mol. Ecol. *17*, 916-929.

Degnan,P.H., Yu,Y., Sisneros,N., Wing,R.A., and Moran,N.A. (2009b). Hamiltonella defensa, genome evolution of protective bacterial endosymbiont from pathogenic ancestors. Proc. Natl. Acad. Sci. U. S. A *106*, 9063-9068.

Dehal,P. and Boore,J.L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. PLoS. Biol. *3*, e314.

Delmotte,F., Rispe,C., Schaber,J., Silva,F.J., and Moya,A. (2006). Tempo and mode of early gene loss in endosymbiotic bacteria from insects. BMC. Evol. Biol. *6*, 56.

DeLong,E.F. and Pace,N.R. (2001). Environmental diversity of bacteria and archaea. Syst. Biol. *50*, 470-478.

Delsuc,F., Brinkmann,H., and Philippe,H. (2005). Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet *6*, 361-375.

DeMoll,E. (1996). Biosynthesis of biotin and lipoic acid. In Escherichia coli and Salmonella cellular and molecular biology, F.C.Neidhardt, ed. (Washington D.C.: American Society for Microbiology), pp. 704-709.

Deng,W., Burland,V., Plunkett,G., III, Boutin,A., Mayhew,G.F., Liss,P., Perna,N.T., Rose,D.J., Mau,B., Zhou,S., Schwartz,D.C., Fetherston,J.D., Lindler,L.E., Brubaker,R.R., Plano,G.V., Straley,S.C., McDonough,K.A., Nilles,M.L., Matson,J.S., Blattner,F.R., and Perry,R.D. (2002). Genome sequence of Yersinia pestis KIM. J. Bacteriol. *184*, 4601-4611.

Dobrindt,U., Hochhut,B., Hentschel,U., and Hacker,J. (2004). Genomic islands in pathogenic and environmental microorganisms. Nat. Rev. Microbiol. *2*, 414-424.

Dobson,S.L., Bourtzis,K., Braig,H.R., Jones,B.F., Zhou,W., Rousset,F., and O'Neill,S.L. (1999). Wolbachia infections are distributed throughout insect somatic and germ line tissues. Insect Biochem. Mol. Biol. *29*, 153-160.

Bibliography

Doolittle,W.F. (1999a). Lateral genomics. Trends Cell Biol. *9*, M5-M8.

Doolittle,W.F. (1999b). Phylogenetic classification and the universal tree. Science *284*, 2124-2129.

Doolittle,W.F. and Bapteste,E. (2007). Pattern pluralism and the Tree of Life hypothesis. Proc. Natl. Acad. Sci. U. S. A *104*, 2043-2049.

Doolittle,W.F. and Sapienza,C. (1980). Selfish genes, the phenotype paradigm and genome evolution. Nature *284*, 601-603.

Dougherty,K.M. and Plague,G.R. (2008). Transposable element loads in a bacterial symbiont of weevils are extremely variable. Appl. Environ. Microbiol. *74*, 7832-7834.

Drummond,D.A., Bloom,J.D., Adami,C., Wilke,C.O., and Arnold,F.H. (2005). Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. U. S. A *102*, 14338-14343.

Drummond,D.A., Raval,A., and Wilke,C.O. (2006). A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. *23*, 327-337.

Duarte,N.C., Herrgard,M.J., and Palsson,B.O. (2004). Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Res. *14*, 1298-1309.

Dubnau,D. (1999). DNA uptake in bacteria. Annu. Rev. Microbiol. *53*, 217-244.

DuBose,R.F., Dykhuizen,D.E., and Hartl,D.L. (1988). Genetic exchange among natural isolates of bacteria: recombination within the phoA gene of Escherichia coli. Proc. Natl. Acad. Sci U. S. A *85*, 7036-7040.

Duchaud,E., Rusniok,C., Frangeul,L., Buchrieser,C., Givaudan,A., Taourit,S., Bocs,S., Boursaux-Eude,C., Chandler,M., Charles,J.F., Dassa,E., Derose,R., Derzelle,S., Freyssinet,G., Gaudriault,S., Medigue,C., Lanois,A., Powell,K., Siguier,P., Vincent,R., Wingate,V., Zouine,M., Glaser,P., Boemare,N., Danchin,A., and Kunst,F. (2003). The genome sequence of the entomopathogenic bacterium Photorhabdus luminescens. Nat. Biotechnol. *21*, 1307-1313.

Dufresne,A., Garczarek,L., and Partensky,F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol. *6*, R14.

Durot,M., Bourguignon,P.Y., and Schachter,V. (2009). Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS Microbiol. Rev. *33*, 164-190.

Dutilh,B.E., Huynen,M.A., Bruno,W.J., and Snel,B. (2004). The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J. Mol. Evol. *58*, 527-539.

Dykhuizen,D.E. and Green,L. (1991). Recombination in Escherichia coli and the definition of biological species. J. Bacteriol. *173*, 7257-7268.

Edwards,J.S., Covert,M., and Palsson,B. (2002). Metabolic modelling of microbes: the flux-balance approach. Environ. Microbiol. *4*, 133-140.

Edwards,J.S., Ibarra,R.U., and Palsson,B.O. (2001). In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat. Biotechnol. *19*, 125-130.

Edwards,J.S. and Palsson,B.O. (2000a). The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc. Natl. Acad. Sci. U. S. A *97*, 5528-5533.

Edwards,J.S. and Palsson,B.O. (2000b). Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. BMC. Bioinformatics. *1*, 1.

Edwards,J.S. and Palsson,B.O. (2000c). Multiple steady states in kinetic models of red cell metabolism. J. Theor. Biol. *207*, 125-127.

Edwards,J.S. and Palsson,B.O. (2000d). Robustness analysis of the Escherichia coli metabolic network. Biotechnol. Prog. *16*, 927-939.

Edwards,R.A. and Rohwer,F. (2005). Viral metagenomics. Nat. Rev. Microbiol. *3*, 504-510.

Edwards,R.J. and Brookfield,J.F. (2003). Transiently beneficial insertions could maintain mobile DNA sequences in variable environments. Mol. Biol. Evol. *20*, 30-37.

Eikmanns,B.J., Rittmann,D., and Sahm,H. (1995). Cloning, sequence analysis, expression, and inactivation of the Corynebacterium glutamicum icd gene encoding isocitrate dehydrogenase and biochemical characterization of the enzyme. J. Bacteriol. *177*, 774-782.

Eisen,J.A. (2000). Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr. Opin. Genet Dev. *10*, 606-611.

Eisen,J.A., Heidelberg,J.F., White,O., and Salzberg,S.L. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol. *1*, RESEARCH0011.

Elbein,A.D., Pan,Y.T., Pastuszak,I., and Carroll,D. (2003). New insights on trehalose: a multifunctional molecule. Glycobiology *13*, 17R-27R.

Bibliography

Erdos,P. and Renyi,A. (1960). {On the evolution of random graphs}. Publ. Math. Inst. Hung. Acad. Sci *5*, 17-61.

Famili,I. and Palsson,B.O. (2003). The convex basis of the left null space of the stoichiometric matrix leads to the definition of metabolically meaningful pools. Biophys. J. *85*, 16-26.

Fang,G., Rocha,E., and Danchin,A. (2005). How essential are nonessential genes? Mol. Biol. Evol. *22*, 2147-2156.

Feil,E.J. and Spratt,B.G. (2001). Recombination and the population structures of bacterial pathogens. Annu. Rev. Microbiol. *55*, 561-590.

Feist,A.M., Henry,C.S., Reed,J.L., Krummenacker,M., Joyce,A.R., Karp,P.D., Broadbelt,L.J., Hatzimanikatis,V., and Palsson,B.O. (2007). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. *3*, 121.

Feist,A.M., Herrgard,M.J., Thiele,I., Reed,J.L., and Palsson,B.O. (2009). Reconstruction of biochemical networks in microorganisms. Nat. Rev. Microbiol. *7*, 129-143.

Fell,D.A. and Small,J.R. (1986). Fat synthesis in adipose tissue. An examination of stoichiometric constraints. Biochem. J. *238*, 781-786.

Fell,D.A. and Wagner,A. (2000). The small world of metabolism. Nat. Biotechnol. *18*, 1121-1122.

Felsenstein,J. (1974). The evolutionary advantage of recombination. Genetics *78*, 737-756.

Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K., Holm,L., Sonnhammer,E.L., Eddy,S.R., and Bateman,A. (2009). The Pfam protein families database. Nucleic Acids Res.

Fischer,E. and Sauer,U. (2005). Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism. Nat. Genet *37*, 636-640.

Fitch,W.M. and Margoliash,E. (1967). Construction of phylogenetic trees. Science *155*, 279-284.

Fitz-Gibbon,S.T. and House,C.H. (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. *27*, 4218-4222.

Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M., and . (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496-512.

Fong,S.S., Burgard,A.P., Herring,C.D., Knight,E.M., Blattner,F.R., Maranas,C.D., and Palsson,B.O. (2005). In silico design and adaptive evolution of Escherichia coli for production of lactic acid. Biotechnol. Bioeng. *91*, 643-648.

Force,A., Lynch,M., Pickett,F.B., Amores,A., Yan,Y.L., and Postlethwait,J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics *151*, 1531-1545.

Forster,J., Famili,I., Fu,P., Palsson,B.O., and Nielsen,J. (2003a). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res. *13*, 244-253.

Forster,J., Famili,I., Palsson,B.O., and Nielsen,J. (2003b). Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. OMICS. *7*, 193-202.

Forsyth,R.A., Haselbeck,R.J., Ohlsen,K.L., Yamamoto,R.T., Xu,H., Trawick,J.D., Wall,D., Wang,L., Brown-Driver,V., Froelich,J.M., KG,C., King,P., McCarthy,M., Malone,C., Misiner,B., Robbins,D., Tan,Z., Zhu Zy,Z.Y., Carr,G., Mosca,D.A., Zamudio,C., Foulkes,J.G., and Zyskind,J.W. (2002). A genome-wide strategy for the identification of essential genes in Staphylococcus aureus. Mol. Microbiol. *43*, 1387-1400.

Foster,J., Ganatra,M., Kamal,I., Ware,J., Makarova,K., Ivanova,N., Bhattacharyya,A., Kapatral,V., Kumar,S., Posfai,J., Vincze,T., Ingram,J., Moran,L., Lapidus,A., Omelchenko,M., Kyrpides,N., Ghedin,E., Wang,S., Goltsman,E., Joukov,V., Ostrovskaya,O., Tsukerman,K., Mazur,M., Comb,D., Koonin,E., and Slatko,B. (2005). The Wolbachia genome of Brugia malayi: endosymbiont evolution within a human pathogenic nematode. PLoS Biol. *3*, e121.

Foster,J.W., Park,Y.K., Penfound,T., Fenger,T., and Spector,M.P. (1990). Regulation of NAD metabolism in Salmonella typhimurium: molecular sequence analysis of the bifunctional nadR regulator and the nadA-pnuC operon. J. Bacteriol. *172*, 4187-4196.

Fournier,G.P., Huang,J., and Gogarten,J.P. (2009). Horizontal gene transfer from extinct and extant lineages: biological innovation and the coral of life. Philos. Trans. R. Soc. Lond B Biol. Sci. *364*, 2229-2239.

Fox,G.E., Wisotzkey,J.D., and Jurtshuk,P., Jr. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int. J. Syst. Bacteriol. *42*, 166-170.

Fralick,J.A. (1996). Evidence that TolC is required for functioning of the Mar/AcrAB efflux pump of Escherichia coli. J. Bacteriol. *178*, 5803-5805.

Fraser,C.M., Casjens,S., Huang,W.M., Sutton,G.G., Clayton,R., Lathigra,R., White,O., Ketchum,K.A., Dodson,R., Hickey,E.K., Gwinn,M., Dougherty,B., Tomb,J.F., Fleischmann,R.D., Richardson,D., Peterson,J., Kerlavage,A.R.,

Bibliography

Quackenbush,J., Salzberg,S., Hanson,M., van,V.R., Palmer,N., Adams,M.D., Gocayne,J., Weidman,J., Utterback,T., Watthey,L., McDonald,L., Artiach,P., Bowman,C., Garland,S., Fuji,C., Cotton,M.D., Horst,K., Roberts,K., Hatch,B., Smith,H.O., and Venter,J.C. (1997). Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature *390*, 580-586.

Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M., Fritchman,R.D., Weidman,J.F., Small,K.V., Sandusky,M., Fuhrmann,J., Nguyen,D., Utterback,T.R., Saudek,D.M., Phillips,C.A., Merrick,J.M., Tomb,J.F., Dougherty,B.A., Bott,K.F., Hu,P.C., Lucier,T.S., Peterson,S.N., Smith,H.O., Hutchison,C.A., III, and Venter,J.C. (1995). The minimal gene complement of Mycoplasma genitalium. Science *270*, 397-403.

Fraser-Liggett,C.M. (2005). Insights on biology and evolution from microbial genome sequencing. Genome Res. *15*, 1603-1610.

Frey,P.A. (1996). The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose. FASEB J. *10*, 461-470.

Friedrich,A.B., Merkert,H., Fendert,T., Hacker,J., Proksch,P., and Hentschel,U. (1999). Microbial diversity in the marine sponge Aplysina cavernicola (formerly Verongia cavernicola) analyzed by fluorescence in situ hybridization (FISH). Marine biology *134*, 461-470.

Frost,L.S., Leplae,R., Summers,A.O., and Toussaint,A. (2005). Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. *3*, 722-732.

Fukatsu,T., Nikoh,N., Kawai,R., and Koga,R. (2000). The secondary endosymbiotic bacterium of the pea aphid Acyrthosiphon pisum (Insecta: homoptera). Appl. Environ. Microbiol. *66*, 2748-2758.

Funk,D.J., Helbling,L., Wernegreen,J.J., and Moran,N.A. (2000). Intraspecific phylogenetic congruence among multiple symbiont genomes. Proc. Biol. Sci. *267*, 2517-2521.

Fuxelius,H.H., Darby,A., Min,C.K., Cho,N.H., and Andersson,S.G. (2007). The genomic and metabolic diversity of Rickettsia. Res. Microbiol. *158*, 745-753.

Fuxelius,H.H., Darby,A.C., Cho,N.H., and Andersson,S.G. (2008). Visualization of pseudogenes in intracellular bacteria reveals the different tracks to gene destruction. Genome Biol. *9*, R42.

Gabaldon,T., Pereto,J., Montero,F., Gil,R., Latorre,A., and Moya,A. (2007). Structural analyses of a hypothetical minimal metabolism. Philos. Trans. R. Soc. Lond B Biol. Sci. *362*, 1751-1762.

Galan,J.E. (1999). Interaction of Salmonella with host cells through the centisome 63 type III secretion system. Curr. Opin. Microbiol. *2*, 46-50.

Galan,J.E. (2001). Salmonella interactions with host cells: type III secretion at work. Annu. Rev. Cell Dev. Biol. *17*, 53-86.

Galan,J.E. and Wolf-Watz,H. (2006). Protein delivery into eukaryotic cells by type III secretion machines. Nature *444*, 567-573.

Galas,D.J. and Chandler,M. (2009). Bacterial Insertion Sequences. In Mobile DNA, D.E.Berg and M.M.Howe, eds. (Washington DC: American Society for Microbiology), pp. 109-163.

Gan,L., Chen,S., and Jensen,G.J. (2008). Molecular organization of Gram-negative peptidoglycan. Proc. Natl. Acad. Sci U. S. A *105*, 18953-18957.

Garcia-Vallve,S., Guzman,E., Montero,M.A., and Romeu,A. (2003). HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res. *31*, 187-189.

Garcia-Vallve,S., Romeu,A., and Palau,J. (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. *10*, 1719-1725.

Gardy,J.L., Spencer,C., Wang,K., Ester,M., Tusnady,G.E., Simon,I., Hua,S., deFays,K., Lambert,C., Nakai,K., and Brinkman,F.S. (2003). PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res. *31*, 3613-3617.

Geiger,A., Cuny,G., and Frutos,R. (2005). Two Tsetse fly species, Glossina palpalis gambiensis and Glossina morsitans morsitans, carry genetically distinct populations of the secondary symbiont Sodalis glossinidius. Appl. Environ. Microbiol. *71*, 8941-8943.

Geiger,A., Ravel,S., Mateille,T., Janelle,J., Patrel,D., Cuny,G., and Frutos,R. (2007). Vector competence of Glossina palpalis gambiensis for Trypanosoma brucei s.l. and genetic diversity of the symbiont Sodalis glossinidius. Mol. Biol. Evol. *24*, 102-109.

Gerdes,S.Y., Scholle,M.D., Campbell,J.W., Balazsi,G., Ravasz,E., Daugherty,M.D., Somera,A.L., Kyrpides,N.C., Anderson,I., Gelfand,M.S., Bhattacharya,A., Kapatral,V., D'Souza,M., Baev,M.V., Grechkin,Y., Mseeh,F., Fonstein,M.Y., Overbeek,R., Barabasi,A.L., Oltvai,Z.N., and Osterman,A.L. (2003). Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J. Bacteriol. *185*, 5673-5684.

Gerlach,R.G. and Hensel,M. (2007). Protein secretion systems and adhesins: the molecular armory of Gram-negative pathogens. Int. J. Med. Microbiol. *297*, 401-415.

Bibliography

Gevers,D., Cohan,F.M., Lawrence,J.G., Spratt,B.G., Coenye,T., Feil,E.J., Stackebrandt,E., Van de,P.Y., Vandamme,P., Thompson,F.L., and Swings,J. (2005). Opinion: Re-evaluating prokaryotic species. Nat. Rev. Microbiol. *3*, 733-739.

Gevers,D., Vandepoele,K., Simillon,C., and Van de,P.Y. (2004). Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. *12*, 148-154.

Gil,R., Belda,E., Gosalbes,M.J., Delaye,L., Vallier,A., Vincent-Monegat,C., Heddi,A., Silva,F.J., Moya,A., and Latorre,A. (2008). Massive presence of insertion sequences in the genome of SOPE, the primary endosymbiont of the rice weevil Sitophilus oryzae. Int. Microbiol. *11*, 41-48.

Gil,R., Sabater-Munoz,B., Latorre,A., Silva,F.J., and Moya,A. (2002). Extreme genome reduction in Buchnera spp.: toward the minimal genome needed for symbiotic life. Proc. Natl. Acad. Sci. U. S. A *99*, 4454-4458.

Gil,R., Silva,F.J., Pereto,J., and Moya,A. (2004). Determination of the core of a minimal bacterial gene set. Microbiol. Mol. Biol. Rev. *68*, 518-537.

Gil,R., Silva,F.J., Zientz,E., Delmotte,F., Gonzalez-Candelas,F., Latorre,A., Rausell,C., Kamerbeek,J., Gadau,J., Holldobler,B., van Ham,R.C., Gross,R., and Moya,A. (2003). The genome sequence of Blochmannia floridanus: comparative analysis of reduced genomes. Proc. Natl. Acad. Sci. U. S. A *100*, 9388-9393.

Giovannoni,S.J., Tripp,H.J., Givan,S., Podar,M., Vergin,K.L., Baptista,D., Bibbs,L., Eads,J., Richardson,T.H., Noordewier,M., Rappe,M.S., Short,J.M., Carrington,J.C., and Mathur,E.J. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. Science *309*, 1242-1245.

Godoy,D., Randle,G., Simpson,A.J., Aanensen,D.M., Pitt,T.L., Kinoshita,R., and Spratt,B.G. (2003). Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, Burkholderia pseudomallei and Burkholderia mallei. J. Clin. Microbiol. *41*, 2068-2079.

Gogarten,J.P., Doolittle,W.F., and Lawrence,J.G. (2002). Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. *19*, 2226-2238.

Gogarten,J.P. and Townsend,J.P. (2005). Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. *3*, 679-687.

Gomez-Valero,L., Latorre,A., and Silva,F.J. (2004a). The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont Buchnera aphidicola. Mol. Biol. Evol. *21*, 2172-2181.

Gomez-Valero,L., Rocha,E.P., Latorre,A., and Silva,F.J. (2007). Reconstructing the ancestor of Mycobacterium leprae: the dynamics of gene loss and genome reduction. Genome Res. *17*, 1178-1185.

Gomez-Valero,L., Soriano-Navarro,M., Perez-Brocal,V., Heddi,A., Moya,A., Garcia-Verdugo,J.M., and Latorre,A. (2004b). Coexistence of Wolbachia with Buchnera aphidicola and a secondary symbiont in the aphid Cinara cedri. J. Bacteriol. *186*, 6626-6633.

Gooding,R.H. and Krafsur,E.S. (2005). Tsetse genetics: contributions to biology, systematics, and control of tsetse flies. Annu. Rev. Entomol. *50*, 101-123.

Gortz,H.D. and Brigge,T. (1998). Intracellular bacteria in protozoa. Naturwissenschaften *85*, 359-368.

Gosalbes,M.J., Lamelas,A., Moya,A., and Latorre,A. (2008). The striking case of tryptophan provision in the cedar aphid Cinara cedri. J. Bacteriol. *190*, 6026-6029.

Gotz,S., Garcia-Gomez,J.M., Terol,J., Williams,T.D., Nagaraj,S.H., Nueda,M.J., Robles,M., Talon,M., Dopazo,J., and Conesa,A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. *36*, 3420-3435.

Green,J.M., Merkel,W.K., and Nichols,B.P. (1992). Characterization and sequence of Escherichia coli pabC, the gene encoding aminodeoxychorismate lyase, a pyridoxal phosphate-containing enzyme. J. Bacteriol. *174*, 5317-5323.

Green,J.M., Nichols,B.P., and Matthews,R.G. (1996). Folate biosynthesis, reduction, and polyglutamylation. In Escherichia coli and Salmonella cellular and molecular biology, F.C.Neidhardt, ed. (Wahington D.C.: American Society for Microbiology), pp. 665-673.

Greenblatt,J. (1996). Control of RNA chain elongation and termination. In Escherichia coli and Salmonella cellular and molecular biology, F.C.Neidhardt, ed. (Washington D.C.: American Society for Microbiology), pp. 582-848.

Greene,R.C. (1996). Biosynthesis of methionine. In Escherichia coli and Salmonella cellular and molecular biology, F.C.Neidhardt, ed. (Washington D.C.: American Society for Microbiology), pp. 542-560.

Grenier,A.M., Nardon,C., and Nardon,P. (1994). The role of symbiotes in flight activity ofSitophilus weevils. Entomologia Experimentalis et Applicata *70*, 201-208.

Gribaldo,S. and Philippe,H. (2002). Ancient phylogenetic relationships. Theor. Popul. Biol. *61*, 391-408.

Groisman,E.A. and Ochman,H. (1997). How Salmonella became a pathogen. Trends Microbiol. *5*, 343-349.

Gu,X. and Zhang,H. (2004). Genome phylogenetic analysis based on extended gene contents. Mol. Biol. Evol. *21*, 1401-1408.

Bibliography

Guell,M., van,N., V, Yus,E., Chen,W.H., Leigh-Bell,J., Michalodimitrakis,K., Yamada,T., Arumugam,M., Doerks,T., Kuhner,S., Rode,M., Suyama,M., Schmidt,S., Gavin,A.C., Bork,P., and Serrano,L. (2009). Transcriptome complexity in a genome-reduced bacterium. Science *326*, 1268-1271.

Guerdoux-Jamet,P., Henaut,A., Nitschke,P., Risler,J.L., and Danchin,A. (1997). Using codon usage to predict genes origin: is the Escherichia coli outer membrane a patchwork of products from different genomes? DNA Res. *4*, 257-265.

Guttman,D.S. and Dykhuizen,D.E. (1994). Detecting selective sweeps in naturally occurring Escherichia coli. Genetics *138*, 993-1003.

Hacker,J., Bender,L., Ott,M., Wingender,J., Lund,B., Marre,R., and Goebel,W. (1990). Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates. Microb. Pathog. *8*, 213-225.

Hacker,J., Blum-Oehler,G., Muhldorfer,I., and Tschape,H. (1997). Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol. Microbiol. *23*, 1089-1097.

Hacker,J. and Carniel,E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep. *2*, 376-381.

Hacker,J. and Kaper,J.B. (2000). Pathogenicity islands and the evolution of microbes. Annu. Rev. Microbiol. *54*, 641-679.

Hanage,W.P., Fraser,C., and Spratt,B.G. (2006a). Sequences, sequence clusters and bacterial species. Philos. Trans. R. Soc. Lond B Biol. Sci. *361*, 1917-1927.

Hanage,W.P., Fraser,C., and Spratt,B.G. (2006b). The impact of homologous recombination on the generation of diversity in bacteria. J. Theor. Biol. *239*, 210-219.

Hansmann,S. and Martin,W. (2000). Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int. J. Syst. Evol. Microbiol. *50 Pt 4*, 1655-1663.

Hatfull,G.F. (2008). Bacteriophage genomics. Curr. Opin. Microbiol. *11*, 447-453.

Haynes,S., Darby,A.C., Daniell,T.J., Webster,G., Van Veen,F.J., Godfray,H.C., Prosser,J.I., and Douglas,A.E. (2003). Diversity of bacteria associated with natural aphid populations. Appl. Environ. Microbiol. *69*, 7216-7223.

Heddi,A., Charles,H., and Khatchadourian,C. (2001). Intracellular bacterial symbiosis in the genus Sitophilus: the 'biological individual' concept revisited. Res. Microbiol. *152*, 431-437.

Heddi,A., Charles,H., Khatchadourian,C., Bonnot,G., and Nardon,P. (1998). Molecular characterization of the principal symbiotic bacteria of the weevil Sitophilus oryzae: a peculiar G + C content of an endocytobiotic DNA. J. Mol. Evol. *47*, 52-61.

Heddi,A., Grenier,A.M., Khatchadourian,C., Charles,H., and Nardon,P. (1999). Four intracellular genomes direct weevil biology: nuclear, mitochondrial, principal endosymbiont, and Wolbachia. Proc. Natl. Acad. Sci. U. S. A *96*, 6814-6819.

Heddi,A., Lestienne,P., Wallace,D.C., and Stepien,G. (1993). Mitochondrial DNA expression in mitochondrial myopathies and coordinated expression of nuclear genes involved in ATP production. J. Biol. Chem. *268*, 12156-12163.

Hegedus,D., Erlandson,M., Gillott,C., and Toprak,U. (2009). New insights into peritrophic matrix synthesis, architecture, and function. Annu. Rev. Entomol. *54*, 285-302.

Heidelberg,J.F., Eisen,J.A., Nelson,W.C., Clayton,R.A., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Umayam,L., Gill,S.R., Nelson,K.E., Read,T.D., Tettelin,H., Richardson,D., Ermolaeva,M.D., Vamathevan,J., Bass,S., Qin,H., Dragoi,I., Sellers,P., McDonald,L., Utterback,T., Fleishmann,R.D., Nierman,W.C., White,O., Salzberg,S.L., Smith,H.O., Colwell,R.R., Mekalanos,J.J., Venter,J.C., and Fraser,C.M. (2000). DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. Nature *406*, 477-483.

Heinrichs,D.E., Yethon,J.A., and Whitfield,C. (1998). Molecular basis for structural diversity in the core regions of the lipopolysaccharides of Escherichia coli and Salmonella enterica. Mol. Microbiol. *30*, 221-232.

Heller,K.B., Lin,E.C., and Wilson,T.H. (1980). Substrate specificity and transport properties of the glycerol facilitator of Escherichia coli. J. Bacteriol. *144*, 274-278.

Hendrix,R.W. (2002). Bacteriophages: evolution of the majority. Theor. Popul. Biol. *61*, 471-480.

Hendrix,R.W., Lawrence,J.G., Hatfull,G.F., and Casjens,S. (2000). The origins and ongoing evolution of viruses. Trends Microbiol. *8*, 504-508.

Hendrix,R.W., Smith,M.C., Burns,R.N., Ford,M.E., and Hatfull,G.F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc. Natl. Acad. Sci. U. S. A *96*, 2192-2197.

Hentschel,U. and Hacker,J. (2001). Pathogenicity islands: the tip of the iceberg. Microbes. Infect. *3*, 545-548.

Bibliography

Hentschel,U., Steinert,M., and Hacker,J. (2000). Common molecular mechanisms of symbiosis and pathogenesis. Trends Microbiol. *8*, 226-231.

Herbeck,J.T., Degnan,P.H., and Wernegreen,J.J. (2005). Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (gamma-Proteobacteria). Mol. Biol. Evol. *22*, 520-532.

Herbeck,J.T., Wall,D.P., and Wernegreen,J.J. (2003). Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont Wigglesworthia. Microbiology *149*, 2585-2596.

Hill,R.E. and Spenser,I.D. (1996). Biosynthesis of vitamin B6. In Escherichia coli and Salmonella cellular and molecular biology, F.C.Neidhardt, ed. (Washington D.C.: American Society for Microbiology), pp. 695-703.

Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.C., and Herrmann,R. (1996). Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res. *24*, 4420-4449.

Hiramatsu,K., Katayama,Y., Yuzawa,H., and Ito,T. (2002). Molecular genetics of methicillin-resistant Staphylococcus aureus. Int. J. Med. Microbiol. *292*, 67-74.

Hirsh,A.E. and Fraser,H.B. (2001). Protein dispensability and rate of evolution. Nature *411*, 1046-1049.

Hochhut,B., Jahreis,K., Lengeler,J.W., and Schmid,K. (1997). CTnscr94, a conjugative transposon found in enterobacteria. J. Bacteriol. *179*, 2097-2102.

Holmes,D.E., Nevin,K.P., and Lovley,D.R. (2004). Comparison of 16S rRNA, nifD, recA, gyrB, rpoB and fusA genes within the family Geobacteraceae fam. nov. Int. J. Syst. Evol. Microbiol. *54*, 1591-1599.

Holt,K.E., Thomson,N.R., Wain,J., Langridge,G.C., Hasan,R., Bhutta,Z.A., Quail,M.A., Norbertczak,H., Walker,D., Simmonds,M., White,B., Bason,N., Mungall,K., Dougan,G., and Parkhill,J. (2009). Pseudogene accumulation in the evolutionary histories of Salmonella enterica serovars Paratyphi A and Typhi. BMC. Genomics *10*, 36.

Holtje,J.V. (1998). Growth of the stress-bearing and shape-maintaining murein sacculus of Escherichia coli. Microbiol. Mol. Biol. Rev. *62*, 181-203.

Homma,K., Fukuchi,S., Kawabata,T., Ota,M., and Nishikawa,K. (2002). A systematic investigation identifies a significant number of probable pseudogenes in the Escherichia coli genome. Gene *294*, 25-33.

Hooper,S.D. and Berg,O.G. (2003b). On the nature of gene innovation: duplication patterns in microbial genomes. Mol. Biol. Evol. *20*, 945-954.

Hooper,S.D. and Berg,O.G. (2003a). Duplication is more common among laterally transferred genes than among indigenous genes. Genome Biol. *4*, R48.

Horn,M. (2008). Chlamydiae as symbionts in eukaryotes. Annu. Rev. Microbiol. *62*, 113-131.

Horn,M. and Wagner,M. (2004). Bacterial endosymbionts of free-living amoebae. J. Eukaryot. Microbiol. *51*, 509-514.

Hosokawa,T., Koga,R., Kikuchi,Y., Meng,X.Y., and Fukatsu,T. (2010). Wolbachia as a bacteriocyte-associated nutritional mutualist. Proc. Natl. Acad. Sci U. S. A *107*, 769-774.

Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J.C., Kitano,H., Arkin,A.P., Bornstein,B.J., Bray,D., Cornish-Bowden,A., Cuellar,A.A., Dronov,S., Gilles,E.D., Ginkel,M., Gor,V., Goryanin,I.I., Hedley,W.J., Hodgman,T.C., Hofmeyr,J.H., Hunter,P.J., Juty,N.S., Kasberger,J.L., Kremling,A., Kummer,U., Le,N.N., Loew,L.M., Lucio,D., Mendes,P., Minch,E., Mjolsness,E.D., Nakayama,Y., Nelson,M.R., Nielsen,P.F., Sakurada,T., Schaff,J.C., Shapiro,B.E., Shimizu,T.S., Spence,H.D., Stelling,J., Takahashi,K., Tomita,M., Wagner,J., and Wang,J. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. *19*, 524-531.

Hueck,C.J. (1998). Type III protein secretion systems in bacterial pathogens of animals and plants. Microbiol. Mol. Biol. Rev. *62*, 379-433.

Hugenholtz,P., Goebel,B.M., and Pace,N.R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. *180*, 4765-4774.

Hughes,D. (2000). Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. Genome Biol. *1*, REVIEWS0006.

Hughes,J.A. (2006). In vivo hydrolysis of S-adenosyl-L-methionine in Escherichia coli increases export of 5-methylthioribose. Can. J. Microbiol. *52*, 599-602.

Hunter,M.S., Perlman,S.J., and Kelly,S.E. (2003). A bacterial symbiont in the Bacteroidetes induces cytoplasmic incompatibility in the parasitoid wasp Encarsia pergandiella. Proc. Biol. Sci. *270*, 2185-2190.

Hurst,L.D. and Smith,N.G. (1999). Do essential genes evolve slowly? Curr. Biol. *9*, 747-750.

Huson,D.H. and Steel,M. (2004). Phylogenetic trees based on gene content. Bioinformatics. *20*, 2044-2049.

Bibliography

Hutchison,C.A., Peterson,S.N., Gill,S.R., Cline,R.T., White,O., Fraser,C.M., Smith,H.O., and Venter,J.C. (1999). Global transposon mutagenesis and a minimal Mycoplasma genome. Science *286*, 2165-2169.

Huynen,M.A. and Bork,P. (1998). Measuring genome evolution. Proc. Natl. Acad. Sci. U. S. A *95*, 5849-5856.

Huynen,M.A. and Snel,B. (2000). Gene and context: integrative approaches to genome analysis. Adv. Protein Chem. *54*, 345-379.

Ibarra,R.U., Edwards,J.S., and Palsson,B.O. (2002). Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature *420*, 186-189.

Inagaki,Y., Susko,E., Fast,N.M., and Roger,A.J. (2004). Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1alpha phylogenies. Mol. Biol. Evol. *21*, 1340-1349.

Itikawa,H., Baumberg,S., and Vogel,H.J. (1968). Enzymic basis for a genetic supression:accumulation and deacylation of N-acetylglutamic gamma-semialdehyde in enterobacterial mutants. Biochim. Biophys. Acta *159*, 547-550.

Ito,T., Katayama,Y., Asada,K., Mori,N., Tsutsumimoto,K., Tiensasitorn,C., and Hiramatsu,K. (2001). Structural comparison of three types of staphylococcal cassette chromosome mec integrated in the chromosome in methicillin-resistant Staphylococcus aureus. Antimicrob. Agents Chemother. *45*, 1323-1336.

Itoh,T., Martin,W., and Nei,M. (2002). Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. Proc. Natl. Acad. Sci. U. S. A *99*, 12944-12948.

Iyer,L.M., Koonin,E.V., and Aravind,L. (2004). Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. Gene *335*, 73-88.

Jackowski,S. and Rock,C.O. (1981). Regulation of coenzyme A biosynthesis. J. Bacteriol. *148*, 926-932.

Jacob,F. and Monod,J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. *3*, 318-356.

Jacobs,M.A., Alwood,A., Thaipisuttikul,I., Spencer,D., Haugen,E., Ernst,S., Will,O., Kaul,R., Raymond,C., Levy,R., Chun-Rong,L., Guenthner,D., Bovee,D., Olson,M.V., and Manoil,C. (2003). Comprehensive transposon mutant library of Pseudomonas aeruginosa. Proc. Natl. Acad. Sci U. S. A *100*, 14339-14344.

Jain,R., Rivera,M.C., and Lake,J.A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. U. S. A *96*, 3801-3806.

Jain,R., Rivera,M.C., Moore,J.E., and Lake,J.A. (2002). Horizontal gene transfer in microbial genome evolution. Theor. Popul. Biol. *61*, 489-495.

Jeong,H., Mason,S.P., Barabasi,A.L., and Oltvai,Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41-42.

Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N., and Barabasi,A.L. (2000). The large-scale organization of metabolic networks. Nature *407*, 651-654.

Jiggins,F.M. (2002). The rate of recombination in Wolbachia bacteria. Mol. Biol. Evol. *19*, 1640-1643.

Jiggins,F.M., Hurst,G.D., Jiggins,C.D., Schulenburg,J.H., and Majerus,M.E. (2000). The butterfly Danaus chrysippus is infected by a male-killing Spiroplasma bacterium. Parasitology *120 (Pt 5)*, 439-446.

Jin,Q., Yuan,Z., Xu,J., Wang,Y., Shen,Y., Lu,W., Wang,J., Liu,H., Yang,J., Yang,F., Zhang,X., Zhang,J., Yang,G., Wu,H., Qu,D., Dong,J., Sun,L., Xue,Y., Zhao,A., Gao,Y., Zhu,J., Kan,B., Ding,K., Chen,S., Cheng,H., Yao,Z., He,B., Chen,R., Ma,D., Qiang,B., Wen,Y., Hou,Y., and Yu,J. (2002). Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. Nucleic Acids Res. *30*, 4432-4441.

Jordan,I.K., Makarova,K.S., Spouge,J.L., Wolf,Y.I., and Koonin,E.V. (2001). Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Res. *11*, 555-565.

Jordan,I.K., Rogozin,I.B., Wolf,Y.I., and Koonin,E.V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. *12*, 962-968.

Joyce,A.R. and Palsson,B.O. (2008). Predicting gene essentiality using genome-scale in silico models. Methods Mol. Biol. *416*, 433-457.

Joyce,A.R., Reed,J.L., White,A., Edwards,R., Osterman,A., Baba,T., Mori,H., Lesely,S.A., Palsson,B.O., and Agarwalla,S. (2006). Experimental and computational assessment of conditionally essential genes in Escherichia coli. J. Bacteriol. *188*, 8259-8271.

Juhala,R.J., Ford,M.E., Duda,R.L., Youlton,A., Hatfull,G.F., and Hendrix,R.W. (2000). Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J. Mol. Biol. *299*, 27-51.

Jurgenson,C.T., Begley,T.P., and Ealick,S.E. (2009). The structural and biochemical foundations of thiamin biosynthesis. Annu. Rev. Biochem. *78*, 569-603.

Bibliography

Kambampati,R. and Lauhon,C.T. (1999). IscS is a sulfurtransferase for the in vitro biosynthesis of 4-thiouridine in Escherichia coli tRNA. Biochemistry *38*, 16561-16568.

Kambampati,R. and Lauhon,C.T. (2000). Evidence for the transfer of sulfane sulfur from IscS to ThiI during the in vitro biosynthesis of 4-thiouridine in Escherichia coli tRNA. J. Biol. Chem. *275*, 10727-10730.

Kanamori,T., Kanou,N., Atomi,H., and Imanaka,T. (2004). Enzymatic characterization of a prokaryotic urea carboxylase. J. Bacteriol. *186*, 2532-2539.

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y., and Hattori,M. (2004). The KEGG resource for deciphering the genome. Nucleic Acids Res. *32*, D277-D280.

Kaneko,T., Nakamura,Y., Sato,S., Asamizu,E., Kato,T., Sasamoto,S., Watanabe,A., Idesawa,K., Ishikawa,A., Kawashima,K., Kimura,T., Kishida,Y., Kiyokawa,C., Kohara,M., Matsumoto,M., Matsuno,A., Mochizuki,Y., Nakayama,S., Nakazaki,N., Shimpo,S., Sugimoto,M., Takeuchi,C., Yamada,M., and Tabata,S. (2000). Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti. DNA Res. *7*, 331-338.

Karaolis,D.K., Somara,S., Maneval,D.R., Jr., Johnson,J.A., and Kaper,J.B. (1999). A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. Nature *399*, 375-379.

Karlin,S. and Burge,C. (1995). Dinucleotide relative abundance extremes: a genomic signature. Trends Genet *11*, 283-290.

Kauffman,K.J., Prakash,P., and Edwards,J.S. (2003). Advances in flux balance analysis. Curr. Opin. Biotechnol. *14*, 491-496.

Kennedy,J. and Kealey,J.T. (2004). Tools for metabolic engineering in Escherichia coli: inactivation of panD by a point mutation. Anal. Biochem. *327*, 91-96.

Keseler,I.M., Bonavides-Martinez,C., Collado-Vides,J., Gama-Castro,S., Gunsalus,R.P., Johnson,D.A., Krummenacker,M., Nolan,L.M., Paley,S., Paulsen,I.T., Peralta-Gil,M., Santos-Zavaleta,A., Shearer,A.G., and Karp,P.D. (2009). EcoCyc: a comprehensive view of Escherichia coli biology. Nucleic Acids Res. *37*, D464-D470.

Kettler,G.C., Martiny,A.C., Huang,K., Zucker,J., Coleman,M.L., Rodrigue,S., Chen,F., Lapidus,A., Ferriera,S., Johnson,J., Steglich,C., Church,G.M., Richardson,P., and Chisholm,S.W. (2007). Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. PLoS Genet *3*, e231.

Kimura,M. and Ota,T. (1974). On some principles governing molecular evolution. Proc. Natl. Acad. Sci U. S. A *71*, 2848-2852.

Klamt,S. and Stelling,J. (2003). Two approaches for metabolic pathway analysis? Trends Biotechnol. *21*, 64-69.

Klappenbach,J.A., Dunbar,J.M., and Schmidt,T.M. (2000). rRNA operon copy number reflects ecological strategies of bacteria. Appl. Environ. Microbiol. *66*, 1328-1333.

Klasson,L. and Andersson,S.G. (2004). Evolution of minimal-gene-sets in host-dependent bacteria. Trends Microbiol. *12*, 37-43.

Klasson,L., Westberg,J., Sapountzis,P., Naslund,K., Lutnaes,Y., Darby,A.C., Veneti,Z., Chen,L., Braig,H.R., Garrett,R., Bourtzis,K., and Andersson,S.G. (2009). The mosaic genome structure of the Wolbachia wRi strain infecting Drosophila simulans. Proc. Natl. Acad. Sci. U. S. A.

Kleckner,N., Chan,R.K., Tye,B.K., and Botstein,D. (1975). Mutagenesis by insertion of a drug-resistance element carrying an inverted repetition. J. Mol. Biol. *97*, 561-575.

Knapp,S., Hacker,J., Jarchau,T., and Goebel,W. (1986). Large, unstable inserts in the chromosome affect virulence properties of uropathogenic Escherichia coli O6 strain 536. J. Bacteriol. *168*, 22-30.

Kobayashi,K., Ehrlich,S.D., Albertini,A., Amati,G., Andersen,K.K., Arnaud,M., Asai,K., Ashikaga,S., Aymerich,S., Bessieres,P., Boland,F., Brignell,S.C., Bron,S., Bunai,K., Chapuis,J., Christiansen,L.C., Danchin,A., Debarbouille,M., Dervyn,E., Deuerling,E., Devine,K., Devine,S.K., Dreesen,O., Errington,J., Fillinger,S., Foster,S.J., Fujita,Y., Galizzi,A., Gardan,R., Eschevins,C., Fukushima,T., Haga,K., Harwood,C.R., Hecker,M., Hosoya,D., Hullo,M.F., Kakeshita,H., Karamata,D., Kasahara,Y., Kawamura,F., Koga,K., Koski,P., Kuwana,R., Imamura,D., Ishimaru,M., Ishikawa,S., Ishio,I., Le,C.D., Masson,A., Mauel,C., Meima,R., Mellado,R.P., Moir,A., Moriya,S., Nagakawa,E., Nanamiya,H., Nakai,S., Nygaard,P., Ogura,M., Ohanan,T., O'Reilly,M., O'Rourke,M., Pragai,Z., Pooley,H.M., Rapoport,G., Rawlins,J.P., Rivas,L.A., Rivolta,C., Sadaie,A., Sadaie,Y., Sarvas,M., Sato,T., Saxild,H.H., Scanlan,E., Schumann,W., Seegers,J.F., Sekiguchi,J., Sekowska,A., Seror,S.J., Simon,M., Stragier,P., Studer,R., Takamatsu,H., Tanaka,T., Takeuchi,M., Thomaides,H.B., Vagner,V., van Dijl,J.M., Watabe,K., Wipat,A., Yamamoto,H., Yamamoto,M., Yamamoto,Y., Yamane,K., Yata,K., Yoshida,K., Yoshikawa,H., Zuber,U., and Ogasawara,N. (2003). Essential Bacillus subtilis genes. Proc. Natl. Acad. Sci. U. S. A *100*, 4678-4683.

Kobayashi,N., Nishino,K., and Yamaguchi,A. (2001). Novel macrolide-specific ABC-type efflux transporter in Escherichia coli. J. Bacteriol. *183*, 5639-5644.

Koch,A.L. (1981). Evolution of antibiotic resistance gene function. Microbiol. Rev. *45*, 355-378.

Bibliography

Kolaczkowski,B. and Thornton,J.W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature *431*, 980-984.

Komaki,K. and Ishikawa,H. (1999). Intracellular bacterial symbionts of aphids possess many genomic copies per bacterium. J. Mol. Evol. *48*, 717-722.

Komaki,K. and Ishikawa,H. (2000). Genomic copy number of intracellular bacterial symbionts of aphids varies in response to developmental stage and morph of their host. Insect Biochem. Mol. Biol. *30*, 253-258.

Kondrashov,F.A., Rogozin,I.B., Wolf,Y.I., and Koonin,E.V. (2002). Selection in the evolution of gene duplications. Genome Biol. *3*, RESEARCH0008.

Konstantinidis,K.T., Ramette,A., and Tiedje,J.M. (2006). The bacterial species definition in the genomic era. Philos. Trans. R. Soc. Lond B Biol. Sci. *361*, 1929-1940.

Konstantinidis,K.T. and Tiedje,J.M. (2005a). Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U. S. A *102*, 2567-2572.

Konstantinidis,K.T. and Tiedje,J.M. (2005b). Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. *187*, 6258-6264.

Konstantinidis,K.T. and Tiedje,J.M. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. Proc. Natl. Acad. Sci. U. S. A *101*, 3160-3165.

Koonin,E.V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. *39*, 309-338.

Koonin,E.V. (2000). How many genes can make a cell: the minimal-gene-set concept. Annu. Rev. Genomics Hum. Genet *1*, 99-116.

Koonin,E.V. and Galperin,M.Y. (1997). Prokaryotic genomes: the emerging paradigm of genome-based microbiology. Curr. Opin. Genet Dev. *7*, 757-763.

Koonin,E.V., Mushegian,A.R., and Bork,P. (1996). Non-orthologous gene displacement. Trends Genet *12*, 334-336.

Koonin,E.V., Tatusov,R.L., and Rudd,K.E. (1995). Sequence similarity analysis of Escherichia coli proteins: functional and evolutionary implications. Proc. Natl. Acad. Sci. U. S. A *92*, 11921-11925.

Koonin,E.V. and Wolf,Y.I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. *36*, 6688-6719.

Korbel,J.O., Jensen,L.J., von,M.C., and Bork,P. (2004). Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat. Biotechnol. *22*, 911-917.

Korbel,J.O., Snel,B., Huynen,M.A., and Bork,P. (2002). SHOT: a web server for the construction of genome phylogenies. Trends Genet. *18*, 158-162.

Kornberg,H.L. (2001). Routes for fructose utilization by Escherichia coli. J. Mol. Microbiol. Biotechnol. *3*, 355-359.

Kresse,A.U., Dinesh,S.D., Larbig,K., and Romling,U. (2003). Impact of large chromosomal inversions on the adaptation and evolution of Pseudomonas aeruginosa chronically colonizing cystic fibrosis lungs. Mol. Microbiol. *47*, 145-158.

Kubi,C., van den,A.J., DE,D.R., Marcotty,T., Dorny,P., and van den,B.P. (2006). The effect of starvation on the susceptibility of teneral and non-teneral tsetse flies to trypanosome infection. Med. Vet. Entomol. *20*, 388-392.

Kuepfer,L., Sauer,U., and Blank,L.M. (2005). Metabolic functions of duplicate genes in Saccharomyces cerevisiae. Genome Res. *15*, 1421-1430.

Kuhle,V. and Hensel,M. (2004). Cellular microbiology of intracellular Salmonella enterica: functions of the type III secretion system encoded by Salmonella pathogenicity island 2. Cell Mol. Life Sci. *61*, 2812-2826.

Kuhner,S., van,N., V, Betts,M.J., Leo-Macias,A., Batisse,C., Rode,M., Yamada,T., Maier,T., Bader,S., Beltran-Alvarez,P., Castano-Diez,D., Chen,W.H., Devos,D., Guell,M., Norambuena,T., Racke,I., Rybin,V., Schmidt,A., Yus,E., Aebersold,R., Herrmann,R., Bottcher,B., Frangakis,A.S., Russell,R.B., Serrano,L., Bork,P., and Gavin,A.C. (2009). Proteome organization in a genome-reduced bacterium. Science *326*, 1235-1240.

Kunin,V., Goldovsky,L., Darzentas,N., and Ouzounis,C.A. (2005). The net of life: reconstructing the microbial phylogenetic network. Genome Res. *15*, 954-959.

Kunin,V. and Ouzounis,C.A. (2003a). GeneTRACE-reconstruction of gene content of ancestral species. Bioinformatics. *19*, 1412-1416.

Kunin,V. and Ouzounis,C.A. (2003b). The balance of driving forces during genome evolution in prokaryotes. Genome Res. *13*, 1589-1594.

Kuo,C.H. and Ochman,H. (2009). The fate of new bacterial genes. FEMS Microbiol. Rev. *33*, 38-43.

Kuo,T.T. and Stocker,B.A. (1969). Suppression of proline requirement of proA and proAB deletion mutants in Salmonella typhimurium by mutation to arginine requirement. J. Bacteriol. *98*, 593-598.

Bibliography

Kurland,C.G. (2000). Something for everyone. Horizontal gene transfer in evolution. EMBO Rep. *1*, 92-95.

Kurland,C.G., Canback,B., and Berg,O.G. (2003). Horizontal gene transfer: a critical view. Proc. Natl. Acad. Sci. U. S. A *100*, 9658-9662.

Kurnasov,O.V., Polanuyer,B.M., Ananta,S., Sloutsky,R., Tam,A., Gerdes,S.Y., and Osterman,A.L. (2002). Ribosylnicotinamide kinase domain of NadR protein: identification and implications in NAD biosynthesis. J. Bacteriol. *184*, 6906-6917.

Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C., and Salzberg,S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol. *5*, R12.

Kyrpides,N., Overbeek,R., and Ouzounis,C. (1999). Universal protein families and the functional content of the last universal common ancestor. J. Mol. Evol. *49*, 413-423.

Labedan,B. and Riley,M. (1995). Widespread protein sequence similarities: origins of Escherichia coli genes. J. Bacteriol. *177*, 1585-1588.

Lafay,B., Lloyd,A.T., McLean,M.J., Devine,K.M., Sharp,P.M., and Wolfe,K.H. (1999). Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res. *27*, 1642-1649.

Lai,C.Y., Baumann,P., and Moran,N. (1996). The endosymbiont (Buchnera sp.) of the aphid Diuraphis noxia contains plasmids consisting of trpEG and tandem repeats of trpEG pseudogenes. Appl. Environ. Microbiol. *62*, 332-339.

Lapierre,P. (2008). Dynamics of Prokaryotic Genome Evolution. In Computational methods for understanding bacterial and archaeal genomes, J.Xu and J.P.Gogarten, eds. (London: Imperial College Press), pp. 99-112.

Lapierre,P. and Gogarten,J.P. (2009). Estimating the size of the bacterial pan-genome. Trends Genet *25*, 107-110.

Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R., Thompson,J.D., Gibson,T.J., and Higgins,D.G. (2007). Clustal W and Clustal X version 2.0. Bioinformatics. *23*, 2947-2948.

Lathe,W.C., III, Snel,B., and Bork,P. (2000). Gene context conservation of a higher order than operons. Trends Biochem. Sci. *25*, 474-479.

Lauhon,C.T. (2002). Requirement for IscS in biosynthesis of all thionucleosides in Escherichia coli. J. Bacteriol. *184*, 6820-6829.

Lauhon,C.T. and Kambampati,R. (2000). The iscS gene in Escherichia coli is required for the biosynthesis of 4-thiouridine, thiamin, and NAD. J. Biol. Chem. *275*, 20096-20103.

Lawrence,J.G. (1999). Gene transfer, speciation, and the evolution of bacterial genomes. Curr. Opin. Microbiol. *2*, 519-523.

Lawrence,J.G. (2003). Gene organization: selection, selfishness, and serendipity. Annu. Rev. Microbiol. *57*, 419-440.

Lawrence,J.G. (2005). Common themes in the genome strategies of pathogens. Curr. Opin. Genet Dev. *15*, 584-588.

Lawrence,J.G. and Hendrickson,H. (2003). Lateral gene transfer: when will adolescence end? Mol. Microbiol. *50*, 739-749.

Lawrence,J.G. and Hendrickson,H. (2005). Genome evolution in bacteria: order beneath chaos. Curr. Opin. Microbiol. *8*, 572-578.

Lawrence,J.G., Hendrix,R.W., and Casjens,S. (2001). Where are the pseudogenes in bacterial genomes? Trends Microbiol. *9*, 535-540.

Lawrence,J.G. and Ochman,H. (1997). Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. *44*, 383-397.

Lawrence,J.G. and Ochman,H. (1998). Molecular archaeology of the Escherichia coli genome. Proc. Natl. Acad. Sci. U. S. A *95*, 9413-9417.

Lawrence,J.G., Ochman,H., and Hartl,D.L. (1992). The evolution of insertion sequences within enteric bacteria. Genetics *131*, 9-20.

Lawrence,J.G. and Roth,J.R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics *143*, 1843-1860.

Lazcano,A. and Forterre,P. (1999). The molecular search for the last common ancestor. J. Mol. Evol. *49*, 411-412.

Lecointre,G., Philippe,H., Van Le,H.L., and Le,G.H. (1993). Species sampling has a major impact on phylogenetic inference. Mol. Phylogenet. Evol. *2*, 205-224.

Ledwidge,R. and Blanchard,J.S. (1999). The dual biosynthetic capability of N-acetylornithine aminotransferase in arginine and lysine biosynthesis. Biochemistry *38*, 3019-3024.

Lee,B.M., Park,Y.J., Park,D.S., Kang,H.W., Kim,J.G., Song,E.S., Park,I.C., Yoon,U.H., Hahn,J.H., Koo,B.S., Lee,G.B., Kim,H., Park,H.S., Yoon,K.O., Kim,J.H., Jung,C.H., Koh,N.H., Seo,J.S., and Go,S.J. (2005). The genome sequence

of Xanthomonas oryzae pathovar oryzae KACC10331, the bacterial blight pathogen of rice. Nucleic Acids Res. *33*, 577-586.

Lee,D.S., Burd,H., Liu,J., Almaas,E., Wiest,O., Barabasi,A.L., Oltvai,Z.N., and Kapatral,V. (2009). Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple Staphylococcus aureus genomes identify novel antimicrobial drug targets. J. Bacteriol. *191*, 4015-4024.

Lee,J.M., Gianchandani,E.P., and Papin,J.A. (2006). Flux balance analysis in the era of metabolomics. Brief. Bioinform. *7*, 140-150.

Lee,M.H., Mulrooney,S.B., Renner,M.J., Markowicz,Y., and Hausinger,R.P. (1992). Klebsiella aerogenes urease gene cluster: sequence of ureD and demonstration that four accessory genes (ureD, ureE, ureF, and ureG) are involved in nickel metallocenter biosynthesis. J. Bacteriol. *174*, 4324-4330.

Lefebure,T. and Stanhope,M.J. (2007). Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome Biol. *8*, R71.

Lehane,M.J., Allingham,P.G., and Weglicki,P. (1996). Composition of the peritrophic matrix of the tsetse fly, Glossina morsitans morsitans. Cell Tissue Res. *283*, 375-384.

Lehmann,C., Begley,T.P., and Ealick,S.E. (2006). Structure of the Escherichia coli ThiS-ThiF complex, a key component of the sulfur transfer system in thiamin biosynthesis. Biochemistry *45*, 11-19.

Leisinger,T. (1987). Biosynthesis of proline. In Escherichia coli and Salmonella typhimurium cellular and molecular biology, F.C.Neidhardt, ed. (Washington, D.C.: American Society for Microbiology), pp. 345-351.

Lemonnier,M. and Lane,D. (1998). Expression of the second lysine decarboxylase gene of Escherichia coli. Microbiology *144 (Pt 3)*, 751-760.

Leonardi,R., Fairhurst,S.A., Kriek,M., Lowe,D.J., and Roach,P.L. (2003). Thiamine biosynthesis in Escherichia coli: isolation and initial characterisation of the ThiGH complex. FEBS Lett. *539*, 95-99.

Lerat,E., Daubin,V., and Moran,N.A. (2003). From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS. Biol. *1*, E19.

Lerat,E., Daubin,V., Ochman,H., and Moran,N.A. (2005). Evolutionary origins of genomic repertoires in bacteria. PLoS. Biol. *3*, e130.

Lerat,E. and Ochman,H. (2004). Psi-Phi: exploring the outer limits of bacterial pseudogenes. Genome Res. *14*, 2273-2278.

Lerat,E. and Ochman,H. (2005). Recognizing the pseudogenes in bacterial genomes. Nucleic Acids Res. *33*, 3125-3132.

Lercher,M.J. and Pal,C. (2008). Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol. Biol. Evol. *25*, 559-567.

Levin,B.R. (1981). Periodic selection, infectious gene exchange and the genetic structure of E. coli populations. Genetics *99*, 1-23.

Li,L., Stoeckert,C.J., Jr., and Roos,D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178-2189.

Liu,Y., Harrison,P.M., Kunin,V., and Gerstein,M. (2004). Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. Genome Biol. *5*, R64.

Lloubes,R., Cascales,E., Walburger,A., Bouveret,E., Lazdunski,C., Bernadac,A., and Journet,L. (2001). The Tol-Pal proteins of the Escherichia coli cell envelope: an energized system required for outer membrane integrity? Res. Microbiol. *152*, 523-529.

Luo,H., Shi,J., Arndt,W., Tang,J., and Friedman,R. (2008). Gene order phylogeny of the genus Prochlorococcus. PLoS. ONE. *3*, e3837.

Luscombe,N.M., Babu,M.M., Yu,H., Snyder,M., Teichmann,S.A., and Gerstein,M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. Nature *431*, 308-312.

Lynch,M. (2006). Streamlining and simplification of microbial genome architecture. Annu. Rev. Microbiol. *60*, 327-349.

Lynch,M. and Conery,J.S. (2004). Response to Comment on "The Origins of Genome Complexity". Science 978b.

Lynch,M. and Conery,J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151-1155.

Lynch,M. and Conery,J.S. (2003a). The evolutionary demography of duplicate genes. J. Struct. Funct. Genomics *3*, 35-44.

Lynch,M. and Conery,J.S. (2003b). The origins of genome complexity. Science *302*, 1401-1404.

Lynch,M. and Force,A. (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics *154*, 459-473.

Bibliography

Lynch,M. and Katju,V. (2004). The altered evolutionary trajectories of gene duplicates. Trends Genet. *20*, 544-549.

Lynch,M., Koskella,B., and Schaack,S. (2006). Mutation pressure and the evolution of organelle genomic architecture. Science *311*, 1727-1730.

Ma,H.W., Kumar,B., Ditges,U., Gunzer,F., Buer,J., and Zeng,A.P. (2004). An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. Nucleic Acids Res. *32*, 6643-6649.

Mahadevan,R., Bond,D.R., Butler,J.E., Esteve-Nunez,A., Coppi,M.V., Palsson,B.O., Schilling,C.H., and Lovley,D.R. (2006). Characterization of metabolism in the Fe(III)-reducing organism Geobacter sulfurreducens by constraint-based modeling. Appl. Environ. Microbiol. *72*, 1558-1568.

Mahadevan,R. and Schilling,C.H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng *5*, 264-276.

Mahillon,J. and Chandler,M. (1998). Insertion sequences. Microbiol. Mol. Biol. Rev. *62*, 725-774.

Maiden,M.C. (2006). Multilocus sequence typing of bacteria. Annu. Rev. Microbiol. *60*, 561-588.

Majewski,J. and Cohan,F.M. (1998). The effect of mismatch repair and heteroduplex formation on sexual isolation in Bacillus. Genetics *148*, 13-18.

Makarova,K.S., Aravind,L., Galperin,M.Y., Grishin,N.V., Tatusov,R.L., Wolf,Y.I., and Koonin,E.V. (1999). Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res. *9*, 608-628.

March,J.C., Eiteman,M.A., and Altman,E. (2002). Expression of an anaplerotic enzyme, pyruvate carboxylase, improves recombinant protein production in Escherichia coli. Appl. Environ. Microbiol. *68*, 5620-5624.

Marchetti,M., Capela,D., Glew,M., Cruveiller,S., Chane-Woon-Ming,B., Gris,C., Timmers,T., Poinsot,V., Gilbert,L.B., Heeb,P., Medigue,C., Batut,J., and Masson-Boivin,C. (2010). Experimental evolution of a plant pathogen into a legume symbiont. PLoS Biol. *8*, e1000280.

Mardis,E.R. (2008b). Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet *9*, 387-402.

Mardis,E.R. (2008a). The impact of next-generation sequencing technology on genetics. Trends Genet *24*, 133-141.

Margulis,L. (1970). Origin of eukaryotic cells. (New Haven: Yale University Press).

Margulis,L. (1981). Symbiosis in cell evolution. (San Francisco: Freeman).

Matthews,P.R. and Stewart,P.R. (1988). Amplification of a section of chromosomal DNA in methicillin-resistant Staphylococcus aureus following growth in high concentrations of methicillin. J. Gen. Microbiol. *134*, 1455-1464.

Maudlin,I. and Welburn,S.C. (1987). Lectin mediated establishment of midgut infections of Trypanosoma congolense and Trypanosoma brucei in Glossina morsitans. Trop. Med. Parasitol. *38*, 167-170.

McCutcheon,J.P., McDonald,B.R., and Moran,N.A. (2009). Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. PLoS. Genet. *5*, e1000565.

McCutcheon,J.P. and Moran,N.A. (2007). Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc. Natl. Acad. Sci. U. S. A *104*, 19392-19397.

Medigue,C. and Moszer,I. (2007). Annotation, comparison and databases for hundreds of bacterial genomes. Res. Microbiol. *158*, 724-736.

Medigue,C., Rouxel,T., Vigier,P., Henaut,A., and Danchin,A. (1991). Evidence for horizontal gene transfer in Escherichia coli speciation. J. Mol. Biol. *222*, 851-856.

Medini,D., Donati,C., Tettelin,H., Masignani,V., and Rappuoli,R. (2005). The microbial pan-genome. Curr. Opin. Genet Dev. *15*, 589-594.

Medrano-Soto,A., Moreno-Hagelsieb,G., Vinuesa,P., Christen,J.A., and Collado-Vides,J. (2004). Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. Mol. Biol. Evol. *21*, 1884-1894.

Mekalanos,J.J. (1983). Duplication and amplification of toxin genes in Vibrio cholerae. Cell *35*, 253-263.

Mengin-Lecreulx,D., Blanot,D., and van,H.J. (1994). Replacement of diaminopimelic acid by cystathionine or lanthionine in the peptidoglycan of Escherichia coli. J. Bacteriol. *176*, 4321-4327.

Meyer,A. and Schartl,M. (1999). Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Curr. Opin. Cell Biol. *11*, 699-704.

Mihok,S., Otieno,L.H., Darji,N., and Munyinyi,D. (1992). Influence of D(+)-glucosamine on infection rates and parasite loads in tsetse flies (Glossina spp.) infected with Trypanosoma brucei. Acta Trop. *51*, 217-228.

Milkman,R. and Crawford,I.P. (1983). Clustered third-base substitutions among wild strains of Escherichia coli. Science *221*, 378-380.

Bibliography

Mira,A., Klasson,L., and Andersson,S.G. (2002). Microbial genome evolution: sources of variability. Curr. Opin. Microbiol. *5*, 506-512.

Mira,A. and Moran,N.A. (2002). Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. Microb. Ecol. *44*, 137-143.

Mira,A., Ochman,H., and Moran,N.A. (2001). Deletional bias and the evolution of bacterial genomes. Trends Genet. *17*, 589-596.

Mira,A., Pushker,R., and Rodriguez-Valera,F. (2006). The Neolithic revolution of bacterial genomes. Trends Microbiol. *14*, 200-206.

Mirkin,B.G., Fenner,T.I., Galperin,M.Y., and Koonin,E.V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC. Evol. Biol. *3*, 2.

Mirold,S., Rabsch,W., Rohde,M., Stender,S., Tschape,H., Russmann,H., Igwe,E., and Hardt,W.D. (1999). Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic Salmonella typhimurium strain. Proc. Natl. Acad. Sci. U. S. A *96*, 9845-9850.

Mirold,S., Rabsch,W., Tschape,H., and Hardt,W.D. (2001). Transfer of the Salmonella type III effector sopE between unrelated phage families. J. Mol. Biol. *312*, 7-16.

Mirza,I.A., Nazi,I., Korczynska,M., Wright,G.D., and Berghuis,A.M. (2005). Crystal structure of homoserine transacetylase from Haemophilus influenzae reveals a new family of alpha/beta-hydrolases. Biochemistry *44*, 15768-15773.

Mobley,H.L. and Hausinger,R.P. (1989). Microbial ureases: significance, regulation, and molecular characterization. Microbiol. Rev. *53*, 85-108.

Monteiro-Vitorello,C.B., de Oliveira,M.C., Zerillo,M.M., Varani,A.M., Civerolo,E., and Van Sluys,M.A. (2005). Xylella and Xanthomonas Mobil'omics. OMICS. *9*, 146-159.

Moran,N.A. (2006). Symbiosis. Curr. Biol. *16*, R866-R871.

Moran,N.A. (2007). Symbiosis as an adaptive process and source of phenotypic complexity. Proc. Natl. Acad. Sci. U. S. A *104 Suppl 1*, 8627-8633.

Moran,N.A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. U. S. A *93*, 2873-2878.

Moran,N.A. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. Curr. Opin. Microbiol. *6*, 512-518.

Moran,N.A. (2002). Microbial minimalism: genome reduction in bacterial pathogens. Cell *108*, 583-586.

Moran,N.A., Degnan,P.H., Santos,S.R., Dunbar,H.E., and Ochman,H. (2005a). The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. Proc. Natl. Acad. Sci. U. S. A *102*, 16919-16926.

Moran,N.A., McCutcheon,J.P., and Nakabachi,A. (2008). Genomics and evolution of heritable bacterial symbionts. Annu. Rev. Genet. *42*, 165-190.

Moran,N.A., McLaughlin,H.J., and Sorek,R. (2009). The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science *323*, 379-382.

Moran,N.A. and Mira,A. (2001). The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol. *2*, RESEARCH0054.

Moran,N.A., Munson,M.A., Baumann,P., and Ishikawa,H. (1993). A molecular clock in endosymbiotic bacteia is calibrated using the insect hosts. Proc. R. Soc. Lond. B *253*, 167-171.

Moran,N.A. and Plague,G.R. (2004). Genomic changes following host restriction in bacteria. Curr. Opin. Genet. Dev. *14*, 627-633.

Moran,N.A., Russell,J.A., Koga,R., and Fukatsu,T. (2005b). Evolutionary relationships of three new species of Enterobacteriaceae living as symbionts of aphids and other insects. Appl. Environ. Microbiol. *71*, 3302-3310.

Moreira,D. and Philippe,H. (2000). Molecular phylogeny: pitfalls and progress. Int. Microbiol. *3*, 9-16.

Moret, B. M., Siepel, A. C., Tang, J, and Liu, T. Inversion Medians Outperform Breakpoint Medians in Phylogeny Reconstruction from Gene-Order Data. Proceedings of the Second International Workshop on Algorithms in Bioinformatics , 521-536. 2002a. Springer-Verlag.
Ref Type: Conference Proceeding

Moret,B.M., Tang,J., Wang,L.S., and Warnow,T. (2002b). Steps toward accurate reconstructions of phylogenies from gene-order data. J. Comput. Syst. Sci. *65*, 508-525.

Moret,B.M., Wang,L.S., Warnow,T., and Wyman,S.K. (2001). New approaches for reconstructing phylogenies from gene order data. Bioinformatics. *17 Suppl 1*, S165-S173.

Morett,E., Saab-Rincon,G., Olvera,L., Olvera,M., Flores,H., and Grande,R. (2008). Sensitive genome-wide screen for low secondary enzymatic activities: the YjbQ family shows thiamin phosphate synthase activity. J. Mol. Biol. *376*, 839-853.

Bibliography

Morgan,G.J., Hatfull,G.F., Casjens,S., and Hendrix,R.W. (2002). Bacteriophage Mu genome sequence: analysis and comparison with Mu-like prophages in Haemophilus, Neisseria and Deinococcus. J. Mol. Biol. *317*, 337-359.

Moriya,Y., Itoh,M., Okuda,S., Yoshizawa,A.C., and Kanehisa,M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. *35*, W182-W185.

Morizono,H., Cabrera-Luque,J., Shi,D., Gallegos,R., Yamaguchi,S., Yu,X., Allewell,N.M., Malamy,M.H., and Tuchman,M. (2006). Acetylornithine transcarbamylase: a novel enzyme in arginine biosynthesis. J. Bacteriol. *188*, 2974-2982.

Moya,A., Pereto,J., Gil,R., and Latorre,A. (2008). Learning how to live together: genomic insights into prokaryote-animal symbioses. Nat. Rev. Genet. *9*, 218-229.

Mrazek,J. and Karlin,S. (1999). Detecting alien genes in bacterial genomes. Ann. N. Y. Acad. Sci. *870*, 314-329.

Muller,H.J. (1964). The relation of recombination to mutational advance. Mutat. Res. *106*, 2-9.

Muller,T. and Vingron,M. (2000). Modeling amino acid replacement. J. Comput. Biol. *7*, 761-776.

Murray,K.D. and Bremer,H. (1996). Control of spoT-dependent ppGpp synthesis and degradation in Escherichia coli. J. Mol. Biol. *259*, 41-57.

Mushegian,A.R. and Koonin,E.V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc. Natl. Acad. Sci. U. S. A *93*, 10268-10273.

Nadeau,J.H. and Sankoff,D. (1998). Counting on comparative maps. Trends Genet. *14*, 495-501.

Nadeau,J.H. and Taylor,B.A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. Proc. Natl. Acad. Sci. U. S. A *81*, 814-818.

Nagai,S. and Flavin,M. (1967). Acetylhomoserine. An intermediate in the fungal biosynthesis of methionine. J. Biol. Chem. *242*, 3884-3895.

Nagakubo,S., Nishino,K., Hirata,T., and Yamaguchi,A. (2002). The putative response regulator BaeR stimulates multidrug resistance of Escherichia coli via a novel multidrug exporter system, MdtABC. J. Bacteriol. *184*, 4161-4167.

Nagy,Z. and Chandler,M. (2004). Regulation of transposition in bacteria. Res. Microbiol. *155*, 387-398.

Nakabachi,A., Yamashita,A., Toh,H., Ishikawa,H., Dunbar,H.E., Moran,N.A., and Hattori,M. (2006). The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science *314*, 267.

Nakagawa,I., Kurokawa,K., Yamashita,A., Nakata,M., Tomiyasu,Y., Okahashi,N., Kawabata,S., Yamazaki,K., Shiba,T., Yasunaga,T., Hayashi,H., Hattori,M., and Hamada,S. (2003). Genome sequence of an M3 strain of Streptococcus pyogenes reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. Genome Res. *13*, 1042-1055.

Nardon,P. and Nardon,C. (1998). Morphology and citology of symbiosis in insects. Ann. Soc. Entomol. (N. C.) *34*, 105-134.

Narita,S., Matsuyama,S., and Tokuda,H. (2004). Lipoprotein trafficking in Escherichia coli. Arch. Microbiol. *182*, 1-6.

Nei,M. (1996). Phylogenetic analysis in molecular evolutionary genetics. Annu. Rev. Genet. *30*, 371-403.

Nelson,K.E., Clayton,R.A., Gill,S.R., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Nelson,W.C., Ketchum,K.A., McDonald,L., Utterback,T.R., Malek,J.A., Linher,K.D., Garrett,M.M., Stewart,A.M., Cotton,M.D., Pratt,M.S., Phillips,C.A., Richardson,D., Heidelberg,J., Sutton,G.G., Fleischmann,R.D., Eisen,J.A., White,O., Salzberg,S.L., Smith,H.O., Venter,J.C., and Fraser,C.M. (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature *399*, 323-329.

Nesbo,C.L., Boucher,Y., and Doolittle,W.F. (2001). Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. J. Mol. Evol. *53*, 340-350.

Newton,I.L., Woyke,T., Auchtung,T.A., Dilly,G.F., Dutton,R.J., Fisher,M.C., Fontanez,K.M., Lau,E., Stewart,F.J., Richardson,P.M., Barry,K.W., Saunders,E., Detter,J.C., Wu,D., Eisen,J.A., and Cavanaugh,C.M. (2007). The Calyptogena magnifica chemoautotrophic symbiont genome. Science *315*, 998-1000.

Nichols,B.P. and Guay,G.G. (1989). Gene amplification contributes to sulfonamide resistance in Escherichia coli. Antimicrob. Agents Chemother. *33*, 2042-2048.

Nierman,W.C., DeShazer,D., Kim,H.S., Tettelin,H., Nelson,K.E., Feldblyum,T., Ulrich,R.L., Ronning,C.M., Brinkac,L.M., Daugherty,S.C., Davidsen,T.D., Deboy,R.T., Dimitrov,G., Dodson,R.J., Durkin,A.S., Gwinn,M.L., Haft,D.H., Khouri,H., Kolonay,J.F., Madupu,R., Mohammoud,Y., Nelson,W.C., Radune,D., Romero,C.M., Sarria,S., Selengut,J., Shamblin,C., Sullivan,S.A., White,O., Yu,Y., Zafar,N., Zhou,L., and Fraser,C.M. (2004). Structural flexibility in the Burkholderia mallei genome. Proc. Natl. Acad. Sci. U. S. A *101*, 14246-14251.

Nikaido,H. (2003). Molecular basis of bacterial outer membrane permeability revisited. Microbiol. Mol. Biol. Rev. *67*, 593-656.

Bibliography

Nogge,G. (1976). Sterility in tsetse flies caused by loss of symbionts. Experientia *32*, 995.

Nogge,G. (1981). Significance of symbionts for the maintenance of an optimal nutritional state for successful reproduction in hematophagous arthropods. Parasitology *82*, 299-304.

Notley-McRobb,L. and Ferenci,T. (2000). Substrate specificity and signal transduction pathways in the glucose-specific enzyme II (EII(Glc)) component of the Escherichia coli phosphotransferase system. J. Bacteriol. *182*, 4437-4442.

Oberhardt,M.A., Puchalka,J., Fryer,K.E., Martins,d.S., V, and Papin,J.A. (2008). Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1. J. Bacteriol. *190*, 2790-2803.

Ochman,H., Elwyn,S., and Moran,N.A. (1999). Calibrating bacterial evolution. Proc. Natl. Acad. Sci. U. S. A *96*, 12638-12643.

Ochman,H. and Groisman,E.A. (1996). Distribution of pathogenicity islands in Salmonella spp. Infect. Immun. *64*, 5410-5412.

Ochman,H., Lawrence,J.G., and Groisman,E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299-304.

Ochman,H., Lerat,E., and Daubin,V. (2005). Examining bacterial species under the specter of gene transfer and exchange. Proc. Natl. Acad. Sci. U. S. A *102 Suppl 1*, 6595-6599.

Ochman,H., Liu,R., and Rocha,E.P. (2007). Erosion of interaction networks in reduced and degraded genomes. J. Exp. Zoolog. B Mol. Dev. Evol. *308*, 97-103.

Ochman,H. and Moran,N.A. (2001). Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. Science *292*, 1096-1099.

Ochman,H. and Selander,R.K. (1984). Evidence for clonal population structure in Escherichia coli. Proc. Natl. Acad. Sci U. S. A *81*, 198-201.

Ochman,H., Soncini,F.C., Solomon,F., and Groisman,E.A. (1996). Identification of a pathogenicity island required for Salmonella survival in host cells. Proc. Natl. Acad. Sci. U. S. A *93*, 7800-7804.

Ochman,H., Whittam,T.S., Caugant,D.A., and Selander,R.K. (1983). Enzyme polymorphism and genetic population structure in Escherichia coli and Shigella. J. Gen. Microbiol. *129*, 2715-2726.

Oda,Y., Larimer,F.W., Chain,P.S., Malfatti,S., Shin,M.V., Vergez,L.M., Hauser,L., Land,M.L., Braatsch,S., Beatty,J.T., Pelletier,D.A., Schaefer,A.L., and Harwood,C.S. (2008). Multiple genome sequences reveal adaptations of a

phototrophic bacterium to sediment microenvironments. Proc. Natl. Acad. Sci. U. S. A *105*, 18543-18548.

Ogata,H., Audic,S., Renesto-Audiffren,P., Fournier,P.E., Barbe,V., Samson,D., Roux,V., Cossart,P., Weissenbach,J., Claverie,J.M., and Raoult,D. (2001). Mechanisms of evolution in Rickettsia conorii and R. prowazekii. Science *293*, 2093-2098.

Ogura,Y., Ooka,T., Iguchi,A., Toh,H., Asadulghani,M., Oshima,K., Kodama,T., Abe,H., Nakayama,K., Kurokawa,K., Tobe,T., Hattori,M., and Hayashi,T. (2009). Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. Proc. Natl. Acad. Sci U. S. A *106*, 17939-17944.

Oh,H., Park,Y., and Park,C. (1999). A mutated PtsG, the glucose transporter, allows uptake of D-ribose. J. Biol. Chem. *274*, 14006-14011.

Ohnishi,M., Kurokawa,K., and Hayashi,T. (2001). Diversification of Escherichia coli genomes: are bacteriophages the major contributors? Trends Microbiol. *9*, 481-485.

Ohno,S. (1970). Evolution by Gene Duplication. (New York: Springer-Verlag).

Oliver,K.M., Russell,J.A., Moran,N.A., and Hunter,M.S. (2003). Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. Proc. Natl. Acad. Sci. U. S. A *100*, 1803-1807.

Onuffer,J.J., Ton,B.T., Klement,I., and Kirsch,J.F. (1995). The use of natural and unnatural amino acid substrates to define the substrate specificity differences of Escherichia coli aspartate and tyrosine aminotransferases. Protein Sci *4*, 1743-1749.

Orgel,L.E., Crick,F.H., and Sapienza,C. (1980). Selfish DNA. Nature *288*, 645-646.

Oshima,K., Kakizawa,S., Nishigawa,H., Jung,H.Y., Wei,W., Suzuki,S., Arashida,R., Nakata,D., Miyata,S., Ugaki,M., and Namba,S. (2004). Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. Nat. Genet *36*, 27-29.

Ouzounis,C.A., Kunin,V., Darzentas,N., and Goldovsky,L. (2006). A minimal estimate for the gene content of the last universal common ancestor--exobiology from a terrestrial perspective. Res. Microbiol. *157*, 57-68.

Overbeek,R., Larsen,N., Pusch,G.D., D'Souza,M., Selkov E Jr, Kyrpides,N., Fonstein,M., Maltsev,N., and Selkov,E. (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res. *28*, 123-125.

Bibliography

Pais,R., Lohs,C., Wu,Y., Wang,J., and Aksoy,S. (2008). The obligate mutualist Wigglesworthia glossinidia influences reproduction, digestion, and immunity processes of its host, the tsetse fly. Appl. Environ. Microbiol. *74*, 5965-5974.

Pal,C., Papp,B., and Hurst,L.D. (2001). Highly expressed genes in yeast evolve slowly. Genetics *158*, 927-931.

Pal,C., Papp,B., and Lercher,M.J. (2006a). An integrated view of protein evolution. Nat. Rev. Genet *7*, 337-348.

Pal,C., Papp,B., and Lercher,M.J. (2005b). Horizontal gene transfer depends on gene content of the host. Bioinformatics. *21 Suppl 2*, ii222-ii223.

Pal,C., Papp,B., and Lercher,M.J. (2005a). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat. Genet *37*, 1372-1375.

Pal,C., Papp,B., Lercher,M.J., Csermely,P., Oliver,S.G., and Hurst,L.D. (2006b). Chance and necessity in the evolution of minimal metabolic networks. Nature *440*, 667-670.

Palacios,C. and Wernegreen,J.J. (2002). A strong effect of AT mutational bias on amino acid usage in Buchnera is mitigated at high-expression genes. Mol. Biol. Evol. *19*, 1575-1584.

Palsson,B.O. (2006). Systems Biology: Properties of reconstructed networks. (New York: Cambridge University Press).

Palsson,B.O., Price,N.D., and Papin,J.A. (2003). Development of network-based pathway definitions: the need to analyze real metabolic networks. Trends Biotechnol. *21*, 195-198.

Papin,J.A., Stelling,J., Price,N.D., Klamt,S., Schuster,S., and Palsson,B.O. (2004). Comparison of network-based pathway analysis methods. Trends Biotechnol. *22*, 400-405.

Parkhill,J., Sebaihia,M., Preston,A., Murphy,L.D., Thomson,N., Harris,D.E., Holden,M.T., Churcher,C.M., Bentley,S.D., Mungall,K.L., Cerdeno-Tarraga,A.M., Temple,L., James,K., Harris,B., Quail,M.A., Achtman,M., Atkin,R., Baker,S., Basham,D., Bason,N., Cherevach,I., Chillingworth,T., Collins,M., Cronin,A., Davis,P., Doggett,J., Feltwell,T., Goble,A., Hamlin,N., Hauser,H., Holroyd,S., Jagels,K., Leather,S., Moule,S., Norberczak,H., O'Neil,S., Ormond,D., Price,C., Rabbinowitsch,E., Rutter,S., Sanders,M., Saunders,D., Seeger,K., Sharp,S., Simmonds,M., Skelton,J., Squares,R., Squares,S., Stevens,K., Unwin,L., Whitehead,S., Barrell,B.G., and Maskell,D.J. (2003). Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. Nat. Genet *35*, 32-40.

Parkhill,J., Wren,B.W., Thomson,N.R., Titball,R.W., Holden,M.T., Prentice,M.B., Sebaihia,M., James,K.D., Churcher,C., Mungall,K.L., Baker,S., Basham,D., Bentley,S.D., Brooks,K., Cerdeno-Tarraga,A.M., Chillingworth,T., Cronin,A., Davies,R.M., Davis,P., Dougan,G., Feltwell,T., Hamlin,N., Holroyd,S., Jagels,K., Karlyshev,A.V., Leather,S., Moule,S., Oyston,P.C., Quail,M., Rutherford,K., Simmonds,M., Skelton,J., Stevens,K., Whitehead,S., and Barrell,B.G. (2001). Genome sequence of Yersinia pestis, the causative agent of plague. Nature *413*, 523-527.

Pastor-Satorras,R., Smith,E., and Sole,R.V. (2003). Evolving protein interaction networks through gene duplication. J. Theor. Biol. *222*, 199-210.

Peacock,L., Ferris,V., Bailey,M., and Gibson,W. (2006). Multiple effects of the lectin-inhibitory sugars D-glucosamine and N-acetyl-glucosamine on tsetse-trypanosome interactions. Parasitology *132*, 651-658.

Pearson,W.R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. *183*, 63-98.

Pedulla,M.L., Ford,M.E., Houtz,J.M., Karthikeyan,T., Wadsworth,C., Lewis,J.A., Jacobs-Sera,D., Falbo,J., Gross,J., Pannunzio,N.R., Brucker,W., Kumar,V., Kandasamy,J., Keenan,L., Bardarov,S., Kriakov,J., Lawrence,J.G., Jacobs,W.R., Jr., Hendrix,R.W., and Hatfull,G.F. (2003). Origins of highly mosaic mycobacteriophage genomes. Cell *113*, 171-182.

Peekhaus,N. and Conway,T. (1998). What's for dinner?: Entner-Doudoroff metabolism in Escherichia coli. J. Bacteriol. *180*, 3495-3502.

Peng,L., rauzo-Bravo,M.J., and Shimizu,K. (2004). Metabolic flux analysis for a ppc mutant Escherichia coli based on 13C-labelling experiments together with enzyme activity assays and intracellular metabolite measurements. FEMS Microbiol. Lett. *235*, 17-23.

Peng,L. and Shimizu,K. (2004). Effect of ppc gene knockout on the metabolism of Escherichia coli in view of gene expressions, enzyme activities and intracellular metabolite concentrations. Appl. Microbiol. Biotechnol.

Pereto,J. (2005). Controversies on the origin of life. Int. Microbiol. *8*, 23-31.

Perez-Brocal,V., Gil,R., Ramos,S., Lamelas,A., Postigo,M., Michelena,J.M., Silva,F.J., Moya,A., and Latorre,A. (2006). A small microbial genome: the end of a long symbiotic relationship? Science *314*, 312-313.

Perlman,S.J., Kelly,S.E., and Hunter,M.S. (2008). Population biology of cytoplasmic incompatibility: maintenance and spread of Cardinium symbionts in a parasitic wasp. Genetics *178*, 1003-1011.

Bibliography

Perna,N.T., Plunkett,G., III, Burland,V., Mau,B., Glasner,J.D., Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A., Posfai,G., Hackett,J., Klink,S., Boutin,A., Shao,Y., Miller,L., Grotbeck,E.J., Davis,N.W., Lim,A., Dimalanta,E.T., Potamousis,K.D., Apodaca,J., Anantharaman,T.S., Lin,J., Yen,G., Schwartz,D.C., Welch,R.A., and Blattner,F.R. (2001). Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature *409*, 529-533.

Pfeiffer,T., Sanchez-Valdenebro,I., Nuno,J.C., Montero,F., and Schuster,S. (1999). METATOOL: for studying metabolic networks. Bioinformatics. *15*, 251-257.

Pharkya,P., Burgard,A.P., and Maranas,C.D. (2004). OptStrain: a computational framework for redesign of microbial production systems. Genome Res. *14*, 2367-2376.

Philippe,H., Zhou,Y., Brinkmann,H., Rodrigue,N., and Delsuc,F. (2005). Heterotachy and long-branch attraction in phylogenetics. BMC. Evol. Biol. *5*, 50.

Plague,G.R., Dunbar,H.E., Tran,P.L., and Moran,N.A. (2008). Extensive proliferation of transposable elements in heritable bacterial symbionts. J. Bacteriol. *190*, 777-779.

Planes,F.J. and Beasley,J.E. (2008). A critical examination of stoichiometric and path-finding approaches to metabolic pathways. Brief. Bioinform. *9*, 422-436.

Podani,J., Oltvai,Z.N., Jeong,H., Tombor,B., Barabasi,A.L., and Szathmary,E. (2001). Comparable system-level organization of Archaea and Eukaryotes. Nat. Genet *29*, 54-56.

Posada,D. and Buckley,T.R. (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst. Biol. *53*, 793-808.

Postma,P.W., Lengeler,J.W., and Jacobson,G.R. (1993). Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. Microbiol. Rev. *57*, 543-594.

Preston,G.M., Haubold,B., and Rainey,P.B. (1998). Bacterial genomics and adaptation to life on plants: implications for the evolution of pathogenicity and symbiosis. Curr. Opin. Microbiol. *1*, 589-597.

Price,M.N., Dehal,P.S., and Arkin,A.P. (2008). Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. Genome Biol. *9*, R4.

Price,N.D., Papin,J.A., Schilling,C.H., and Palsson,B.O. (2003). Genome-scale microbial in silico models: the constraints-based approach. Trends Biotechnol. *21*, 162-169.

Price,N.D., Reed,J.L., and Palsson,B.O. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat. Rev. Microbiol. *2*, 886-897.

Priest,F.G., Barker,M., Baillie,L.W., Holmes,E.C., and Maiden,M.C. (2004). Population structure and evolution of the Bacillus cereus group. J. Bacteriol. *186*, 7959-7970.

Proudfoot,M., Kuznetsova,E., Brown,G., Rao,N.N., Kitagawa,M., Mori,H., Savchenko,A., and Yakunin,A.F. (2004). General enzymatic screens identify three new nucleotidases in Escherichia coli. Biochemical characterization of SurE, YfbR, and YjjG. J. Biol. Chem. *279*, 54687-54694.

Puchalka,J., Oberhardt,M.A., Godinho,M., Bielecka,A., Regenhardt,D., Timmis,K.N., Papin,J.A., and Martins,d.S., V (2008). Genome-scale reconstruction and analysis of the Pseudomonas putida KT2440 metabolic network facilitates applications in biotechnology. PLoS Comput. Biol. *4*, e1000210.

Pugsley,A.P., Francetic,O., Driessen,A.J., and de,L., V (2004). Getting out: protein traffic in prokaryotes. Mol. Microbiol. *52*, 3-11.

Puigbo,P., Bravo,I.G., and Garcia-Vallve,S. (2008a). E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). BMC. Bioinformatics. *9*, 65.

Puigbo,P., Romeu,A., and Garcia-Vallve,S. (2008b). HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection. Nucleic Acids Res. *36*, D524-D527.

Pushker,R., Mira,A., and Rodriguez-Valera,F. (2004). Comparative genomics of gene-family size in closely related bacteria. Genome Biol. *5*, R27.

Raetz,C.R.H. (1996). Bacterial lipopolysaccharides: A remarkable family of bioactive macroamphiphiles. In Escherichia coli and Salmonella cellular and molecular biology, F.C.Neidhardt, ed. (Washington D.C.: American Society for Microbiology), pp. 1035-1063.

Raffaelli,N., Lorenzi,T., Mariani,P.L., Emanuelli,M., Amici,A., Ruggieri,S., and Magni,G. (1999). The Escherichia coli NadR regulator is endowed with nicotinamide mononucleotide adenylyltransferase activity. J. Bacteriol. *181*, 5509-5511.

Ragan,M.A. (2001). On surrogate methods for detecting lateral gene transfer. FEMS Microbiol. Lett. *201*, 187-191.

Rain,J.C., Selig,L., De,R.H., Battaglia,V., Reverdy,C., Simon,S., Lenzen,G., Petel,F., Wojcik,J., Schachter,V., Chemama,Y., Labigne,A., and Legrain,P. (2001). The protein-protein interaction map of Helicobacter pylori. Nature *409*, 211-215.

Bibliography

Ramakrishna,R., Edwards,J.S., McCulloch,A., and Palsson,B.O. (2001). Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints. Am. J. Physiol Regul. Integr. Comp Physiol *280*, R695-R704.

Raman,K., Rajagopalan,P., and Chandra,N. (2005). Flux balance analysis of mycolic acid pathway: targets for anti-tubercular drugs. PLoS Comput. Biol. *1*, e46.

Ranea,J.A., Buchan,D.W., Thornton,J.M., and Orengo,C.A. (2004). Evolution of protein superfamilies and bacterial genome size. J. Mol. Biol. *336*, 871-887.

Raoult,D., Ogata,H., Audic,S., Robert,C., Suhre,K., Drancourt,M., and Claverie,J.M. (2003). Tropheryma whipplei Twist: a human pathogenic Actinobacteria with a reduced genome. Genome Res. *13*, 1800-1809.

Rasko,D.A., Rosovitz,M.J., Myers,G.S., Mongodin,E.F., Fricke,W.F., Gajer,P., Crabtree,J., Sebaihia,M., Thomson,N.R., Chaudhuri,R., Henderson,I.R., Sperandio,V., and Ravel,J. (2008). The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J. Bacteriol. *190*, 6881-6893.

Record,M.T., Reznikoff,W.S., Craig,M.L., McQuade,K.L., and Schlax,P.J. (1996). Escherichia coli RNA polymerase (Esigma70), promoters, and the kinetics of the steps of transcription initiation. In Escherichia coli and Salmonella Cellular and Molecular Biology, F.C.Neidhardt, ed. (Washington,D.C.: American Society for Microbiology), pp. 792-821.

Reder,C. (1988). Metabolic control theory: a structural approach. J. Theor. Biol. *135*, 175-201.

Reed,J.L., Vo,T.D., Schilling,C.H., and Palsson,B.O. (2003). An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol. *4*, R54.

Reitzer,L.J. and Magasanik,B. (1987). Ammonia assimilation and the biosynthesis of Glutamine, Glutamate, Aspartate, Asparragine, L-Alanine, and D-Alanine. In Escherichia coli and Salmonella typhimurium cellular and molecular biology, F.C.Neidhardt, ed. (Washington D.C.: American Society for Microbiology), pp. 302-320.

Reznikoff,W.S. (1993). The Tn5 transposon. Annu. Rev. Microbiol. *47*, 945-963.

Rice,P., Longden,I., and Bleasby,A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet *16*, 276-277.

Riehle,M.M., Bennett,A.F., and Long,A.D. (2001). Genetic architecture of thermal adaptation in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A *98*, 525-530.

Rigby,P.W., Burleigh,B.D., Jr., and Hartley,B.S. (1974). Gene duplication in experimental enzyme evolution. Nature *251*, 200-204.

Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R., Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T., Mori,H., Perna,N.T., Plunkett,G., III, Rudd,K.E., Serres,M.H., Thomas,G.H., Thomson,N.R., Wishart,D., and Wanner,B.L. (2006). Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. Nucleic Acids Res. *34*, 1-9.

Rio,R.V., Hu,Y., and Aksoy,S. (2004). Strategies of the home-team: symbioses exploited for vector-borne disease control. Trends Microbiol. *12*, 325-336.

Rio,R.V., Lefevre,C., Heddi,A., and Aksoy,S. (2003). Comparative genomics of insect-symbiotic bacteria: influence of host environment on microbial genome composition. Appl. Environ. Microbiol. *69*, 6825-6832.

Rio,R.V., Wu,Y.N., Filardo,G., and Aksoy,S. (2006). Dynamics of multiple symbiont density regulation during host development: tsetse fly and its microbial flora. Proc. Biol. Sci. *273*, 805-814.

Rispe,C., Delmotte,F., van Ham,R.C., and Moya,A. (2004). Mutational and selective pressures on codon and amino acid usage in Buchnera, endosymbiotic bacteria of aphids. Genome Res. *14*, 44-53.

Rocha,E.P. (2006). Inference and analysis of the relative stability of bacterial chromosomes. Mol. Biol. Evol. *23*, 513-522.

Rocha,E.P. (2003a). An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. Genome Res. *13*, 1123-1132.

Rocha,E.P. (2003b). DNA repeats lead to the accelerated loss of gene order in bacteria. Trends Genet. *19*, 600-603.

Rocha,E.P. (2004). Order and disorder in bacterial genomes. Curr. Opin. Microbiol. *7*, 519-527.

Rocha,E.P. (2008a). Evolutionary patterns in prokaryotic genomes. Curr. Opin. Microbiol. *11*, 454-460.

Rocha,E.P. (2008b). The organization of the bacterial genome. Annu. Rev. Genet. *42*, 211-233.

Rocha,E.P. and Danchin,A. (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol. Biol. Evol. *21*, 108-116.

Rocha,E.P. and Danchin,A. (2003a). Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat. Genet. *34*, 377-378.

Bibliography

Rocha,E.P. and Danchin,A. (2003b). Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res. *31*, 6570-6577.

Rocha,E.P., Danchin,A., and Viari,A. (1999). Functional and evolutionary roles of long repeats in prokaryotes. Res. Microbiol. *150*, 725-733.

Rocha,E.P., Danchin,A., and Viari,A. (2001). Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. Genome Res. *11*, 946-958.

Rocha,I., Forster,J., and Nielsen,J. (2008). Design and application of genome-scale reconstructed metabolic models. Methods Mol. Biol. *416*, 409-431.

Rockafellar,R.T. (1970). Convex analysis. (Princeton,NJ: Princeton University Press).

Roditi,I. and Lehane,M.J. (2008). Interactions between trypanosomes and tsetse flies. Curr. Opin. Microbiol. *11*, 345-351.

Rodriguez-Valera,F., Martin-Cuadrado,A.B., Rodriguez-Brito,B., Pasic,L., Thingstad,T.F., Rohwer,F., and Mira,A. (2009). Explaining microbial population genomics through phage predation. Nat. Rev. Microbiol. *7*, 828-836.

Rokas,A., Williams,B.L., King,N., and Carroll,S.B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature *425*, 798-804.

Romero,D. and Palacios,R. (1997). Gene amplification and genomic plasticity in prokaryotes. Annu. Rev. Genet. *31*, 91-111.

Rosenberg,E.Y., Ma,D., and Nikaido,H. (2000). AcrD of Escherichia coli is an aminoglycoside efflux pump. J. Bacteriol. *182*, 1754-1756.

Ross-Macdonald,P., Coelho,P.S., Roemer,T., Agarwal,S., Kumar,A., Jansen,R., Cheung,K.H., Sheehan,A., Symoniatis,D., Umansky,L., Heidtman,M., Nelson,F.K., Iwasaki,H., Hager,K., Gerstein,M., Miller,P., Roeder,G.S., and Snyder,M. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature *402*, 413-418.

Rouhbakhsh,D., Lai,C.Y., von Dohlen,C.D., Clark,M.A., Baumann,L., Baumann,P., Moran,N.A., and Voegtlin,D.J. (1996). The tryptophan biosynthetic pathway of aphid endosymbionts (Buchnera): genetics and evolution of plasmid-associated anthranilate synthase (trpEG) within the aphididae. J. Mol. Evol. *42*, 414-421.

Rush,D., Karibian,D., Karonovsky,M.L., and Magasanik,B. (1957). Pathways of glycerol dissimilation in two strains of Aerobacter aerogenes; enzymatic and tracer studies. J. Biol. Chem. *226*, 891-899.

Russell,J.A., Latorre,A., Sabater-Munoz,B., Moya,A., and Moran,N.A. (2003). Side-stepping secondary symbionts: widespread horizontal transfer across and beyond the Aphidoidea. Mol. Ecol. *12*, 1061-1075.

Russell,J.A. and Moran,N.A. (2006). Costs and benefits of symbiont infection in aphids: variation among symbionts and across temperatures. Proc. Biol. Sci. *273*, 603-610.

Sahara,T., Takada,Y., Takeuchi,Y., Yamaoka,N., and Fukunaga,N. (2002). Cloning, sequencing, and expression of a gene encoding the monomeric isocitrate dehydrogenase of the nitrogen-fixing bacterium, Azotobacter vinelandii. Biosci. Biotechnol. Biochem. *66*, 489-500.

Saier,M.H., Jr. and Paulsen,I.T. (1999). Paralogous genes encoding transport proteins in microbial genomes. Res. Microbiol. *150*, 689-699.

Saitou,N. and Nei,M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406-425.

Salgado,H., Gama-Castro,S., Peralta-Gil,M., az-Peredo,E., Sanchez-Solano,F., Santos-Zavaleta,A., Martinez-Flores,I., Jimenez-Jacinto,V., Bonavides-Martinez,C., Segura-Salazar,J., Martinez-Antonio,A., and Collado-Vides,J. (2006a). RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res. *34*, D394-D397.

Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Peralta-Gil,M., Penaloza-Spinola,M.I., Martinez-Antonio,A., Karp,P.D., and Collado-Vides,J. (2006b). The comprehensive updated regulatory network of Escherichia coli K-12. BMC. Bioinformatics. *7*, 5.

Sallstrom,B. and Andersson,S.G. (2005). Genome reduction in the alpha-Proteobacteria. Curr. Opin. Microbiol. *8*, 579-585.

Sandstrom,J.P., Russell,J.A., White,J.P., and Moran,N.A. (2001). Independent origins and horizontal transfer of bacterial symbionts of aphids. Mol. Ecol. *10*, 217-228.

Sankoff,D. and Nadeau,J.H. (2003). Chromosome rearrangements in evolution: From gene order to genome sequence and back. Proc. Natl. Acad. Sci. U. S. A *100*, 11188-11189.

Sargent,F., Bogsch,E.G., Stanley,N.R., Wexler,M., Robinson,C., Berks,B.C., and Palmer,T. (1998). Overlapping functions of components of a bacterial Sec-independent protein export pathway. EMBO J. *17*, 3640-3650.

Satishchandran,C. and Boyle,S.M. (1986). Purification and properties of agmatine ureohydrolyase, a putrescine biosynthetic enzyme in Escherichia coli. J. Bacteriol. *165*, 843-848.

Bibliography

Sauer,C., Stackebrandt,E., Gadau,J., Holldobler,B., and Gross,R. (2000). Systematic relationships and cospeciation of bacterial endosymbionts and their carpenter ant host species: proposal of the new taxon Candidatus Blochmannia gen. nov. Int. J. Syst. Evol. Microbiol. *50 Pt 5*, 1877-1886.

Sawyer,S.A., Dykhuizen,D.E., DuBose,R.F., Green,L., Mutangadura-Mhlanga,T., Wolczyk,D.F., and Hartl,D.L. (1987). Distribution and abundance of insertion sequences among natural isolates of Escherichia coli. Genetics *115*, 51-63.

Scarborough,C.L., Ferrari,J., and Godfray,H.C. (2005). Aphid protected from pathogen by endosymbiont. Science *310*, 1781.

Schaber,J., Rispe,C., Wernegreen,J., Buness,A., Delmotte,F., Silva,F.J., and Moya,A. (2005). Gene expression levels influence amino acid usage and evolutionary rates in endosymbiotic bacteria. Gene *352*, 109-117.

Schauder,S., Shokat,K., Surette,M.G., and Bassler,B.L. (2001). The LuxS family of bacterial autoinducers: biosynthesis of a novel quorum-sensing signal molecule. Mol. Microbiol. *41*, 463-476.

Schilling,C.H., Covert,M.W., Famili,I., Church,G.M., Edwards,J.S., and Palsson,B.O. (2002). Genome-scale metabolic model of Helicobacter pylori 26695. J. Bacteriol. *184*, 4582-4593.

Schilling,C.H., Edwards,J.S., Letscher,D., and Palsson,B.O. (2000a). Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. Biotechnol. Bioeng. *71*, 286-306.

Schilling,C.H., Letscher,D., and Palsson,B.O. (2000b). Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. J. Theor. Biol. *203*, 229-248.

Schilling,C.H. and Palsson,B.O. (2000). Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis. J. Theor. Biol. *203*, 249-283.

Schilling,C.H., Schuster,S., Palsson,B.O., and Heinrich,R. (1999). Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. Biotechnol. Prog. *15*, 296-303.

Schmid,M.B. and Roth,J.R. (1987). Gene location affects expression level in Salmonella typhimurium. J. Bacteriol. *169*, 2872-2875.

Schmidt,H.A., Strimmer,K., Vingron,M., and von,H.A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. *18*, 502-504.

Schneider,D., Duperchy,E., Coursange,E., Lenski,R.E., and Blot,M. (2000). Long-term experimental evolution in Escherichia coli. IX. Characterization of insertion sequence-mediated mutations and rearrangements. Genetics *156*, 477-488.

Schneider,D. and Lenski,R.E. (2004). Dynamics of insertion sequence elements during experimental evolution of bacteria. Res. Microbiol. *155*, 319-327.

Schneiker,S., Perlova,O., Kaiser,O., Gerth,K., Alici,A., Altmeyer,M.O., Bartels,D., Bekel,T., Beyer,S., Bode,E., Bode,H.B., Bolten,C.J., Choudhuri,J.V., Doss,S., Elnakady,Y.A., Frank,B., Gaigalat,L., Goesmann,A., Groeger,C., Gross,F., Jelsbak,L., Jelsbak,L., Kalinowski,J., Kegler,C., Knauber,T., Konietzny,S., Kopp,M., Krause,L., Krug,D., Linke,B., Mahmud,T., Martinez-Arias,R., McHardy,A.C., Merai,M., Meyer,F., Mormann,S., Munoz-Dorado,J., Perez,J., Pradella,S., Rachid,S., Raddatz,G., Rosenau,F., Ruckert,C., Sasse,F., Scharfe,M., Schuster,S.C., Suen,G., Treuner-Lange,A., Velicer,G.J., Vorholter,F.J., Weissman,K.J., Welch,R.D., Wenzel,S.C., Whitworth,D.E., Wilhelm,S., Wittmann,C., Blocker,H., Puhler,A., and Muller,R. (2007). Complete genome sequence of the myxobacterium Sorangium cellulosum. Nat. Biotechnol. *25*, 1281-1289.

Schoen,C., Blom,J., Claus,H., Schramm-Gluck,A., Brandt,P., Muller,T., Goesmann,A., Joseph,B., Konietzny,S., Kurzai,O., Schmitt,C., Friedrich,T., Linke,B., Vogel,U., and Frosch,M. (2008). Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis. Proc. Natl. Acad. Sci. U. S. A *105*, 3473-3478.

Schofield,M.J. and Hsieh,P. (2003). DNA mismatch repair: molecular mechanisms and biological function. Annu. Rev. Microbiol. *57*, 579-608.

Schuster,S., Dandekar,T., and Fell,D.A. (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends Biotechnol. *17*, 53-60.

Schuster,S., Fell,D.A., and Dandekar,T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat. Biotechnol. *18*, 326-332.

Schwarz,G. (2005). Molybdenum cofactor biosynthesis and deficiency. Cell Mol. Life Sci *62*, 2792-2810.

Segre,D., Vitkup,D., and Church,G.M. (2002). Analysis of optimality in natural and perturbed metabolic networks. Proc. Natl. Acad. Sci U. S. A *99*, 15112-15117.

Serres,M.H. and Riley,M. (2000). MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. Microb. Comp Genomics *5*, 205-222.

Bibliography

Seshasayee,A.S., Bertone,P., Fraser,G.M., and Luscombe,N.M. (2006). Transcriptional regulatory networks in bacteria: from input signals to output responses. Curr. Opin. Microbiol. *9*, 511-519.

Sharp,P.M. (1991). Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: codon usage, map position, and concerted evolution. J. Mol. Evol. *33*, 23-33.

Sharp,P.M. and Li,W.H. (1987a). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. *15*, 1281-1295.

Sharp,P.M. and Li,W.H. (1987b). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. *4*, 222-230.

Sharples,G.J. (2009). For absent friends: life without recombination in mutualistic gamma-proteobacteria. Trends Microbiol. *17*, 233-242.

Shaw,M.K. and Moloo,S.K. (1991). Comparative study on Rickettsia-like organisms in the midgut epithelial cells of different Glossina species. Parasitology *102 Pt 2*, 193-199.

Shigenobu,S., Watanabe,H., Hattori,M., Sakaki,Y., and Ishikawa,H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature *407*, 81-86.

Shlomi,T., Berkman,O., and Ruppin,E. (2005). Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proc. Natl. Acad. Sci U. S. A *102*, 7695-7700.

Sicheritz-Ponten,T. and Andersson,S.G. (2001). A phylogenomic approach to microbial evolution. Nucleic Acids Res. *29*, 545-552.

Sigrist,C.J., Cerutti,L., de,C.E., Langendijk-Genevaux,P.S., Bulliard,V., Bairoch,A., and Hulo,N. (2009). PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res.

Siguier,P., Filee,J., and Chandler,M. (2006a). Insertion sequences in prokaryotic genomes. Curr. Opin. Microbiol. *9*, 526-531.

Siguier,P., Perochon,J., Lestrade,L., Mahillon,J., and Chandler,M. (2006b). ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. *34*, D32-D36.

Silva,F.J. and Latorre,A. (2008). Genome Reduction During Prokaryotic Evolution. In Computational Methods for Understanding Bacterial and Archaeal Genomes, Y.Xu and J.P.Gogarten, eds. (London: Imperial College Press), pp. 153-184.

Silva,F.J., Latorre,A., Gomez-Valero,L., and Moya,A. (2007). Genomic changes in bacteria: From free-living to endosymbiotic life. In Structural approaches to sequence evolution, U.Bastolla, M.Porto, E.Roman, and M.Vendruscolo, eds. Springelink), pp. 149-165.

Silva,F.J., Latorre,A., and Moya,A. (2001). Genome size reduction through multiple events of gene disintegration in Buchnera APS. Trends Genet. *17*, 615-618.

Silva,F.J., Latorre,A., and Moya,A. (2003). Why are the genomes of endosymbiotic bacteria so stable? Trends Genet. *19*, 176-180.

Smallbone,K., Simeonidis,E., Broomhead,D.S., and Kell,D.B. (2007). Something from nothing: bridging the gap between constraint-based and kinetic modelling. FEBS J. *274*, 5576-5585.

Snel,B., Bork,P., and Huynen,M.A. (1999). Genome phylogeny based on gene content. Nat. Genet. *21*, 108-110.

Snel,B., Bork,P., and Huynen,M.A. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res. *12*, 17-25.

Snel,B., Huynen,M.A., and Dutilh,B.E. (2005). Genome trees and the nature of genome evolution. Annu. Rev. Microbiol. *59*, 191-209.

Sonnleitner,B. and Kappeli,O. (1986). Growth of Saccharomyces cerevisiae is controlled by its limited respiratory capacity: Formulation and verification of a hypothesis. Biotechnol. Bioeng. *28*, 927-937.

Sonti,R.V. and Roth,J.R. (1989). Role of gene duplications in the adaptation of Salmonella typhimurium to growth on limiting carbon sources. Genetics *123*, 19-28.

Sousa,C., de,L., V, and Cebolla,A. (1997). Modulation of gene expression through chromosomal positioning in Escherichia coli. Microbiology *143 (Pt 6)*, 2071-2078.

Stackebrandt,E., Frederiksen,W., Garrity,G.M., Grimont,P.A., Kampfer,P., Maiden,M.C., Nesme,X., Rossello-Mora,R., Swings,J., Truper,H.G., Vauterin,L., Ward,A.C., and Whitman,W.B. (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. *52*, 1043-1047.

Steinert,M., Hentschel,U., and Hacker,J. (2000). Symbiosis and pathogenesis: evolution of the microbe-host interaction. Naturwissenschaften *87*, 1-11.

Stelling,J., Klamt,S., Bettenbrock,K., Schuster,S., and Gilles,E.D. (2002). Metabolic network structure determines key aspects of functionality and regulation. Nature *420*, 190-193.

Bibliography

Stinear,T.P., Seemann,T., Harrison,P.F., Jenkin,G.A., Davies,J.K., Johnson,P.D., Abdellah,Z., Arrowsmith,C., Chillingworth,T., Churcher,C., Clarke,K., Cronin,A., Davis,P., Goodhead,I., Holroyd,N., Jagels,K., Lord,A., Moule,S., Mungall,K., Norbertczak,H., Quail,M.A., Rabbinowitsch,E., Walker,D., White,B., Whitehead,S., Small,P.L., Brosch,R., Ramakrishnan,L., Fischbach,M.A., Parkhill,J., and Cole,S.T. (2008). Insights from the complete genome sequence of Mycobacterium marinum on the evolution of Mycobacterium tuberculosis. Genome Res. *18*, 729-741.

Stock,J.B., Waygood,E.B., Meadow,N.D., Postma,P.W., and Roseman,S. (1982). Sugar transport by the bacterial phosphotransferase system. The glucose receptors of the Salmonella typhimurium phosphotransferase system. J. Biol. Chem. *257*, 14543-14552.

Stouthamer,R., Breeuwer,J.A., and Hurst,G.D. (1999). Wolbachia pipientis: microbial manipulator of arthropod reproduction. Annu. Rev. Microbiol. *53*, 71-102.

Strimmer,K. and von,H.A. (1996). Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. *13*, 964-969.

Strimmer,K. and von,H.A. (2003). Nucleotide substitution models. In The phylogenetic handbook, M.Salemi and A.M.Vandamme, eds. (New York: Cambridge University Press), pp. 72-88.

Sullivan,J.T. and Ronson,C.W. (1998). Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. Proc. Natl. Acad. Sci. U. S. A *95*, 5145-5149.

Sullivan,M.B., Coleman,M.L., Weigele,P., Rohwer,F., and Chisholm,S.W. (2005). Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. PLoS Biol. *3*, e144.

Summer,E.J., Gonzalez,C.F., Carlisle,T., Mebane,L.M., Cass,A.M., Savva,C.G., LiPuma,J., and Young,R. (2004). Burkholderia cenocepacia phage BcepMu and a family of Mu-like phages encoding potential pathogenesis factors. J. Mol. Biol. *340*, 49-65.

Susko,E., Inagaki,Y., and Roger,A.J. (2004). On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. Mol. Biol. Evol. *21*, 1629-1642.

Susskind,M.M. and Botstein,D. (1978). Molecular genetics of bacteriophage P22. Microbiol. Rev. *42*, 385-413.

Suyama,M. and Bork,P. (2001). Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet *17*, 10-13.

Szathmary,E. (2005). Life: in search of the simplest cell. Nature *433*, 469-470.

Tamames,J. (2001). Evolution of gene order conservation in prokaryotes. Genome Biol. *2*, RESEARCH0020.

Tamames,J., Moya,A., and Valencia,A. (2007). Modular organization in the reductive evolution of protein-protein interaction networks. Genome Biol. *8*, R94.

Tamas,I., Klasson,L., Canback,B., Naslund,A.K., Eriksson,A.S., Wernegreen,J.J., Sandstrom,J.P., Moran,N.A., and Andersson,S.G. (2002). 50 million years of genomic stasis in endosymbiotic bacteria. Science *296*, 2376-2379.

Tamas,I., Klasson,L.M., Sandstrom,J.P., and Andersson,S.G. (2001). Mutualists and parasites: how to paint yourself into a (metabolic) corner. FEBS Lett. *498*, 135-139.

Tamura,K., Dudley,J., Nei,M., and Kumar,S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. *24*, 1596-1599.

Tang,J. and Moret,B.M. (2003). Scaling up accurate phylogenetic reconstruction from gene-order data. Bioinformatics. *19 Suppl 1*, i305-i312.

Teichmann,S.A. and Babu,M.M. (2004). Gene regulatory network growth by duplication. Nat. Genet *36*, 492-496.

Tekaia,F., Lazcano,A., and Dujon,B. (1999). The genomic tree as revealed from whole proteome comparisons. Genome Res. *9*, 550-557.

Tesler,G. (2002). GRIMM: genome rearrangements web server. Bioinformatics. *18*, 492-493.

Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., Crabtree,J., Jones,A.L., Durkin,A.S., Deboy,R.T., Davidsen,T.M., Mora,M., Scarselli,M., Ros,I., Peterson,J.D., Hauser,C.R., Sundaram,J.P., Nelson,W.C., Madupu,R., Brinkac,L.M., Dodson,R.J., Rosovitz,M.J., Sullivan,S.A., Daugherty,S.C., Haft,D.H., Selengut,J., Gwinn,M.L., Zhou,L., Zafar,N., Khouri,H., Radune,D., Dimitrov,G., Watkins,K., O'Connor,K.J., Smith,S., Utterback,T.R., White,O., Rubens,C.E., Grandi,G., Madoff,L.C., Kasper,D.L., Telford,J.L., Wessels,M.R., Rappuoli,R., and Fraser,C.M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. U. S. A *102*, 13950-13955.

Tettelin,H., Riley,D., Cattuto,C., and Medini,D. (2008). Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. *11*, 472-477.

Thiele,I., Vo,T.D., Price,N.D., and Palsson,B.O. (2005). Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. J. Bacteriol. *187*, 5818-5830.

Bibliography

Thomas,C.A., Jr. (1971). The genetic organization of chromosomes. Annu. Rev. Genet *5*, 237-256.

Thomas,G.H., Zucker,J., Macdonald,S.J., Sorokin,A., Goryanin,I., and Douglas,A.E. (2009). A fragile metabolic network adapted for cooperation in the symbiotic bacterium Buchnera aphidicola. BMC. Syst. Biol. *3*, 24.

Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F., and Higgins,D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. *25*, 4876-4882.

Thomson,N., Baker,S., Pickard,D., Fookes,M., Anjum,M., Hamlin,N., Wain,J., House,D., Bhutta,Z., Chan,K., Falkow,S., Parkhill,J., Woodward,M., Ivens,A., and Dougan,G. (2004). The role of prophage-like elements in the diversity of Salmonella enterica serovars. J. Mol. Biol. *339*, 279-300.

Tillier,E.R. and Collins,R.A. (2000). Genome rearrangement by replication-directed translocation. Nat. Genet *26*, 195-197.

Toh,H., Weiss,B.L., Perkin,S.A., Yamashita,A., Oshima,K., Hattori,M., and Aksoy,S. (2006). Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of Sodalis glossinidius in the tsetse host. Genome Res. *16*, 149-156.

Touchon,M. and Rocha,E.P. (2007). Causes of insertion sequences abundance in prokaryotic genomes. Mol. Biol. Evol. *24*, 969-981.

Tringe,S.G. and Rubin,E.M. (2005). Metagenomics: DNA sequencing of environmental samples. Nat. Rev. Genet *6*, 805-814.

Tringe,S.G., von,M.C., Kobayashi,A., Salamov,A.A., Chen,K., Chang,H.W., Podar,M., Short,J.M., Mathur,E.J., Detter,J.C., Bork,P., Hugenholtz,P., and Rubin,E.M. (2005). Comparative metagenomics of microbial communities. Science *308*, 554-557.

Tsuchida,T., Koga,R., Shibao,H., Matsumoto,T., and Fukatsu,T. (2002). Diversity and geographic distribution of secondary endosymbiotic bacteria in natural populations of the pea aphid, Acyrthosiphon pisum. Mol. Ecol. *11*, 2123-2135.

Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S., and Banfield,J.F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature *428*, 37-43.

Uchiyama,I. (2003). MBGD: microbial genome database for comparative analysis. Nucleic Acids Res. *31*, 58-62.

van Ham,R.C., Kamerbeek,J., Palacios,C., Rausell,C., Abascal,F., Bastolla,U., Fernandez,J.M., Jimenez,L., Postigo,M., Silva,F.J., Tamames,J., Viguera,E., Latorre,A., Valencia,A., Moran,F., and Moya,A. (2003). Reductive genome evolution in Buchnera aphidicola. Proc. Natl. Acad. Sci. U. S. A *100*, 581-586.

van Heijenoort,J. (1996). Murein synthesis. In Escherichia coli and Salmonella cellular and molecular biology, F.C.Neidhardt, ed. (Washington D.C.: American Society for Microbiology), pp. 1025-1035.

Van Sluys,M.A., de Oliveira,M.C., Monteiro-Vitorello,C.B., Miyaki,C.Y., Furlan,L.R., Camargo,L.E., da Silva,A.C., Moon,D.H., Takita,M.A., Lemos,E.G., Machado,M.A., Ferro,M.I., da Silva,F.R., Goldman,M.H., Goldman,G.H., Lemos,M.V., El-Dorry,H., Tsai,S.M., Carrer,H., Carraro,D.M., de Oliveira,R.C., Nunes,L.R., Siqueira,W.J., Coutinho,L.L., Kimura,E.T., Ferro,E.S., Harakava,R., Kuramae,E.E., Marino,C.L., Giglioti,E., Abreu,I.L., Alves,L.M., do Amaral,A.M., Baia,G.S., Blanco,S.R., Brito,M.S., Cannavan,F.S., Celestino,A.V., da Cunha,A.F., Fenille,R.C., Ferro,J.A., Formighieri,E.F., Kishi,L.T., Leoni,S.G., Oliveira,A.R., Rosa,V.E., Jr., Sassaki,F.T., Sena,J.A., de Souza,A.A., Truffi,D., Tsukumo,F., Yanai,G.M., Zaros,L.G., Civerolo,E.L., Simpson,A.J., Almeida,N.F., Jr., Setubal,J.C., and Kitajima,J.P. (2003). Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of Xylella fastidiosa. J. Bacteriol. *185*, 1018-1026.

Vandamme,P., Pot,B., Gillis,M., De,V.P., Kersters,K., and Swings,J. (1996). Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol. Rev. *60*, 407-438.

Vander Horn,P.B., Backstrom,A.D., Stewart,V., and Begley,T.P. (1993). Structural genes for thiamine biosynthetic enzymes (thiCEFGH) in Escherichia coli K-12. J. Bacteriol. *175*, 982-992.

Varma,A., Boesch,B.W., and Palsson,B.O. (1993a). Biochemical production capabilities of escherichia coli. Biotechnol. Bioeng. *42*, 59-73.

Varma,A., Boesch,B.W., and Palsson,B.O. (1993b). Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates. Appl. Environ. Microbiol. *59*, 2465-2473.

Varma,A. and Palsson,B.O. (1994a). Metabolic Flux Balancing:Basic Concepts, Scientific and Practical Use. Biotechnology *12*, 994-998.

Varma,A. and Palsson,B.O. (1994b). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Appl. Environ. Microbiol. *60*, 3724-3731.

Ventura,M., Canchaya,C., Kleerebezem,M., de Vos,W.M., Siezen,R.J., and Brussow,H. (2003). The prophage sequences of Lactobacillus plantarum strain WCFS1. Virology *316*, 245-255.

Bibliography

Vernikos,G.S. and Parkhill,J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics. *22*, 2196-2203.

Vernikos,G.S. and Parkhill,J. (2008). Resolving the structural features of genomic islands: a machine learning approach. Genome Res. *18*, 331-342.

Vollmer,W. and Bertsche,U. (2008). Murein (peptidoglycan) structure, architecture and biosynthesis in Escherichia coli. Biochim. Biophys. Acta *1778*, 1714-1734.

von Dohlen,C.D. and Moran,N.A. (2000). Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation. Biological Journal of the Linnean Society *71*, 289-717.

Vulic,M., Dionisio,F., Taddei,F., and Radman,M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc. Natl. Acad. Sci U. S. A *94*, 9763-9767.

Wagner,A. (2003). How the global structure of protein interaction networks evolves. Proc. Biol. Sci. *270*, 457-466.

Wagner,A. (2000). Robustness against mutations in genetic networks of yeast. Nat. Genet *24*, 355-361.

Wagner,A. (2006). Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. Mol. Biol. Evol. *23*, 723-733.

Wagner,A. and de la Chaux N. (2008). Distant horizontal gene transfer is rare for multiple families of prokaryotic insertion sequences. Mol. Genet. Genomics *280*, 397-408.

Wagner,A. and Fell,D.A. (2001). The small world inside large metabolic networks. Proc. Biol. Sci. *268*, 1803-1810.

Wagner,A., Lewis,C., and Bichsel,M. (2007). A survey of bacterial insertion sequences using IScan. Nucleic Acids Res. *35*, 5284-5293.

Waldor,M.K. and Mekalanos,J.J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. Science *272*, 1910-1914.

Wallace,D.C. and Morowitz,H.J. (1973). Genome size and evolution. Chromosoma *40*, 121-126.

Wang, L. S. and Warnow, T. Estimating true evolutionary distances between genomes. 637-646. 2001. ACM. STOC '01: Proceedings of the thirty-third annual ACM symposium on Theory of computing.
Ref Type: Conference Proceeding

Wang,M.D., Buckley,L., and Berg,C.M. (1987). Cloning of genes that suppress an Escherichia coli K-12 alanine auxotroph when present in multicopy plasmids. J. Bacteriol. *169*, 5610-5614.

Ward,D.M. (1998). A natural species concept for prokaryotes. Curr. Opin. Microbiol. *1*, 271-277.

Waters,E., Hohn,M.J., Ahel,I., Graham,D.E., Adams,M.D., Barnstead,M., Beeson,K.Y., Bibbs,L., Bolanos,R., Keller,M., Kretz,K., Lin,X., Mathur,E., Ni,J., Podar,M., Richardson,T., Sutton,G.G., Simon,M., Soll,D., Stetter,K.O., Short,J.M., and Noordewier,M. (2003). The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism. Proc. Natl. Acad. Sci. U. S. A *100*, 12984-12988.

Watts,D.J. and Strogatz,S.H. (1998). Collective dynamics of 'small-world' networks. Nature *393*, 440-442.

Wayne,L.G., Brenner,D.J., Colwell,R.R., Grimont,P.A.D., Kandler,O., Krychevsky,M.I., Moore,L.H., Moore,W.E.C., Murray,R.G.E., Stackebrandt,E., Starr,M.P., and Truper,H.G. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. Int. J. Syst. Bacteriol. *37*, 463-464.

Wei,J., Goldberg,M.B., Burland,V., Venkatesan,M.M., Deng,W., Fournier,G., Mayhew,G.F., Plunkett,G., III, Rose,D.J., Darling,A., Mau,B., Perna,N.T., Payne,S.M., Runyen-Janecky,L.J., Zhou,S., Schwartz,D.C., and Blattner,F.R. (2003). Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T. Infect. Immun. *71*, 2775-2786.

Weiss,B.L., Mouchotte,R., Rio,R.V., Wu,Y.N., Wu,Z., Heddi,A., and Aksoy,S. (2006). Interspecific transfer of bacterial endosymbionts between tsetse fly species: infection establishment and effect on host fitness. Appl. Environ. Microbiol. *72*, 7013-7021.

Weiss,B.L., Wu,Y., Schwank,J.J., Tolwinski,N.S., and Aksoy,S. (2008). An insect symbiosis is influenced by bacterium-specific polymorphisms in outer-membrane protein A. Proc. Natl. Acad. Sci. U. S. A *105*, 15088-15093.

Welburn,S.C., Arnold,K., Maudlin,I., and Gooday,G.W. (1993). Rickettsia-like organisms and chitinase production in relation to transmission of trypanosomes by tsetse flies. Parasitology *107 (Pt 2)*, 141-145.

Welburn,S.C. and Maudlin,I. (1992). The nature of the teneral state in Glossina and its role in the acquisition of trypanosome infection in tsetse. Ann. Trop. Med. Parasitol. *86*, 529-536.

Welburn,S.C. and Maudlin,I. (1999). Tsetse-trypanosome interactions: rites of passage. Parasitol. Today *15*, 399-403.

Bibliography

Welburn,S.C., Maudlin,I., and Ellis,D.S. (1987). In vitro cultivation of rickettsia-like-organisms from Glossina spp. Ann. Trop. Med. Parasitol. *81*, 331-335.

Welburn,S.C., Maudlin,I., and Ellis,D.S. (1989). Rate of trypanosome killing by lectins in midguts of different species and strains of Glossina. Med. Vet. Entomol. *3*, 77-82.

Welch,R.A., Burland,V., Plunkett,G., III, Redford,P., Roesch,P., Rasko,D., Buckles,E.L., Liou,S.R., Boutin,A., Hackett,J., Stroud,D., Mayhew,G.F., Rose,D.J., Zhou,S., Schwartz,D.C., Perna,N.T., Mobley,H.L., Donnenberg,M.S., and Blattner,F.R. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc. Natl. Acad. Sci. U. S. A *99*, 17020-17024.

Wellner,A., Lurie,M.N., and Gophna,U. (2007). Complexity, connectivity, and duplicability as barriers to lateral gene transfer. Genome Biol. *8*, R156.

Wernegreen,J.J. (2002). Genome evolution in bacterial endosymbionts of insects. Nat. Rev. Genet. *3*, 850-861.

Wernegreen,J.J. (2004). Endosymbiosis: lessons in conflict resolution. PLoS. Biol. *2*, E68.

Wernegreen,J.J. (2005). For better or worse: genomic consequences of intracellular mutualism and parasitism. Curr. Opin. Genet. Dev. *15*, 572-583.

Werren,J.H., Baldo,L., and Clark,M.E. (2008). Wolbachia: master manipulators of invertebrate biology. Nat. Rev. Microbiol. *6*, 741-751.

Werren,J.H. and Bartos,J.D. (2001). Recombination in Wolbachia. Curr. Biol. *11*, 431-435.

Whalen,W.A. and Berg,C.M. (1982). Analysis of an avtA::Mu d1(Ap lac) mutant: metabolic role of transaminase C. J. Bacteriol. *150*, 739-746.

Whitfield,C. (2006). Biosynthesis and assembly of capsular polysaccharides in Escherichia coli. Annu. Rev. Biochem. *75*, 39-68.

Whittam,T.S., Ochman,H., and Selander,R.K. (1983a). Geographic components of linkage disequilibrium in natural populations of Escherichia coli. Mol. Biol. Evol. *1*, 67-83.

Whittam,T.S., Ochman,H., and Selander,R.K. (1983b). Multilocus genetic structure in natural populations of Escherichia coli. Proc. Natl. Acad. Sci U. S. A *80*, 1751-1755.

Wiback,S.J., Famili,I., Greenberg,H.J., and Palsson,B.O. (2004). Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. J. Theor. Biol. *228*, 437-447.

Wicker,C. (1983). Differential vitamin and choline requirements of symbiotic and aposymbiotic s. oryzae (coleoptera: curculionidae). Comparative Biochemistry and Physiology Part A: Physiology *76*, 177-182.

Wilcox,J.L., Dunbar,H.E., Wolfinger,R.D., and Moran,N.A. (2003). Consequences of reductive evolution for gene expression in an obligate endosymbiont. Mol. Microbiol. *48*, 1491-1500.

Wilson,A.C., Carlson,S.S., and White,T.J. (1977). Biochemical evolution. Annu. Rev. Biochem. *46*, 573-639.

Wissenbach,U., Six,S., Bongaerts,J., Ternes,D., Steinwachs,S., and Unden,G. (1995). A third periplasmic transport system for L-arginine in Escherichia coli: molecular characterization of the artPIQMJ genes, arginine binding and transport. Mol. Microbiol. *17*, 675-686.

Woese,C.R. (1987). Bacterial evolution. Microbiol. Rev. *51*, 221-271.

Woese,C.R., Kandler,O., and Wheelis,M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U. S. A *87*, 4576-4579.

Woese,C.R., Olsen,G.J., Ibba,M., and Soll,D. (2000). Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol. Mol. Biol. Rev. *64*, 202-236.

Wolf,Y.I., Aravind,L., Grishin,N.V., and Koonin,E.V. (1999). Evolution of aminoacyl-tRNA synthetases--analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res. *9*, 689-710.

Wolf,Y.I., Rogozin,I.B., Grishin,N.V., and Koonin,E.V. (2002). Genome trees and the tree of life. Trends Genet *18*, 472-479.

Wolfe,K.H. and Shields,D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. Nature *387*, 708-713.

Wood,J.M. (1987). Membrane association of proline dehydrogenase in Escherichia coli is redox dependent. Proc. Natl. Acad. Sci U. S. A *84*, 373-377.

Woods,S.A., Schwartzbach,S.D., and Guest,J.R. (1988). Two biochemically distinct classes of fumarase in Escherichia coli. Biochim. Biophys. Acta *954*, 14-26.

Bibliography

Wu,D., Daugherty,S.C., Van Aken,S.E., Pai,G.H., Watkins,K.L., Khouri,H., Tallon,L.J., Zaborsky,J.M., Dunbar,H.E., Tran,P.L., Moran,N.A., and Eisen,J.A. (2006). Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS. Biol. *4*, e188.

Wu,M., Sun,L.V., Vamathevan,J., Riegler,M., Deboy,R., Brownlie,J.C., McGraw,E.A., Martin,W., Esser,C., Ahmadinejad,N., Wiegand,C., Madupu,R., Beanan,M.J., Brinkac,L.M., Daugherty,S.C., Durkin,A.S., Kolonay,J.F., Nelson,W.C., Mohamoud,Y., Lee,P., Berry,K., Young,M.B., Utterback,T., Weidman,J., Nierman,W.C., Paulsen,I.T., Nelson,K.E., Tettelin,H., O'Neill,S.L., and Eisen,J.A. (2004). Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: a streamlined genome overrun by mobile genetic elements. PLoS. Biol. *2*, E69.

Xi,J., Ge,Y., Kinsland,C., McLafferty,F.W., and Begley,T.P. (2001). Biosynthesis of the thiazole moiety of thiamin in Escherichia coli: identification of an acyldisulfide-linked protein--protein conjugate that is functionally analogous to the ubiquitin/E1 complex. Proc. Natl. Acad. Sci U. S. A *98*, 8513-8518.

Xu,M., Struck,D.K., Deaton,J., Wang,I.N., and Young,R. (2004). A signal-arrest-release sequence mediates export and control of the phage P1 endolysin. Proc. Natl. Acad. Sci. U. S. A *101*, 6415-6420.

Yang,F., Yang,J., Zhang,X., Chen,L., Jiang,Y., Yan,Y., Tang,X., Wang,J., Xiong,Z., Dong,J., Xue,Y., Zhu,Y., Xu,X., Sun,L., Chen,S., Nie,H., Peng,J., Xu,J., Wang,Y., Yuan,Z., Wen,Y., Yao,Z., Shen,Y., Qiang,B., Hou,Y., Yu,J., and Jin,Q. (2005). Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic Acids Res. *33*, 6445-6458.

Yang,Y., Tsui,H.C., Man,T.K., and Winkler,M.E. (1998a). Identification and function of the pdxY gene, which encodes a novel pyridoxal kinase involved in the salvage pathway of pyridoxal 5'-phosphate biosynthesis in Escherichia coli K-12. J. Bacteriol. *180*, 1814-1821.

Yang,Y., Zhao,G., Man,T.K., and Winkler,M.E. (1998b). Involvement of the gapA- and epd (gapB)-encoded dehydrogenases in pyridoxal 5'-phosphate coenzyme biosynthesis in Escherichia coli K-12. J. Bacteriol. *180*, 4294-4299.

Yang,Y., Zhao,G., and Winkler,M.E. (1996). Identification of the pdxK gene that encodes pyridoxine (vitamin B6) kinase in Escherichia coli K-12. FEMS Microbiol. Lett. *141*, 89-95.

Yang,Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586-1591.

Yang,Z. and Nielsen,R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. *17*, 32-43.

Yoon,S.H., Hur,C.G., Kang,H.Y., Kim,Y.H., Oh,T.K., and Kim,J.F. (2005). A computational approach for identifying pathogenicity islands in prokaryotic genomes. BMC. Bioinformatics. *6*, 184.

Yus,E., Maier,T., Michalodimitrakis,K., van,N., V, Yamada,T., Chen,W.H., Wodke,J.A., Guell,M., Martinez,S., Bourgeois,R., Kuhner,S., Raineri,E., Letunic,I., Kalinina,O.V., Rode,M., Herrmann,R., Gutierrez-Gallego,R., Russell,R.B., Gavin,A.C., Bork,P., and Serrano,L. (2009). Impact of genome reduction on bacterial metabolism and its regulation. Science *326*, 1263-1268.

Zhang,J. and He,X. (2005). Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol. Biol. Evol. *22*, 1147-1155.

Zhang,W., Zhou,Y., and Becker,D.F. (2004). Regulation of PutA-membrane associations by flavin adenine dinucleotide reduction. Biochemistry *43*, 13165-13174.

Zhang,Y., Thiele,I., Weekes,D., Li,Z., Jaroszewski,L., Ginalski,K., Deacon,A.M., Wooley,J., Lesley,S.A., Wilson,I.A., Palsson,B., Osterman,A., and Godzik,A. (2009). Three-dimensional structural view of the central metabolic network of Thermotoga maritima. Science *325*, 1544-1549.

Zhao,G., Pease,A.J., Bharani,N., and Winkler,M.E. (1995). Biochemical characterization of gapB-encoded erythrose 4-phosphate dehydrogenase of Escherichia coli K-12 and its possible role in pyridoxal 5'-phosphate biosynthesis. J. Bacteriol. *177*, 2804-2812.

Zhaxybayeva,O., Lapierre,P., and Gogarten,J.P. (2004). Genome mosaicism and organismal lineages. Trends Genet. *20*, 254-260.

Zhaxybayeva,O., Lapierre,P., and Gogarten,J.P. (2005). Ancient gene duplications and the root(s) of the tree of life. Protoplasma *227*, 53-64.

Zhu,N., Olivera,B.M., and Roth,J.R. (1991). Activity of the nicotinamide mononucleotide transport system is regulated in Salmonella typhimurium. J. Bacteriol. *173*, 1311-1320.

Zientz,E., Dandekar,T., and Gross,R. (2004). Metabolic interdependence of obligate intracellular bacteria and their insect hosts. Microbiol. Mol. Biol. Rev. *68*, 745-770.

Zientz,E., Silva,F.J., and Gross,R. (2001). Genome interdependence in insect-bacterium symbioses. Genome Biol. *2*, REVIEWS1032.

Zimmer,D.P., Soupene,E., Lee,H.L., Wendisch,V.F., Khodursky,A.B., Peter,B.J., Bender,R.A., and Kustu,S. (2000). Nitrogen regulatory protein C-controlled genes of Escherichia coli: scavenging as a defense against nitrogen limitation. Proc. Natl. Acad. Sci. U. S. A *97*, 14674-14679.

Bibliography

Zuckerkandl,E. and Pauling,L. (1965). Molecules as documents of evolutionary history. J. Theor. Biol. *8*, 357-366.

9 Breve resumen en castellano

Breve resumen en castellano

CAPÍTULO 1: Introducción general

El principal objetivo de esta tesis doctoral se centra en el estudio de las dinámicas de evolución genómica en procariotas, y de forma más específica en el proceso de evolución reductiva que experimentan bacterias endosimbiontes de insectos a lo largo de su evolución asociada al hospedador. Para ello se han utilizado dos aproximaciones metodológicas principales: la genómica comparada y la biología de sistemas. La primera ha permitido estudiar la evolución de la organización genómica en γ-proteobacterias, donde se incluyen la mayoría de endosimbiontes bacterianos con genoma completamente secuenciado, mientras que la segunda ha permitido estudiar todo el proceso de evolución reductiva en *Sodalis glossinidius*, el endosimbionte secundario de la mosca tsé-tsé que se encuentra en etapas muy tempranas de este proceso. Para ello se ha llevado a cabo una re-anotación exhaustiva de este genoma con el objetivo de determinar el impacto de la proliferación de diferentes tipos de elementos genéticos móviles y del proceso de inactivación génica masiva que está teniendo lugar en este genoma sobre sus capacidades funcionales considerando tanto *S. glossinidius* de forma individual así como el consorcio endosimbionte completo que se establece en la mosca tsé-tsé y que incluye el endosimbionte primario ancestral *Wigglesworthia glossinidia*.

El análisis evolutivo a escala genómica ha sido posible gracias al incremento exponencial de genomas completamente secuenciados asociado con el desarrollo de técnicas de secuenciación masiva cada vez mas baratas y rápidas, y si hay un grupo de organismos en los que el análisis genómico se ha beneficiado de manera especialmente significativa de estos avances ese es el de los procariotas. En etapas iniciales a partir de la secuenciación del primer genoma bacteriano completo, *Haemophilus influenzae* (Fleischmann et al., 1995), el objetivo era secuenciar genomas de bacterias modelo como *Escherichia coli* o *Bacillus subtilis* así como de determinadas bacterias patógenas causantes de enfermedades de interés mundial como *Mycoplasma pneumoniae* M129, *Helycobacter pylori*, o *Borrelia burgdorferi*, donde el conjunto de genes presentes en sus genomas permitió inferir sus mecanismos de patogenicidad así como sus capacidades funcionales diferenciales . Sin embargo, el desarrollo de plataformas de secuenciación masiva ha supuesto una revolución en el campo de la genómica que ha permitido abordar estudios de carácter metagenómico donde el objetivo es secuenciar comunidades bacterianas completas características de determinados ambientes, lo que permite estudiar la estructura poblacional de las comunidades en términos de diversidad de especies, abundancia relativa de las mismas, y potencial funcional en términos de capacidades metabólicas (Tringe and Rubin, 2005). La secuenciación de diferentes cepas o individuos de una misma especie o una misma población permite también abordar estudios de genómica poblacional, donde las poblaciones bacterianas y sus individuos se caracterizan por propiedades genómicas como la presencia o ausencia

de genes o variaciones en los parámetros de diversidad nucleotídica a lo largo de los genomas (Fraser-Liggett, 2005). Asimismo, la disponibilidad de genomas completos con diferentes grados de proximidad evolutiva (cepas de una misma especie conjuntamente con especies evolutivamente próximas) ha permitido comparar las relaciones evolutivas inferidas a partir de marcadores evolutivos clásicos como el gen 16S ribosomal con relaciones evolutivas inferidas a partir de caracteres de similitud a escala genómica (Konstantinidis et al., 2006).

Una de las conclusiones más importantes que se han extraído a partir de la comparación de múltiples genomas microbianos es la gran diversidad existente en procariotas a todos los niveles de la escala evolutiva tanto a nivel de tamaños genómicos como de contenido génico. Actualmente, con más de 800 genomas bacteriano completamente secuenciados, el rango de tamaños genómicos en procariotas oscila entre las 0,16 kilobases del endosimbionte primario de psílidos *Carsonella rudii* (Nakabachi et al., 2006) a las 13.33 megabases dela mixobacteria de vida libre *Sorangium cellulosum* (Schneiker et al., 2007). A pesar de estas variaciones en el tamaño genómico, la mayoría de bacterias presentan densidades codificantes similares, con un gen por kilobase aproximadamente en contraste con la predominancia de DNA no codificante característica de genomas eucariotas, de tal forma que existe una clara correlación positiva entre tamaño de genoma y número de genes en procariotas (Figura 1.1). Bacterias de genomas de gran tamaño se corresponden con bacterias de vida libre que viven en ambientes altamente fluctuantes en los que el organismo necesita un amplio repertorio de capacidades funcionales para sobrevivir, de tal forma que estos genomas aparecen enriquecidos en genes implicados en regulación y actividades del metabolismo secundario. Por el contrario, los genomas más reducidos aparecen asociados con bacterias que viven en asociación con diversos huéspedes eucariotas, con los cuales se establecen relaciones tanto de tipo patógeno como simbionte, en las cuales las bacterias habitan nichos ecológico muy estables con muy pocas fluctuaciones (Konstantinidis and Tiedje, 2004). Sin embargo, estas variaciones no son exclusivas de niveles taxonómicos elevados, ya que variaciones importantes en tamaño genómico y número de genes se observan también incluso entre diferentes cepas de una misma especie. Un claro ejemplo lo encontramos al comparar el contenido génico de tres cepas de *E. coli* patogénicas (CFT073 y EDL933) y no patogénicas (MG1655), donde se vio que los tres genomas compartían únicamente el 40% del total de genes presentes en las tres cepas, mientras que la mayoría del resto de genes se correspondía con genes específicos de cada genoma que eran responsables de sus capacidades funcionales diferenciales y que explicaban las diferencias ecológicas existentes entre ellos (Welch et al., 2002). Esto hace que en un genoma particular se pueda diferenciar entre el endogenoma constituido por genes esenciales para el funcionamiento de cualquier sistema celular, como pueden ser genes implicados en la maquinaria de replicación y transcripción o en funciones metabólicas básicas, que aparecen presentes en todos los genomas de cepas o especies próximas, y un

exogenoma en el que se incluyen aquellos genes específicos de un determinado individuo así como genes con presencia/ausencia variable entre genomas próximos (Casjens, 1998).

La caracterización de ambos subconjuntos de genes en diferentes especies de bacterias es un tópico recurrente en estudios de evolución genómica, y resulta esencial para comprender la evolución bacteriana en su conjunto más allá de las propiedades individuales de genomas particulares. Mientras que la caracterización del endogenoma o *core* genómico a diferentes escalas evolutivas permite aproximarnos al contenido génico de los ancestros evolutivos de las especies bacterianas actuales, siendo clave para el desarrollo de aproximaciones filogenómicas al estudio de la evolución bacteriana capaces de superar las limitaciones asociadas a los estudios evolutivos basados en genes individuales (Comas et al., 2007; Delsuc et al., 2005), el conjunto de genes accesorios o auxiliares que residen en el exogenoma es responsable de la enorme diversidad ecológica que existe entre especies procariotas ya que incluye genes implicados en funciones celulares que, aunque no sean esenciales para la supervivencia de la especie, proporcionan importantes ventajas selectivas a los individuos que los poseen a nivel de capacidad de adaptación a un determinado nicho ecológico, de colonización de un determinado hospedador, o de resistencia a antibióticos (Abby and Daubin, 2007). Sin embargo, una proporción significativa del exogenoma se corresponde con diferentes tipos de elementos genético móviles que proliferan a lo largo de los genomas y que tienen una influencia decisiva en la evolución de genomas bacterianos debido a su efecto sobre la estructura y organización de los genomas (Rocha, 2008b). La enorme plasticidad en el contenido génico que se observa entre genomas evolutivamente próximos es consecuencia de diferentes procesos de ganancia y pérdida de genes que tienen lugar a lo largo de la evolución de diferentes linajes bacterianos, y que hacen que un genoma individual sean una representación incompleta del contenido génico de la especie procariota a la que pertenece, al cual se hace referencia con el concepto de pangenoma, que se define como el conjunto de genes que conforman todos los genomas una determinada especie (Medini et al., 2005). Las diferencias en el tamaño y contenido de los pangenomas de diferentes especies son un reflejo de las diferencias en sus perfiles ecológicos, pudiendo encontrar pangenomas cerrados característicos de especies que habitan ambientes restringidos y estables donde se encuentran aislados respecto a otras poblaciones bacterianas, con pocas oportunidades de adquirir material genético a través de transferencia genética horizontal, siendo un claro ejemplo los genomas de diferentes cepas del endosimbionte primario de pulgones *Buchnera aphidicola*, con pangenomas cerrados extremadamente estables consecuencia de su estricta localización en el interior de células especializadas del pulgón donde no tienen oportunidad de adquirir material genético desde otras especies bacterianas (Tamas et al., 2002). Por el contrario, especies bacterianas funcionalmente más versátiles capaces de colonizar diferentes ambientes y con múltiples mecanismos para

intercambiar material genético con especies de su entorno mediante diferentes tipos de elementos genéticos móviles tienen pangenomas abiertos, que aumentan en tamaño de forma significativa con cada genoma que se incorpora al conjunto de la especie, siendo un ejemplo claro el pangenoma de *Streptococcus agalactiae* donde simulaciones matemáticas indican que el tamaño de su pangenoma incrementa en un promedio de 33 genes por cada genoma incorporado (Tettelin et al., 2005).

Esta enorme plasticidad en el contenido génico es consecuencia del papel predominante que tienen los mecanismos de ganancia y pérdida génica sobre la evolución de procariotas. La evolución de genomas procariotas es consecuencia de la adquisición de genes selectivamente beneficiosos o de diferentes tipos de elementos genéticos móviles a través de sucesos de duplicación génica y transferencia genética horizontal, combinados con la pérdida de genes no esenciales consecuencia de sucesos de mutación, deleción, y deriva genética en poblaciones bacterianas de tamaño efectivo reducido, de tal forma que el contenido génico de genomas procariotas es un reflejo de las diferentes presiones selectivas que actúan en diferentes momentos de su trayectoria evolutiva. El papel de cada uno de estos factores en la evolución de diferentes linajes bacterianos ha sido revelado a través de la genómica comparada, como la importancia de la duplicación génica en la evolución de *Vibrio cholerae* (Heidelberg et al., 2000), la importancia de la transferencia genética horizontal en linajes bacterianos como las Xanthomonadales (Comas et al., 2006), y de manera específicamente relevante en el contexto de esta tesis el papel predominante de sucesos de pérdida génica en la evolución de genomas de bacterias que establecen asociaciones de tipo endosimbionte o patógena con diferentes huéspedes eucariotas (Silva and Latorre, 2008; Moya et al., 2008). En todos estos procesos, la proliferación de elementos genéticos móviles tienen una influencia decisiva tanto para la inserción de material genético como por su papel en sucesos de deleción, actuando como puntos de recombinación que también generan cambios en la estructura del genoma mediante reordenaciones cromosómicos como inversiones y translocaciones (Rocha, 2004). La genómica comparada ha revelado una clara correlación entre la presencia de elementos genéticos móviles y el nivel de estabilidad genómica en genomas procariotas que se explica por las presiones selectivas diferenciales que afectan a la evolución de diferentes linajes bacterianos. En un extremo del espectro, bacterias de vida libre con genomas de tamaños elevados presentan gran variabilidad en contenido y orden génico asociados a la presencia de elementos genéticos móviles dispersos a lo largo del genoma, mientras que en el otro extremo se encuentran los endosimbiontes bacterianos de insectos, que presentan genomas altamente reducidos y estables asociados con una ausencia prácticamente total de elementos genéticos móviles (Bentley and Parkhill, 2004). Estos linajes endosimbiontes con genomas altamente reducidos han evolucionado a partir de bacterias de vida libre mediante un proceso de pérdida génica masiva que ha tenido lugar repetidas veces a lo largo de la evolución de diferentes linajes de insectos, y cuyo estudio es uno de los principales objetivos de la presente tesis

doctoral, centrado de manera específica en el linaje de *S. glossinidius*, el endosimbionte secundario de la mosca tsé-tsé, que con un genoma de más de 4 megabases se encuentra en una etapa muy inicial de este proceso de evolución reductiva.

En los siguientes apartados de esta introducción se van a tratar los diferentes tipos de elementos genéticos móviles presentes en genomas procariotas, el papel de los procesos de ganancia y pérdida génica en diferentes etapas del proceso de evolución reductiva desde bacterias de vida libre a bacterias endosimbiontes obligadas, y como los avances en la biología de sistemas permiten una nueva aproximación a este proceso de evolución reductiva integrando la información genómica en un contexto sistémico que nos permite evaluar de forma cuantitativa la funcionalidad global de sistemas biológicos.

Elementos genéticos móviles como mediadores de plasticidad genómica

Una fracción significativa del conjunto de genes variables en genomas procariotas se corresponde con diferentes tipos de elementos genéticos móviles, que se definen como segmentos de DNA que contienen genes responsables de llevar a cabo su propia transferencia tanto dentro de un mismo genoma como ente genomas. El conjunto de elementos genéticos móviles presente en un genoma concreto recibe el nombre de mobiloma, y consiste básicamente de secuencias de inserción (IS) y transposones, bacteriófagos, plásmidos y diferentes tipos de islas genómicas (Frost et al., 2005). Mediante genómica comparada es posible identificar huellas de la actividad de los diferentes elementos genéticos móviles sobre la estructura y capacidades funcionales codificadas en un determinado genoma. Por ejemplo, la presencia de diferentes copias de un mismo IS o elemento fágico a lo largo de un genoma es una fuente potencial de reordenaciones genómicas por recombinación homóloga entre copias de estas secuencias repetidas. En este caso, la orientación de las secuencias repetidas determina el tipo de reordenación que se produce, de tal forma que si ambas repeticiones se encuentran en la misma orientación la recombinación homóloga lleva a la deleción del segmento de DNA comprendido entre ambas, mientras que si las repeticiones se encuentran en orientación invertida la recombinación homóloga lleva a la inversión del segmento de DNA comprendido entre ellas. Esto se ha observado por ejemplo al comparar diferentes cepas de la bacteria patógena de plantas *Xylella fastidiosa*, donde cinco de los 6 sitios de recombinación responsables de tres inversiones cromosómicas están localizados en profagos duplicados, mientras que en la especie próxima *Xanthomonas campestris*, diferentes tipos de IS han proliferado en diferentes cepas (Monteiro-Vitorello et al., 2005), así como en el patógeno *Streptococcus pyogenes*, donde recombinación homóloga entre profagos duplicados es responsable de una inversión cromosómica

de 1 megabase alrededor del origen de replicación (Nakagawa et al., 2003). No obstante, muchos de las reordenación cromosómicas son eliminadas por selección natural porque tienen efectos deletéreos tanto a nivel de la posible deleción de genes funcionalmente importantes como por la alternación de la organización cromosómica en las regiones reordenadas (Rocha, 2004). En este contexto, la proliferación de elementos genéticos móviles, especialmente IS, es masiva en bacterias que han sufrido una transición reciente a nichos ecológicos restringidos o especializados desde un ancestro ecológicamente más versátil, como es el caso de bacterias que establecen asociaciones obligadas simbióticas o patogénicas con determinados huéspedes eucariotas desde ancestros de vida libre. Un ejemplo es el incremento progresivo en el número de IS que se produce en β-proteobacterias patógenas del género *Bordetella* desde un patógeno como *Bordetella bronchiseptica* que coloniza el tracto respiratorio de una gran variedad de mamíferos y que contiene cero copias de IS hasta un patógeno como *Bordetella pertusis* restringido exclusivamente a humanos y donde se identifican más de 260 copias. En este último caso, la proliferación de IS aparece asociada a la generación de reordenaciones genómicas y deleciones responsables de la reducción genómica que se produce asociada a esta transición evolutiva (Parkhill et al., 2003). Por el contrario, los genomas mínimos de bacterias endosimbiontes de insectos con asociaciones ancestrales con su correspondiente hospedador carecen de este tipo de elementos genéticos móviles, que se han perdido a lo largo de su evolución desde la asociación inicial con el insecto huésped (Shigenobu et al., 2000; Tamas et al., 2002; Akman et al., 2002; Gil et al., 2003; van Ham et al., 2003; Degnan et al., 2005; Perez-Brocal et al., 2006). Estudios comparativos entre genomas de especies próximas revelan una gran homogeneidad entre las copias de diferentes IS de un mismo tipo o familia dentro de cada genoma, así como una gran variabilidad a nivel del tipo de IS presentes y su abundancia entre genomas incluso entre cepas de una misma especie, que se ha explicado por sucesivos ciclos de expansión y extinción donde los IS proliferan rápidamente después de su transferencia horizontal al genoma receptor hasta un punto en el que la eficacia del hospedador se ve comprometida, lo que hace que se eliminen del genoma (Wagner, 2006; Wagner and de la Chaux N., 2008). Esta proliferación de IS es mayor en bacterias con genomas de gran tamaño en las cuales la densidad de sitios de inserción potencialmente deletéreos (genes esenciales donde la inserción de un IS tiene consecuencias letales para la viabilidad de la bacteria) es menor que en bacterias con genomas reducidos, y además se ve incrementada en bacterias que sufren una transición reciente a un modo de vida dependiente de hospedador debido a la disminución en la presión selectiva sobre un gran número de genes que pasan a ser no esenciales en el nicho ecológico asociado al hospedado (Moran and Plague, 2004; Touchon and Rocha, 2007).

Además de su efecto sobre la estructura y organización de los genomas, los elementos genéticos móviles son importantes agentes en la transferencia horizontal de genes, siendo responsables de muchas actividades especificas de una determinada

especie o cepa. En este contexto, los bacteriófagos juegan un papel esencial en la evolución y virulencia de un gran número de bacterias patógenas y son una fuente importante de variabilidad en el contenido génico de genomas bacterianos, pudiendo representar más del 50% de contenido génico específico de un gran número de bacterias patógenas (Casjens, 2003; Hatfull, 2008). De hecho, los bacteriófagos son considerados como las entidades más abundantes y diversas de la naturaleza, siendo estimados como 10 veces más abundantes que los procariotas en muestras ambientales y con una enorme diversidad que se comienza a revelar por estudios de metagenómica viral que sugiere que una fracción muy significativa de la diversidad fágica permanece por explorar (Breitbart et al., 2002). Sin embargo, desde un punto de vista evolutivo, los profagos que aparecen integrados en los genomas procariotas son elementos transitorios que en última instancia conducen a la muerte de la célula hospedadora debido a su actividad lítica potencial como virus procariotas, a pesar de los posibles beneficios que puedan inducir en un periodo de tiempo inmediato, de tal forma que muchos de estos profagos aparecen inactivados por diferentes sucesos mutacionales que conducen a su eliminación final por sucesos delecionales, siendo habitual observar la presencia de remanentes fágicos en genomas bacterianos que se encontrarían en vías de eliminación (Lawrence et al., 2001).

Este es el caso también de los diferentes tipos de islas genómicas que proliferan en genomas bacterianos que albergan genes codificantes de actividades específicas capaces de mediar adaptación a determinados nichos ecológicos, como genes de resistencia a antibióticos, o genes capaces de inducir patogenicidad o capaces de facilitar interacciones ecológicas entre bacterias y huéspedes eucariotas. Un claro ejemplo lo encontramos en la capacidad de fijación de nitrógeno en bacterias simbiontes de plantas del género *Rhizobium* que reside en una isla genómica que contiene los genes responsables de la colonización de tejidos de raíz por parte de la bacteria y los genes responsables de fijación de nitrógeno a amonio (Sullivan and Ronson, 1998), o en las islas genómicas presentes en diferentes cepas de *Staphilococcus aureus* que contienen genes de resistencia a meticilina que incrementan su supervivencia en hospitales o en ambientes donde coexisten con otras bacterias productoras de antibióticos (Hiramatsu et al., 2002). En el contexto de bacterias simbiontes o patógenas que viven en asociación con huéspedes eucariotas, dicha interacción depende en muchos casos de la presencia de diferentes tipos de sistemas de secreción codificados por islas genómicas que median la interacción física entre bacteria y célula huésped así como la transferencia de diferentes moléculas efectoras que determinan el tipo de interacción entre bacteria y hospedador (Hueck, 1998; Dobrindt et al., 2004). Un ejemplo lo encontramos en *Salmonella enterica*, donde dos islas de patogenicidad diferentes codificantes de sistemas de secreción tipo III son responsables de diferentes etapas de la infección de células del hospedador, de tal forma que una de ellas es responsable de la interacción inicial e invasión de células epiteliales del huésped mientras que la otra es necesaria en etapas posteriores de la infección una vez la bacteria ha colonizado

el citoplasma de la célula hospedadora (Ochman and Groisman, 1996; Galan, 2001; Kuhle and Hensel, 2004). Estas mismas islas genómicas que confieren un carácter patogénico *S. enterica* se encuentran también presentes conservando la misma estructura en *S. glossinidius*, donde desempeñan un papel similar aunque sin conferir carácter patógeno a la bacteria (Dale et al., 2001; Dale et al., 2005), lo que representa un ejemplo de la plasticidad de este tipo de elementos genéticos móviles dependiendo del contexto ecológico de la bacteria que los alberga en el sentido que una misma isla genómica responsable del carácter patógeno de *S. enterica* actúa como una isla de eficacia en *S. glossinidius* favoreciendo su supervivencia en un contexto simbiótico.

En el contexto de esta tesis, uno de los objetivos que nos hemos planteado consiste en determinar el impacto de la proliferación de diferentes tipos de elementos genéticos móviles sobre la organización del genoma y las capacidades funcionales de *S. glossinidius*, el endosimbionte secundario de la mosca tsé-tsé que, en contra de los endosimbiontes mutualistas con asociaciones ancestrales con sus insectos hospedador, presenta una muy reciente transición al modo de vida dependiente de huésped que se refleja por su tamaño genómico próximo al de una bacteria de vida libre como *E. coli*, sin sesgos composicionales hacia AT típicos de bacterias mutualistas obligadas, y en el que se está produciendo un proceso de inactivación génica masiva que hace que únicamente el 50% de su genoma sea codificante (Toh et al., 2006).

Evolución genómica por ganancia de genes: Duplicación génica y transferencia genética horizontal

Los principales mecanismos responsables del incremento de tamaño en genomas bacterianos son la duplicación génica y la adquisición de DNA exógeno mediante sucesos de transferencia genética horizontal, donde se incluyen la transferencia de los diferentes tipos de elementos genéticos móviles descritos anteriormente. Asimismo, la distinción entre genes ortólogos que evolucionan verticalmente por especiación frente genes parálogos que evolucionan por duplicación génica seguido de divergencia entre copias y genes xenólogos originados a partir de sucesos de transferencia genética horizontal desde otra bacteria es esencial a la hora de entender la evolución de los procariotas (Zhaxybayeva et al., 2004; Koonin, 2005). En el caso de la duplicación génica, su importancia evolutiva en la generación de nuevas funciones génicas fue postulada por Ohno en 1970 en su trabajo seminal *Evolution by gene duplication*, donde se propone que el gen duplicado se encuentra libre de presiones selectivas para el mantenimiento de función de tal forma que puede generar nuevas funciones génicas por acumulación de mutaciones (Ohno, 1970). Sin embargo, mientras en eucariotas el papel esencial de la duplicación génica en la

generación de nuevas funciones génicas es mayoritario, en procariotas su importancia es considerada menor debido al papel predominante de la transferencia genética horizontal en la adquisición de nuevas funciones génicas. No obstante, familias de genes parálogos originados por duplicación génica se han identificado en la mayoría de genomas procariotas secuenciados, existiendo una clara correlación positiva entre el tamaño del genoma y el número de genes duplicados, de tal forma que los genomas reducidos de bacterias endosimbiontes obligadas presentan el menor número de genes duplicados mientras que en genomas de gran tamaño como *Streptomyces coelicor* la fracción de genes parálogos llega hasta el 50% del genoma (Jordan et al., 2001; Gevers et al., 2004; Pushker et al., 2004). Estudios similares revelan que no todas las familias génicas son igualmente proclives a sucesos de duplicación, pudiendo encontrar desde genes implicados en funciones metabólicas que presentan una correlación linear entre tamaño de genoma y número de copias del gen hasta genes implicados en regulación de la expresión génica y genes asociados a sucesos de transferencia genética horizontal, que aparecen ausentes en genomas de tamaño reducido mientras que en genomas de tamaños elevados forman familias génicas de gran tamaño y complejidad (Hooper and Berg, 2003a; Hooper and Berg, 2003b; Ranea et al., 2004). La expansión diferencial de diferentes familias génicas en diferentes bacterias se ha postulado como un mecanismo adaptativo que permite la rápida sobreexpresión de determinados genes que pueden ser necesarios para la supervivencia bacteriana bajo determinadas presiones ambientales como ausencia de nutrientes (Sonti and Roth, 1989), presencia de antibióticos (Nichols and Guay, 1989), o adaptación a temperaturas elevadas (Riehle et al., 2001), siendo los elementos genéticos móviles importantes mediadores de estos sucesos de duplicación (Romero and Palacios, 1997). Incluso en los genomas altamente reducidos de endosimbiontes bacterianos de insectos se detectan sucesos de amplificación génica relacionados con diferentes aspectos de la interacción ecológica con el insecto hospedador, como es el caso de la amplificación de genes codificantes de la enzima antranilato sintasa implicada en la síntesis de triptófano en plásmidos del endosimbionte primario de pulgones *B. aphidicola* en determinadas especies que presentan alta demanda de este aminoácido (Rouhbakhsh et al., 1996; Lai et al., 1996). Sin embargo, la inactivación de genes duplicados es un proceso común en la mayoría de genomas de bacterias que se encuentran en un proceso de reducción genómica bien a través de inserción de elementos genéticos móviles como en la evolución de *Shigella flexneri* a partir de *E. coli* (Jin et al., 2002; Wei et al., 2003) o a través de sucesos de inactivación génica masiva como en la evolución de *Myocobacterium leprae* desde su divergencia de *Mycobacterium tuberculosis* (Cole, 1998), de tal forma que no solo incrementos sino también reducciones en el tamaño genómico se explican parcialmente por expansiones o reducciones en familias génicas.

Sin embargo, la capacidad de las bacterias de intercambiar material genético tanto entre especies como entre cepas de una misma especie a través de sucesos de

transferencia genética horizontal es probablemente la mayor fuente de variabilidad genética en bacterias y la responsable de gran parte de la variación en contenido génico entre bacterias. La importancia de la transferencia genética horizontal en la evolución de bacterias ya se consideraba antes del desarrollo de la genómica moderna, aunque inicialmente se consideraba un fenómeno residual que tenía lugar bajo determinadas presiones selectivas dentro del modelo clásico de evolución clonal de bacterias (Levin, 1981; Cohan, 2001). Esta visión tradicional de la evolución de bacterias cambia a medida que el análisis evolutivo molecular empezó a demostrar que diferentes genes de una misma especie presentan diferentes historias evolutivas en reconstrucciones filogenéticas (DuBose et al., 1988; Dykhuizen and Green, 1991) y que la recombinación entre linajes bacterianos es mucho más común de lo que se consideró inicialmente en base a la ausencia de reproducción sexual en bacterias (Feil and Spratt, 2001). Sin embargo, ha sido la genómica comparada la que ha revelado el papel predominante de la transferencia genética horizontal en la evolución de bacterias a todos los niveles de la escala evolutiva, aunque su importancia relativa depende en gran medida del contexto ecológico y las presiones selectivas que afectan a la evolución de diferentes linajes bacterianos. Los ejemplos anteriormente mencionados de proliferación de bacteriófagos y diferentes tipos de islas genómicas son claros ejemplos de transferencia genética horizontal, que también se ha documentado a escalas taxonómicas superiores como es el caso de las bacterias hipertermófilas *Aquifex aeolicus* y *Thermathoga maritima*, donde se estima que un 25% de sus genes proviene de transferencia genética horizontal desde arqueas (Aravind et al., 1998; Nelson et al., 1999), o dentro del grupo de las Xanthomonadales, donde análisis filogenéticos han demostrado un elevado grado de mosaicismo en estos genomas consecuencia de transferencias genéticas horizontales de diferentes subdivisiones de las proteobacterias (Comas et al., 2006). Por el contrario, los genomas de bacterias intracelulares obligadas representan ejemplos clásicos de ausencia de genes producto de transferencia genética horizontal consecuencia de su estricto asilamiento en el citoplasma de las células hospedadoras (Ochman et al., 2000). El hecho de que un gran número de sucesos de transferencia genética horizontal tengan lugar entre bacterias evolutivamente distantes incrementa la heterogeneidad composicional de los genomas bacterianos y representa la principal forma de identificar estos genes transferidos horizontalmente tanto por métodos composicionales (Karlin and Burge, 1995; Lawrence and Ochman, 1998; Mrazek and Karlin, 1999) como por métodos filogenéticos (Doolittle, 1999a; Bapteste et al., 2004). No obstante, los métodos filogenéticos se pueden ver afectados por los problemas clásicos asociados a la reconstrucción filogenética basada en secuencias, como la heterogeneidad en las tasas de evolución que puede generar agrupamientos erróneos de especies con evolución acelerada, problemas de resolución de nodos ancestrales, o problemas de identificación precisa de ortólogos (Moreira and Philippe, 2000), mientras que los métodos composicionales están limitados a detectar transferencias entre especies alejadas con composiciones nucleotídicas diferenciadas, de tal forma que transferencias entre organismos

próximos con similares características composicionales son difíciles de detectar, viéndose también afectados por el proceso de *amelioration* por el cual los genes transferidos horizontalmente tienden a adoptar las características composicionales de los genomas receptores, con lo cual se dificulta la detección de transferencias ancestrales (Lawrence and Ochman, 1997).

Se ha encontrado también que no todos los genes del genoma son igual de proclives a ser transferidos horizontalmente, y que genes informacionales implicados en procesos de replicación, transcripción o traducción tienden a transferirse con mucha menor frecuencia que genes operacionales implicados en metabolismo, transporte, u otras funciones celulares. Esto se ha explicado con lo que se llamó la hipótesis de la complejidad, por la cual los genes informacionales aparecen más restringidos a sucesos de transferencia horizontal debido a su mayor grado de coadaptación con otros genes con los que interactúan para formar los complejos proteicos de la maquinaria de replicación o de transcripción, lo que hace muy poco probable que puedan ser reemplazados por genes exógenos con la misma eficacia, mientras este tipo de restricciones sería menor en genes operacionales, lo que facilita su transferencia e integración entre linajes procariotas (Jain et al., 1999). No obstante, se han detectado casos de transferencia genética horizontal incluso a nivel de genes informacionales como aminoacil tRNA sintetasas (Woese et al., 2000) o proteínas ribosomales (Brochier et al., 2000), lo que demuestra que ningún gen es completamente inmune a sucesos de transferencia genética horizontal. No obstante, a pesar de que la importancia de la transferencia genética horizontal en la evolución de procariotas es indiscutible, existe una intensa discusión acerca de su impacto global en el sentido de si el paradigma darwiniano clásico de evolución por árboles bifurcados representando herencia vertical sería válido para representar la evolución de procariotas o constituye una representación incompleta de la evolución procariótica real debido a la presencia masiva de la transferencia genética horizontal. Por un lado están autores que proponen una clasificación más natural de linajes procariotas que considere la importancia de la transferencia genética horizontal entre linajes de forma cuantitativa, una aproximación a la que se refieren como "síntesis de la vida" (Bapteste et al., 2004) que postula que la evolución procariota se ajusta mejor a una red filogenética donde la herencia vertical se combina con transferencias horizontales entre linajes de diferente intensidad según el nivel de genes transferidos (Bapteste et al., 2004; Kunin et al., 2005; Dagan and Martin, 2006; Doolittle and Bapteste, 2007). Por otro lado, otros autores postulan que a pesar de la importancia de la transferencia genética horizontal en la evolución de procariotas, la fracción de genes transferidos que permanecen fijados a lo largo de la evolución es mínima, de tal forma que es posible identificar en los genomas un conjunto de genes conservados a lo largo de la evolución con herencia predominantemente vertical válidos para estudiar la evolución bacteriana a través de sus filogenias y otro conjunto de genes con presencia variable entre linajes que se corresponde con genes producto de transferencia genética horizontal como pueden

ser diferentes tipos de elementos genéticos móviles (Lerat et al., 2003; Daubin et al., 2003; Ochman et al., 2005). Recientemente se ha propuesto un modelo que trata de aunar las dos visiones bajo el cual una herencia predominantemente vertical se combinaría con lo que denominan "rutas de transferencia génica" entre linajes (Beiko et al., 2005).

En vista de todos estos resultados, la gran variabilidad en el contenido génico que se observa entre genomas procariotas a todos los niveles es una prueba clara de la importancia de la transferencia genética horizontal en la evolución de procariotas, que permite adaptaciones rápidas a cambios ambientales mediante la adquisición de determinados genes o islas genómicas que confieren funciones beneficiosas a los organismos receptores. Sin embargo, una fracción importante de estas transferencias se corresponde con genes de función desconocida o diferentes tipos de elementos genéticos móviles como bacteriófagos o IS que son predominantemente eliminados en la evolución futura del linaje receptor, de tal forma que una fracción significativamente reducida de los genes transferidos se mantiene de manera estable a lo largo de la evolución de los linajes bacterianos. Además, la importancia de la transferencia genética horizontal varía de manera significativa entre linajes en función de sus condiciones ecológicas y las presiones selectivas que afectan su evolución de tal forma que cualquier generalización a nivel global del efecto de la transferencia genética horizontal debe considerar esta heterogeneidad. Además, un modelo realista de evolución y especiación procariota debe considerar no sólo los efectos de la transferencia genética horizontal sino también los procesos de pérdida génica que aparecen como predominantes en la evolución de un gran número de linajes bacterianos, de tal forma que el balance entre ganancia y pérdida de genes es un parámetro esencial a la hora de entender la evolución de genomas procariotas (Snel et al., 2002; Kunin and Ouzounis, 2003b; Mirkin et al., 2003).

Evolución genómica por pérdida de genes: Evolución reductiva en bacterias intracelulares obligadas

A pesar de la importancia de duplicación génica y especialmente la transferencia genética horizontal, el fenómeno opuesto, la pérdida de genes, es igualmente importante siendo incluso predominante en la evolución de bacterias que establecen diferentes tipos de asociación desde simbiótica a patogénica con diferentes huéspedes eucariotas. De hecho, una de las primeras conclusiones que se extrajeron a partir de la comparación de los tamaños genómicos de diferentes linajes bacterianos es que las bacterias que establecen este tipo de asociaciones con huéspedes eucariotas presentan sistemáticamente tamaños genómicos reducidos en comparación con bacterias próximas de vida libre (Moran, 2007; Moya et al., 2008). Aunque en un principio se propuso que estos genomas reducidos eran reflejo del

genoma de los ancestros evolutivos a partir de los cuales los genomas de gran tamaño evolucionarían por duplicación o adquisición de genes (Wallace and Morowitz, 1973), actualmente está claro que estos genomas mínimos han evolucionado a partir de genomas ancestrales de mayor tamaño consecuencia de cambios ecológicos drásticos asociados a la transición a un modo de vida dependiente de huésped que se traducen en cambios en el balance entre selección y deriva genética que afectan a la evolución de estos genomas (Andersson and Kurland, 1998; Andersson and Andersson, 1999b). Este proceso de evolución reductiva ha tenido lugar repetidas veces a lo largo de la evolución de diferentes linajes procariotas, como en bacterias parasíticas del género *Wolbachia* y *Ricketsia* pertenecientes a la subdivisión alfa de las proteobacterias (Andersson and Andersson, 1999a; Ogata et al., 2001; Sallstrom and Andersson, 2005; Werren et al., 2008), dentro de los Mollicutes, que incluyen varios genomas de tamaño reducido como por ejemplo *Mycoplasma genitalium* (Fraser et al., 1995), en bacterias patógenas de la clase Clamidia (Horn and Wagner, 2004; Horn, 2008), en bacterias gram positivas patógenas de plantas como *Phytoplasma asteris* (Oshima et al., 2004), o incluso en bacterias de vida libre como es el caso de la cianobacteria *Prochlorococcus marinus*, donde variaciones de tamaño de 1 megabase se han detectado en diferentes cepas correspondientes a diferentes ecotipos adaptados a altas y bajas intensidades lumínicas en océanos (Dufresne et al., 2005; Kettler et al., 2007). Sin embargo, entre todos los linajes bacterianos, la subdivisión gamma de las proteobacterias es la que incluye el mayor número y los casos más extremos de genomas reducidos que se conocen, que aparecen asociados con bacterias endosimbiontes que viven en asociaciones intracelulares obligadas con diferentes insectos hospedador, siendo el grupo donde estas dinámicas de evolución reductiva se han caracterizado en gran detalle conjuntamente con parásitos intracelulares del género *Ricketsia* (α-proteobacteria) y patógenos intracelulares del género *Mycobacterium*, gracias a la disponibilidad de múltiples genomas de diferentes cepas de la misma especie conjuntamente con genomas de especies próximas de vida libre (Moran and Mira, 2001; Silva et al., 2001; Gomez-Valero et al., 2004a; Gomez-Valero et al., 2007; Darby et al., 2007).

Tanto en bacterias endosimbiontes como en bacterias patógenas o parasíticas, el proceso de reducción genómica está asociado con la pérdida de la capacidad de sobrevivir fuera del hospedador como bacteria de vida libre consecuencia de la pérdida de genes necesarios para tal fin, aunque el contenido génico final dependerá del tipo de asociación que se establece con el huésped eucariota. En bacterias endosimbiontes de insectos con genomas altamente reducidos se establece predominantemente una asociación de tipo mutualista en la cual tanto el insecto como la bacteria se benefician mutuamente de la interacción. Estas asociaciones tienen predominantemente un propósito nutricional, donde la bacteria endosimbionte le proporciona al insecto hospedador diferentes nutrientes que es incapaz de sintetizar por sí mismo o asimilar a partir de la dieta, como es el caso del suplemento

de aminoácidos por parte de *B. aphidicola* a su pulgón hospedador (Shigenobu et al., 2000; Tamas et al., 2002; van Ham et al., 2003; Perez-Brocal et al., 2006), el suplemento de diferentes cofactores por parte de *Wigglesworthia glossinidia* a la mosca tsé-tsé (Akman et al., 2002), o el suplemento de aminoácidos y la capacidad de reciclar urea por parte de *Blochmannia spp*. en el contexto de su asociación con hormigas carpintero (Gil et al., 2003; Degnan et al., 2005). En todos estos casos, la bacteria retiene genes cuyas funciones son más beneficiosas para el hospedador que para la bacteria misma. Por contra, en bacterias parásitas o patógenas, la bacteria no proporciona ningún beneficio directo al huésped, y retiene predominantemente genes implicados en interacciones celulares, mecanismos de patogenicidad y estrategias de defensa para asegurar su supervivencia en el hospedador (Ochman and Moran, 2001; Lawrence, 2005). No obstante, a pesar de estas diferencias marcadas por el tipo de asociación que se establece con el hospedador, tanto bacterias endosimbiontes como parásitas o patógenas intracelulares presentan características comunes asociadas al proceso de evolución reductiva como son genomas altamente reducidos con capacidades codificantes mínimas, altas tasas de evolución a nivel de secuencia que generan importantes sesgos composicionales hacia AT, aunque existen excepciones recientemente descritas en el caso de *Hodgkinia cicadicola*, endosimbionte de cicadas, que presenta un 58% de GC,(McCutcheon et al., 2009), y con una ausencia prácticamente total de sucesos de transferencia genética horizontal que indican mecanismos comunes en el proceso de reducción genómica asociados a cambios drásticos en la estructura poblacional y en las presiones selectivas que afectan a estas bacterias endosimbiontes en la transición desde a un modo de vida libre a un modo de vida intracelular dependiente de hospedador (Wernegreen, 2002; Wernegreen, 2004). Por un lado, los tejidos del hospedador proporcionan un ambiente mucho más estable, con un suplemento de nutrientes más o menos estable en comparación al modo de vida libre, lo que hace que disminuya significativamente la presión selectiva sobre un gran número de genes importantes para la supervivencia de la bacteria en un contexto de vida libre, de tal forma que mutaciones de pérdida de función sobre estos genes no serán eliminadas de forma efectiva por la selección natural, dando lugar a la acumulación de pseudogenes. Este proceso de inactivación génica se ve incrementado como consecuencia de una reducción drástica en el tamaño efectivo de las población endosimbionte en comparación con su ancestro de vida libre como consecuencia de los cuellos de botella que sufren las poblaciones de bacterias endosimbiontes durante su transmisión por línea vertical a través de las madres hospedadoras a sus descendientes, lo que hace que aumente el efecto de la deriva genética aleatoria y disminuya el efecto de la selección natural de tal forma que mutaciones ligeramente deletéreas en genes beneficiosos se pueden fijar en la población como consecuencia de la deriva genética. Estas mutaciones ligeramente deletéreas no se pueden recuperar a través de transferencia horizontal desde otro linaje debido a la estricta localización intracelular de estas bacterias endosimbiontes, dando lugar a una disminución en la eficacia de las poblaciones endosimbiontes como consecuencia de

la acumulación de mutaciones ligeramente deletéreas que no se pueden revertir por recombinación, un proceso que se conoce como Trinquete de Muller (Muller, 1964; Andersson and Hughes, 1996; Moran, 1996). La combinación de estos dos procesos (incremento del efecto de deriva genética y disminución de la presión selectiva para mantenimiento de función) hace que en etapas iniciales de la asociación con el hospedador, estas bacterias intracelulares obligadas acumulen un gran número de pseudogenes además de sufrir una proliferación masiva de elementos genéticos móviles que no serán eliminados por selección (Moran and Plague, 2004; Dale and Moran, 2006). Este proceso masivo de inactivación génica es seguido por una pérdida masiva de DNA en etapas más avanzadas de la asociación con el hospedador que conduce a los tamaños genómicos extremadamente reducidos que se observan en bacterias endosimbiontes con asociaciones ancestrales con su insecto hospedador como es el caso de los endosimbiontes bacterianos de insectos descritos anteriormente, siendo un ejemplo extremo el genoma de 160 kilobases del endosimbionte primario de psílidos *Carsonella rudii* (Nakabachi et al., 2006). Este proceso de pérdida génica es consecuencia predominantemente de múltiples eventos de deleción que llevan a la desintegración y pérdida de los pseudogenes ancestrales, pudiéndose detectar huellas de sucesos de pérdida incluso en los genomas de bacterias en etapas avanzadas de reducción genómica (Silva et al., 2001; Gomez-Valero et al., 2004a; Andersson and Andersson, 1999a; Andersson and Andersson, 2001), complementado con pérdidas en bloque de múltiples genes a través de deleciones de gran tamaño por sucesos de recombinación entre secuencias repetidas ancestrales (Moran and Mira, 2001).

Con respecto a las fuerzas evolutivas que gobiernan el proceso de reducción genómica, se han propuesto diferentes teorías que se pueden clasificar de manera general en teorías seleccionistas y teorías mutacionales. Las teorías seleccionistas postulan que la naturaleza compacta de los genomas bacterianos es consecuencia de su mayor ventaja selectiva en términos de tasas de replicación más rápidas y menor gasto energético dedicado a la replicación de DNA no codificante, de tal forma que los sucesos delecionales que eliminen DNA no codificante serán selectivamente ventajosos (Cavalier-Smith, 2005; Giovannoni et al., 2005). Bajo esta teoría, genomas compactos con poca cantidad de DNA no codificante serán selectivamente favorecidos en poblaciones naturales lo suficientemente grandes como para evitar los efectos de la deriva genética aleatoria, como es el caso de poblaciones naturales de bacterias de vida libre, cuyo tamaño es varios órdenes de magnitud superior al de poblaciones eucariotas. Sin embargo, esta teoría falla a la hora de explicar la evolución reductiva de bacterias intracelulares obligadas, donde la reducción genómica aparece asociada a una disminución del tamaño poblacional efectivo de las poblaciones en comparación con bacterias de vida libre (Mira and Moran, 2002), aunque el carácter poliploide de muchos de estos endosimbiontes intracelulares obligados podría aumentar el tamaño poblacional efectivo si fuera el genoma y no la célula la unidad de selección (Komaki and Ishikawa, 1999; Komaki and Ishikawa,

2000). No obstante, existen fuertes argumentos en contra de esta teoría seleccionista, como la ausencia de correlación entre tamaño genómico y tiempo de división que se observa al analizar diferentes linajes bacterianos o las altas tasas de crecimiento en organismos con tamaños genómicos grandes como *E. coli* mientras que en organismos con genomas reducidos como *Borrelia burgdorferi* las tasas de crecimiento son mucho más lentas (Bergthorsson and Ochman, 1998; Mira et al., 2001), lo que indica que las tasas de crecimiento celular no están limitadas por las tasas de replicación. Recientemente, una nueva teoría basada en principios genético-poblacionales con similitudes con esta teoría seleccionista fue propuesta por Michael Lynch y colaboradores que trata de explicar el incremento de tamaño y complejidad genómica que se observa en el continuo evolutivo desde procariotas a eucariotas (Lynch and Conery, 2003b). Esta teoría asume los excesos de DNA no codificante representan una carga mutacional para los organismos que lo poseen, de tal forma que es la estructura poblacional definida en base al tamaño poblacional efectivo la que determina la eficacia con la que este exceso de DNA es eliminado del genoma. Bajo esta teoría, la expansión de tamaños genómicos que se observa en eucariotas es una respuesta pasiva debido a tamaños poblacionales reducidos que hacen que la selección natural no pueda eliminar este exceso de DNA de forma eficiente, mientras que el elevado grado de compactación que se observa en genomas bacterianos con tamaños poblacionales mucho mayores sería consecuencia de una mayor eficacia de la selección natural a la hora de eliminar el DNA no funcional (Lynch and Conery, 2003b). Sin embargo, esta teoría no se ajusta a la situación que se observa en bacterias intracelulares obligadas, donde tamaños poblacionales efectivos reducidos aparecen asociados con genomas altamente reducidos (Daubin and Moran, 2004), aunque Lynch propone que la poliploidía asociada a estas bacterias intracelulares hace que el tamaño genómico efectivo sea mucho mayor que el tamaño poblacional efectivo (Lynch and Conery, 2004).

En contra de estas teorías seleccionistas y genético-poblacionales, las teorías mutacionales proponen que los tamaños genómicos reducidos que se observan en bacterias intracelulares obligadas son consecuencia de un sesgo mutacional que favorece las deleciones sobre las inserciones en los genomas bacterianos (Mira et al., 2001). Este sesgo mutacional hacia las deleciones es equilibrado por la selección natural favoreciendo el mantenimiento de la función génica, y es el balance entre estas dos fuerzas selectivas lo que determina el tamaño genómico final, de tal forma que en bacterias de vida libre la acción de la selección natural favoreciendo el mantenimiento de muchas funciones génicas conjuntamente con la adquisición de material genético exógeno mediante transferencia genética horizontal evita fenómenos de reducción genómica masiva, pero en bacterias intracelulares obligadas, la reducción de esta presión selectiva para el mantenimiento de funciones génicas que dejan de ser importantes en el nicho ecológico asociado al hospedador conjuntamente con el incremento de la deriva genética sobre la selección natural consecuencia de la reducción en los tamaños poblacionales efectivos de las bacterias

intracelulares hace que se incremente la cantidad de DNA no funcional en estos genomas que será eliminado como consecuencia del sesgo delecional inherente en genomas procariotas tanto por grandes deleciones de múltiples genes como por pérdidas gen a gen a través de múltiples deleciones de pequeño tamaño (Mira et al., 2001; Silva et al., 2001). Evidencias a favor de la existencia de este sesgo mutacional hacia las deleciones se han detectado tanto en bacterias de vida libre como en bacterias intracelulares comparando el tamaño y la frecuencia de inserciones y deleciones en pseudogenes en comparación con ortólogos funcionales de otros genomas, que indican que las deleciones son el suceso mutacional predominante (Andersson and Andersson, 1999a; Andersson and Andersson, 2001). También se ha propuesto que este sesgo delecional puede tener cierto valor selectivo ya que evita la acumulación de elementos genéticos móviles con efectos deletéreos como transposones, IS o especialmente bacteriófagos que entran en los genomas a través de sucesos de transferencia genética horizontal (Lawrence et al., 2001), de tal forma que en bacterias de vida libre altas tasas de deleción será selectivamente favorables ya que evitan la acumulación de material genético deletéreo, mientras que en bacterias intracelulares obligadas el incremento del efecto de la deriva genética conjuntamente con la ausencia de transferencia genética horizontal hacen que las tasas de deleción se reduzcan, lo que explica la presencia de un gran número de pseudogenes especialmente en bacterias en etapas iniciales de la asociación con el hospedador, como los más de 1100 pseudogenes presentes en el genoma de *Mycobacterium leprae* (Cole, 1998; Gomez-Valero et al., 2007), que serán eliminados del genoma a largo plazo como consecuencia del sesgo mutacional hacia las deleciones dando lugar a los genomas mínimos que se observan en bacterias endosimbiontes con asociaciones ancestrales como *B. aphidicola* en pulgones o *Carsonella rudii* en psílidos (Lawrence et al., 2001).

En el contexto de esta tesis, el endosimbionte secundario de la mosca tsé-tsé *S. glossinidius* representa uno de los pocos ejemplos de una bacteria en etapas muy tempranas de la asociación con el insecto hospedador, con un genoma de tamaño similar al de una bacteria de vida libre como *E. coli* pero con un gran número de pseudogenes consecuencia del proceso de reducción genómica que está sufriendo como consecuencia de la transición a un modo de vida dependiente de huésped. Esto hace que tengamos tanto la huella del genoma ancestral como el estado actual del genoma, lo que nos permite abordar el estudio de esta transición evolutiva ancestral desde el ancestro de vida libre hasta el estado actual con el objetivo de determinar los efectos de la proliferación de pseudogenes y elementos genéticos móviles sobre la funcionalidad y organización del genoma. Además, si consideramos el estado actual de *S. glossinidius* como un estado transitorio dentro del proceso de evolución reductiva, es posible simular este proceso reductivo "*in-silico*" desde el hipotético ancestro de vida libre hasta los posibles genomas mínimos que se podrían producir a partir del estado funcional del genoma actual mediante de herramientas de la biología de sistemas, que permiten llevar a cabo aproximaciones computacionales a

diferentes aspectos de la funcionalidad global de los sistemas biológicos a través de la reconstrucción y análisis funcional de diferentes tipos de redes celulares.

Genómica y biología de sistemas

La disponibilidad de genomas completamente secuenciados y anotados permite una aproximación a las capacidades funcionales del organismo correspondiente a través del conjunto de genes presentes en el genoma. A partir de la información residente en la anotación de genes y pseudogenes es posible determinar el nivel de autonomía de un determinado organismo, cuáles son sus capacidades biosintéticas y degradativas, y en el contexto de la evolución reductiva, cual es el nivel de integración de la bacteria endosimbionte con su correspondiente hospedador y cuáles pueden ser las posibles vías de integración con la biología del huésped (Zientz et al., 2004). Sin embargo, las principales características biológicas son consecuencia de interacciones complejas entre diferentes constituyentes celulares, como la interacción entre factores de transcripción con sus correspondientes secuencias reguladoras para controlar la expresión génica, la interacción entre subunidades de diferentes complejos proteicos celulares necesaria para su correcto funcionamiento, o la integración de diferentes reacciones metabólicas catalizadas por diferentes enzimas celulares en diferentes rutas metabólicas que determinan las capacidades metabólicas del organismo. Estas interacciones se pueden representar desde el punto de vista computacional a través de diferentes tipos de redes como puedan ser redes reguladoras, de interacción proteína-proteína o redes metabólicas, de tal forma que el análisis de estas redes permite una aproximación a la organización y funcionalidad real de los sistemas biológicos (Barabasi and Oltvai, 2004).

Diversos estudios teóricos acerca de la organización estructural y la topología de redes biológicas reconstruidas en diferentes organismos tanto procariotas como eucariotas han revelado que, a pesar de la diversidad de redes biológicas existentes en la naturaleza, todas se organizan de forma similar siguiendo lo que se denomina una ley de potencias en la conectividad de sus elementos o nodos (Barabasi and Albert, 1999; Jeong et al., 2000). Esto hace que la conectividad de los nodos de la red no sea uniforme, como se esperaría en redes aleatorias, sino que aparece gobernada por unos pocos nodos que soportan un gran número de interacciones y que son los que soportan el peso estructural de la red ("*hubs*") mientras que el resto de nodos de la red presenta muy pocas conexiones, principalmente a estos nodos principales. Esta estructura genera una gran diversidad en los valores de conectividad de los nodos de la red, definidos como el número de interacciones con otros elementos, lo que hace que no se pueda definir un nodo característico que represente al resto de nodos de la red y por este motivo se considere a estas redes

como redes libres de escala (Barabasi and Oltvai, 2004) frente a la organización de redes aleatorias donde todos los nodos presentan el mismo número de interacciones promedio como consecuencia de la distribución aleatoria de conectividad entre nodos de la red (Erdos and Renyi, 1960). La principal consecuencia de esta organización libre de escala es que las redes biológicas son más robustas frente a alteraciones aleatorias de su estructura, que en redes biológicas se traduce como una mayor tolerancia sucesos de inactivación o pérdida de genes ya que la mayoría de estos sucesos afectaran a nodos con pocas conexiones de tal forma que la organización global de la red permanecerá inalterada (Albert et al., 2000; Albert, 2005). Esto supone una mayor adaptabilidad del sistema a cambios en su estructura, pero al mismo tiempo una mayor vulnerabilidad del mismo frente a ataques directos a los pocos nodos con un elevado número de interacciones que soportan el peso estructural de la red.

Este análisis topológico de las redes biológicas presenta limitaciones intrínsecas en el sentido de que aunque su estudio es esencial para entender la estructura de la red, no proporciona información acerca del perfil funcional del sistema, lo cual es especialmente importante en el estudio de redes metabólicas donde la topología de la red no permite inferir las capacidades funcionales del sistema bajo diferentes condiciones de entorno. Estas capacidades metabólicas se pueden definir mediante las distribuciones de flujos metabólicos a través de la red, que se definen como la cantidad de sustrato que es convertido a su correspondiente producto en cada reacción de la red por unidad de tiempo (Edwards et al., 2002; Barabasi and Oltvai, 2004). Matemáticamente es posible predecir el conjunto de flujos metabólicos que opera en una red metabólica bajo diferentes condiciones de entorno a partir únicamente de la estequiometria de las reacciones incluidas en la red y asumiendo que el sistema se encuentra en estado estacionario, lo que significa que el balance neto entre la producción y consumo de cualquier metabolito de la red es igual a cero (Varma and Palsson, 1994a; Edwards et al., 2002; Price et al., 2003; Feist et al., 2009). La restricción impuesta por el estado estacionario hace que se genere desde el punto de vista matemático un conjunto de soluciones o posibles distribuciones de flujos metabólicos a través de las reacciones de la red que se ajusten a dicha restricción, y este conjunto de soluciones se puede explorar a través de diferentes aproximaciones computacionales. Una posibilidad es llevar a cabo un muestreo aleatorio de este conjunto de soluciones (distribuciones de flujo metabólico que satisfacen la restricción del estado estacionario) para tener una visión general de las capacidades funcionales de la red metabólica bajo unas determinadas condiciones de entorno (Almaas et al., 2004; Wiback et al., 2004), o determinar los valores de flujo máximo y mínimo a través de cada reacción de la red bajo los cuales se satisface la asunción de estado estacionario mediante Análisis de Variabilidad de Flujos o FVA (Mahadevan and Schilling, 2003). Sin embargo, una de las principales finalidades de la reconstrucción de redes metabólicas a escala genómica es ser capaz de predecir cuantitativamente la viabilidad del sistema en términos de crecimiento bajo unas

determinadas condiciones de entorno de forma equivalente a medidas experimentales de crecimiento celular, y esto se puede llevar a cabo mediante el Análisis de Balance de Flujos o FBA, que consiste en utilizar herramientas de programación lineal para determinar qué distribución de flujos metabólicos de todas las posibles bajo estado estacionario optimiza una determinada función objetivo, que se suele definir como una ecuación de producción de biomasa donde se reflejan todos los metabolitos que el sistema tiene que producir para sobrevivir en unas condiciones de entorno determinadas (Schilling et al., 2000a; Palsson, 2006; Durot et al., 2009). FBA se basa en la asunción de que los organismos han optimizado su capacidad de crecimiento bajo determinadas condiciones ambientales a lo largo de su evolución, de tal forma que la maximización de la producción de biomasa representa una buena aproximación a la funcionalidad global del sistema (Varma and Palsson, 1994a). Esta asunción se ha comprobado experimentalmente en diferentes organismos capaces de ser cultivados in vitro, donde las predicciones de producción de biomasa por FBA sobre el modelo metabólico en diferentes condiciones de entorno se ajustan con medidas experimentales de crecimiento en diferentes medios de cultivo (Edwards et al., 2001).

Del mismo modo, predicciones acerca del carácter esencial de genes mediante simulaciones de deleción sobre redes metabólicas se han demostrado altamente coincidentes con medidas de esencialidad experimentales determinadas a partir de experimentos de inactivación génica (Joyce and Palsson, 2008), lo cual abre un espectro de posibles aplicaciones de este tipo de análisis en el contexto de la evolución de genomas bacterianos que no se pueden abordar mediante análisis genómico clásico. Por ejemplo, FBA se ha utilizado para evaluar la contribución de genes transferidos horizontalmente a la funcionalidad del sistema en redes metabólicas de bacterias, revelando que estos genes son metabólicamente activos sólo bajo ciertas condiciones de entorno (Pal et al., 2005a). Del mismo modo, simulaciones de evolución reductiva sobre la red metabólica de *E. coli* mediante FBA predicen con alta sensibilidad el contenido génico de genomas mínimos reales como *W. glossinidia* y *B. aphidicola* bajo condiciones de entorno similares a las que la bacteria encuentra en su asociación con su correspondiente hospedador (Pal et al., 2006b).

CAPÍTULO 2: Objetivos generales

Los objetivos de esta tesis son:

- Reconstruir las filogenias de orden génico en gamma-proteobacterias a partir de un conjunto de genes conservados en todos los genomas y comparar con filogenias obtenidas a partir de datos de secuencia.

- Analizar los patrones de evolución del orden génico en diferentes linajes dentro de las gamma-proteobacterias, incluyendo diferentes linajes de bacterias endosimbiontes de insectos con asociaciones ancestrales (*B. aphidicola*, *W. glossinidia* y *B. floridanus*) y más recientes (*S. glossinidius*) con sus correspondientes insectos huésped.

- Caracterización de pseudogenes y elementos genéticos móviles en el genoma de *S. glossinidius*, el endosimbionte secundario de la mosca tsé-tsé, y evaluación de su impacto sobre la estructura genómica y las capacidades funcionales de esta bacteria endosimbionte en estadios muy tempranos de la asociación con el hospedador. Integración con las capacidades metabólicas del endosimbionte primario *W. glossinidia* en el contexto ecológico de la asociación con la mosca tsé-tsé.

- Reconstrucción de la red metabólica de *S. glossinidius* a diferentes estadios del proceso de evolución reductiva y análisis funcional por FBA para determinar los sucesos clave en la transición a un modo de vida dependiente de huésped y evaluar el impacto de la pérdida génica sobre la robustez y capacidad funcional de *S. glossinidius*.

- Predicción de la posible evolución futura de *S. glossinidius* en el contexto de la reducción genómica mediante simulaciones de evolución reductiva sobre su red metabólica funcional y correlación entre esencialidad y evolución a nivel de secuencia a partir de genes presentes y ausentes en redes metabólicas mínimas

CAPÍTULO 3: Distancias de reordenaciones genómicas y evolución del orden génico en gamma-proteobacterias

Los cromosomas bacterianos presentan características estructurales y de organización comunes que han sido reveladas mediante genómica comparada, estableciéndose un compromiso evolutivo entre el mantenimiento de una organización cromosómica funcionalmente beneficiosa para el organismo y la

generación de nuevas variantes mediante reordenaciones cromosómicas y sucesos de ganancia y pérdida de genes que puedan tener valor adaptativo bajo ciertas condiciones de entorno. Un claro ejemplo de estructuras cromosómicas conservadas es la organización de genes funcionalmente relacionados en operones que facilitan una regulación común de la expresión génica (Jacob and Monod, 1961) así como la conservación de cierta organización cromosómica a nivel supra-operónico (Lathe, III et al., 2000; Korbel et al., 2004). Además, la organización de los cromosomas bacterianos, con un único origen de replicación, genera un gradiente de dosis génica que permite la sobreexpresión de genes cercanos al origen de replicación, una característica especialmente relevante en bacterias de crecimiento rápido (Schmid and Roth, 1987; Sousa et al., 1997). También se ha comprobado que existe una tendencia significativa hacia la localización de los genes esenciales o funcionalmente importantes en la cadena líder del cromosoma bacteriano para evitar colisiones entre la maquinaria de replicación y la de transcripción que generarían transcritos truncados con consecuencias especialmente letales en genes funcionalmente importantes (Rocha and Danchin, 2003b; Rocha and Danchin, 2003a). En contra de esta conservación de la organización cromosómica, los genomas bacterianos contienen un gran número de secuencias repetidas capaces de generar diferentes tipos de reordenaciones cromosómicas mediante recombinación homóloga o ilegítima (Achaz et al., 2002; Achaz et al., 2003a). Estos sucesos de reordenación pueden tener cierto valor adaptativo como estrategias de adaptación a cambios drásticos en las condiciones de entorno de una bacteria, generando nuevas variantes cromosómicas beneficiosas (Rocha, 2004).

Antes de la era genómica, la construcción de mapas genéticos detallados en diferentes bacterias ya reveló que el orden génico no se conservaba entre bacterias evolutivamente alejadas. Posteriormente, la secuenciación y comparación de múltiples genomas bacterianos ha mostrado que especies evolutivamente próximas acumulan un menor número de reordenaciones que especies evolutivamente más alejadas, aunque existen excepciones respecto a esta tendencia general tanto a nivel de linajes evolutivos con una significativa conservación del orden génico como linajes que acumulan un elevado número de reordenaciones (Casjens, 1998; Nadeau and Sankoff, 1998). Existen cuatro tipos principales de cambios que pueden alterar el orden de los genes en un cromosoma. En primer lugar, inversiones y transposiciones se detectan de manera habitual cuando se comparan genomas evolutivamente próximos (Hughes, 2000). Las inversiones suelen ser simétricas alrededor del origen de replicación para minimizar el efecto sobre la organización global del cromosoma (Tillier and Collins, 2000; Eisen et al., 2000), mientras que las translocaciones se pueden producir tanto dentro de un mismo cromosoma como entre cromosomas. Los genes también se pueden delecionar tanto a través de un único suceso de recombinación como a través de un proceso progresivo de desintegración génica, dando lugar a huecos cuando los genomas son comparados (Andersson and Andersson, 2001; Silva et al., 2001; Moran, 2002). En tercer lugar,

la transferencia genética horizontal genera inserciones de material genético a lo largo del genoma que pueden llegar a representar hasta el 20%-30% del genoma en comparaciones entre especies próximas, como en cepas uropatogénicas de *E. coli* en comparación con cepas no patogénicas de laboratorio (Welch et al., 2002), aunque en otros linajes evolutivos como las bacterias intracelulares obligadas su influencia es prácticamente nula (Ochman et al., 2000). Por último, duplicaciones parciales de segmentos cromosómicos también alteran el orden génico del genoma.

Las reordenaciones genómicas se han estudiado en diferentes linajes bacterianos, entre ellos en Gamma (γ)-proteobacteria, donde se ha detectado que las inversiones cromosómicas son comunes al comparar los genomas de especies próximas (Hughes, 2000). Desde el punto de vista evolutivo y dentro de los objetivos de esta tesis, las γ-proteobacterias son particularmente interesantes debido al hecho de que se dispone de un gran número de genomas completamente secuenciados con diferentes grados de proximidad evolutiva (desde cepas de una misma especie hasta especies de un mismo género o grupo), y además incluyen un gran número de genomas de bacterias endosimbiontes de insectos con genomas altamente reducidos, como los del endosimbionte primario de pulgones *B. aphidicola* de diferentes especies de hospedador (Shigenobu et al., 2000; Tamas et al., 2002; van Ham et al., 2003; Perez-Brocal et al., 2006; Moran et al., 2009), los genomas del endosimbionte primario y secundario de la mosca tsé-tsé *W. glossinidia* y *S. glossinidius* (Akman et al., 2002; Toh et al., 2006), el endosimbionte primario de diferentes especies de hormiga carpintero *Blochmannia spp.* (Gil et al., 2003; Degnan et al., 2005), o el genoma de *Carsonella rudii*, el endosimbionte primario de psílidos (Nakabachi et al., 2006). El caso de *B. aphidicola* ejemplifica las particulares características evolutivas de los endosimbiontes bacterianos de insectos. Este endosimbionte primario se transmite de manera estrictamente vertical por vía materna (Baumann et al., 1995), y el tiempo de divergencia entre las primeras tres cepas secuenciadas se estimó en 164 millones de años basado en la divergencia entre sus correspondientes pulgones hospedadores (von Dohlen and Moran, 2000). Durante todo este periodo se mantiene una estabilidad genómica prácticamente total, con ningún suceso de transferencia genética horizontal ni duplicación génica y con solo cuatro pequeñas reordenaciones en el genoma de *B. aphidicola* del pulgón *Baizongia pistaciae* correspondientes a dos pequeñas inversiones de uno y seis genes y dos pequeñas translocaciones desde dos plásmidos al genoma de dos y cuatro genes (Tamas et al., 2002; Silva et al., 2003; van Ham et al., 2003). Sin embargo, a pesar de esta estabilidad cromosómica, al menos 164 pérdidas génicas han tenido lugar a lo largo de la evolución de estos tres linajes (Silva et al., 2003; Gomez-Valero et al., 2004a). Este periodo de estabilidad genómica contrasta con el resto de la evolución de *B. aphidicola* desde su divergencia de una bacteria de vida libre próxima como *E. coli*. En este primer periodo se produjeron un gran número de reordenaciones y más de mil pérdidas génicas asociadas tanto a pequeñas como a grandes deleciones (Moran and Mira, 2001; Silva et al., 2001).

Dadas las particulares características evolutivas de estos linajes endosimbiontes, con elevadas tasas de mutación desde el establecimiento de la asociación endosimbionte (Moran, 1996; Itoh et al., 2002) y sesgos composicionales hacia altos contenidos de AT que afectan también a su uso de aminoácidos (Moran, 1996; Clark et al., 1999; Shigenobu et al., 2000; Palacios and Wernegreen, 2002; Rispe et al., 2004), su posición dentro del árbol filogenético de las γ-proteobacterias y en particular su relación con las bacterias entéricas de vida libre, que se postulan como sus ancestros evolutivos más próximos, es una cuestión de debate dentro del campo de la evolución molecular. El agrupamiento de las tres especies de endosimbiontes bacterianos de insectos (*B. aphidicola*, *W. glossinidia* y *B. floridanus*) se ha propuesto en base a filogenias reconstruidas a partir del gen 16S ribosomal (Sauer et al., 2000) y en filogenias inferidas a partir de un alineamiento concatenado de 61 genes (Gil et al., 2003). El monofiletismo de *B. aphidicola* y *W. glossindia* también se ha observado en aproximaciones filogenómicas (Daubin et al., 2003; Canback et al., 2004). Sin embargo, también existen filogenias basadas en el gen 16S ribosomal que rechazan el monofiletismo de las tres especies de endosimbiontes (Charles et al., 2001). La dificultad de las reconstrucciones filogenéticas con endosimbiontes se pone de manifiesto por el hecho que en la mayor parte de las filogenias de genes individuales, el linaje de *B. aphidicola*muestra una posición basal dentro de las γ-proteobacterias y bastante alejadas de las bacterias entéricas. La disponibilidad de genomas completamente secuenciados hace posible reconstruir filogenias utilizando la información completa de todos los genes del genoma para tratar de solucionar los problemas de incongruencia filogenética que se observan al comparar filogenias de genes individuales (Wolf et al., 2002; Bapteste et al., 2004). Estas aproximaciones filogenómicas se pueden basar en datos de secuencia utilizando alineamientos concatenados de un conjunto de genes (Hansmann and Martin, 2000; Brown et al., 2001), o bien pueden utilizar la reconstrucción de superárboles a partir del conjunto de filogenias de genes individuales (Sicheritz-Ponten and Andersson, 2001; Bininda-Emonds, 2004). Para minimizar el posible efecto distorsionador de la transferencia genética horizontal en las filogenias, se puede utilizar un conjunto de genes conservados en todos los genomas de interés (Daubin et al., 2002). Por último, también se puede utilizar el contenido génico y el orden génico como caracteres para la reconstrucción filogenética (Fitz-Gibbon and House, 1999; Suyama and Bork, 2001).

Los objetivos principales de este capítulo son, en primer lugar, estimar distancias evolutivas basadas en reordenaciones genómicas entre genomas de γ-proteobacterias utilizando dos diferentes medidas de estabilidad genómica que son el número de inversiones y el número de puntos de rotura (BP) entre cromosomas. Para ello se utilizará un subconjunto de genes presentes en todos los genomas que estudiamos, que se corresponden con genes esenciales o funcionalmente importantes que evolucionarán más lentamente a nivel de orden génico ya que cambios de posición en estos genes pueden tener consecuencias deletéreas. Además, de esta forma

tratamos de minimizar la incorporación de genes producto de transferencia genética horizontal. A partir de las distancias de orden génico se tratará de determinar qué linajes evolucionan más rápidamente o más lentamente a nivel de reordenaciones genómicas, y analizaremos el patrón evolutivo de los diferentes linajes de bacterias endosimbiontes de insectos en etapas avanzadas (*B. aphidicola*, *B. floridanus*, y *W. glossinidia*) y iniciales (*S. glossinidius*) del proceso de reducción genómica. Por último, las distancias de orden génico se utilizarán para reconstruir la filogenia de γ-proteobacterias y se compararan los resultados con filogenias obtenidas a partir de datos de secuencia para analizar las diferencias en la evolución de ambos caracteres (secuencia y orden génico) y como estas diferencias afectan al posible monofiletismo de los endosimbiontes bacterianos de insectos.

Resultados y discusión

Identificación de genes ortólogos y cálculo de distancias de orden génico entre genomas de γ-proteobacterias

El primer paso en este análisis consistió en la reconstrucción de una tabla de genes ortólogos en los 31 genomas de γ-proteobacterias con los que vamos a trabajar (Ver Tabla 3.1). Para que un gen se incluya en la tabla, se exige que un ortólogo esté presente en los 31 genomas tanto como gen o como pseudogen, ya que desde el punto de vista del cálculo de distancias de orden génico solo nos interesa la posición del gen, independientemente de si está o no está activo. El resultado de este análisis es una tabla de 244 genes presentes en los 31 genomas de γ-proteobacterias con los que trabajamos donde se refleja la posición y orientación de cada gen en cada genoma. Estos 244 genes son sometidos a un segundo análisis para testar si alguno de ellos es producto de transferencia genética horizontal que consiste en mapear estos genes en la base de datos Horizontal Gene Transfer Database (HGT-DB)(Garcia-Vallve et al., 2003), así como en un conjunto de genes identificados como xenólogos en base a diferentes criterios incluyendo validación filogenética y composicional (Medrano-Soto et al., 2004). En ninguno de los 244 genes encontramos evidencia significativa de transferencia genética horizontal.

Para calcular el número de reordenaciones entre genomas se utilizan dos diferentes distancias de orden génico. Por un lado se calculan la distancia BP entre pares de genomas, que se define como el número de adyacencias génicas presentes en un genoma y ausentes en el otro (Nadeau and Taylor, 1984), la cual normalizamos por el total de genes considerados (244 genes). Por otro lado se calcula la distancia de inversiones entre pares de genomas, que se define como el mínimo número de inversiones necesarias para pasar de un genoma a otro (Tesler, 2002), normalizando también por el número de genes considerados. En ambos casos, la información que se considera es la posición y la orientación (positiva o negativa) de cada uno de los 244 genes en los 31 genomas de γ-proteobacterias. Es importante tener en cuenta

que tanto las distancias BP como las distancias de inversiones hacen referencia al número de BP e inversiones observado entre dos genomas, y por tanto, al igual que sucede con distancias basadas en datos de secuencias, pueden diferir de la distancia evolutiva real, ya que, por ejemplo, en el caso de las distancias BP, un mismo punto de rotura puede ser utilizado por diferentes sucesos de reordenación, un hecho que no es posible inferir a partir de las distancias observadas. Con datos de secuencia, esta subestima de la distancia evolutiva real se puede corregir mediante la aplicación de diferentes modelos evolutivos (Posada and Buckley, 2004), pero con datos de orden génico estos modelos se encuentran mucho menos desarrollados (Nadeau and Taylor, 1984; Wang and Warnow, 2001). Sin embargo, diferentes estudios de simulación computacional comparando distancias estimadas con distancias reales de inversiones y puntos de rotura muestran que estas desviaciones son mínimas en el rango de distancias normalizadas con el que trabajamos en este estudio (Bourque and Pevzner, 2002; Moret et al., 2002b). Al comparar las estimas de ambos tipos de distancias de orden génico, se observó una gran correlación entre las distancias de inversiones y las distancias BP (r=0.996; Ver Figura 3.1) que indica que las inversiones son el suceso de reordenación predominante en la evolución de γ-proteobacterias, lo que confirma resultados previos de otros estudios (Hughes, 2000).

Distancias basadas en orden génico contra distancias basadas en secuencias aminoacídicas a lo largo de la evolución de γ-proteobacterias

Con el objetivo de detectar si las tasas de reordenaciones cromosómicas han sido constantes a lo largo de la evolución de los diferentes linajes dentro de las γ-proteobacterias, las distancias de orden génico basadas en inversiones y puntos de rotura entre pares de genomas se compararon con distancias evolutivas basadas en datos de secuencia. Para ellos se utilizó un alineamiento concatenado de las proteínas codificadas por 10 genes informacionales de evolución lenta conservados en los 31 genomas con los que trabajamos (3670 posiciones), a partir del cual se calcularon distancias de sustitución aminoacídica entre pares de genomas por máxima verosimilitud asumiendo el modelo VT de evolución proteica (Muller and Vingron, 2000). La comparación entre las distancias de orden génico y las distancias basadas en secuencias aminoacídicas muestran que aunque existe una tendencia general hacia un incremento del número de reordenaciones genómicas a lo largo del tiempo evolutivo, existen fuertes variaciones tanto dentro de cada linaje como entre diferentes linajes de γ-proteobacterias (Ver Figura 3.3). Este es el caso del linaje de las Pasteurellas (*H. ducreyi*, *H. influenzae* and *P. multocida*), donde se observa un número significativamente elevado de puntos de rotura e inversiones frente al resto de genomas de γ-proteobacterias estudiados, lo que indica que este linaje presenta tasas de reordenaciones aceleradas en comparación con el resto de γ-proteobacterias. Esta aceleración también tiene lugar aunque en menor intensidad en *S. glossinidius*, el endosimbionte secundario de la mosca tsé-tsé, cuyas distancias de orden génico

frente al resto de γ-proteobacterias aparecen por encima de la tendencia general (Ver Figura 3.3). Situación opuesta a la de *S. glossinidius* la encontramos en *B. floridanus* cuando la comparamos con el resto de bacterias entéricas de vida libre, donde la aceleración en las tasas de sustitución característica de endosimbiontes bacterianos, que afectaría también a las distancias de sustitución aminoacídica, explican los altos valores de distancias de secuencia observados para este linaje en comparación con las distancias de orden génico. Por último, las comparaciones entre los tres genomas de *B. aphidicola* muestran una ausencia prácticamente total de reordenaciones entre ellas a pesar de las altas tasas de sustitución aminoacídica que se observan, que al igual que en el caso de *B. floridanus*, se explica por la aceleración en las tasas de sustitución que sufren estos linajes endosimbiontes desde su transición al modo de vida dependiente de hospedador (Moran, 1996; Itoh et al., 2002).

Estos resultados indican que, si bien para algunas especies se observa un incremento paralelo del numero de reordenaciones genómicas con las distancias de sustitución aminoacídica, existen diferentes linajes que muestran un marcado desequilibrio o heterogeneidad entre ambos parámetros, pudiendo encontrar desde linajes como las Pasteurellas donde se produce una aceleración en las tasas de reordenación en comparación con la evolución a nivel de secuencia hasta linajes como *B. aphidicola* donde se observa la tendencia inversa, una aceleración en las tasas de sustitución aminoacídica conjuntamente con una ausencia prácticamente total de reordenaciones entre los tres genomas de *B. aphidicola* comparados.

Distancias de inversión relativas

Con el objetivo de confirmar la aceleración en las tasas de reordenaciones genómicas en los linajes de las Pasteurellas y en el linaje de *S. glossinidius* así como para caracterizar la evolución de bacterias endosimbiontes a nivel de orden génico frente a bacterias entéricas de vida libre se llevó a cabo una aproximación similar al test de tasas relativas de Tajima comparando las tasas de inversión en los linajes de *B. aphidicola BAp*, *B. floridanus*, *W. glossinidia*, *S. glossinidius*, *V. cholerae*, *V. vulnificus*, *V. parahaemolyticus*, *H. ducreyi*, *H. influenzae* y *P. multocida* con los de bacterias entéricas de vida libre *E. coli* K12, *S. thypimurium* LT2, *S. enterica* subsp. *Enterica* serovar Thipi str. CT18, *S. enterica* subsp. *Enterica* serovar Typhi Ty2, *S. flexneri* 2a str. 301, *S. flexneri* 2a str. 2457T, *Y. pestis* CO92 y *Y. pestis* KIM desde su divergencia a partir del ancestro común utilizando como grupo externo *S. oneidensis* MR-1, *P. aeruginosa*, *P. putida*, y *P. syringae* (Ver Figura 3.4). Los resultados de este análisis confirman que los linajes de las Pasteurellas y de *S. glossinidius* están evolucionando a unas tasas de inversiones que son aproximadamente el doble que las de bacterias entéricas de vida libre, lo que explica su posición divergente respecto a la tendencia general cuando se comparan las distancias de orden génico con las distancias de sustitución aminoacídica. En el caso de los endosimbiontes primarios su comportamiento es ligeramente diferente en los

tres linajes analizados, ya que si bien *B. aphidicola BAp* y *W. glossinidia* están evolucionando a unas tasas de inversiones ligeramente superiores que las bacterias entéricas de vida libre (1.38 y 1.35 veces superior respectivamente), *B. floridanus* parece evolucionar a la misma tasa que bacterias entéricas de vida libre (1.04). Sin embargo, la correcta interpretación de estas tasas de inversiones en los endosimbiontes primarios requiere tener en cuenta que las inversiones calculadas para cada uno de ellos en este análisis se corresponden con el número total de inversiones que han acumulado desde la divergencia del grupo de bacterias entéricas de vida libre. En el caso de *B. aphidicola*, se sabe que durante los últimos 100-150 millones de años sus genomas no han sufrido prácticamente ninguna reordenación (Tamas et al., 2002), de tal forma que si asumimos que la divergencia entre *B. aphidicola* y el grupo de bacterias entéricas de vida libre tuvo lugar en algún momento entre 200 y 300 millones de años atrás, se definirían dos periodos claramente diferenciados en la evolución genómica de *B. aphidicola*. En un primer periodo evolutivo que se correspondería con etapas iniciales de la adaptación al modo de vida endosimbionte desde el ancestro de vida libre, la tasa de inversiones sería aproximadamente el doble (2.76) del valor promedio que estimamos para toda la evolución de *B. aphidicola BAp* desde en ancestro común con las entéricas de vida libre (1.38 veces superior). Este periodo inicial de aceleración en las tasas de inversiones se continuaría con un segundo periodo correspondiente a la divergencia entre las diferentes cepas de *B. aphidicola* durante el cual la tasa de reordenaciones es prácticamente nula a la luz del elevado grado de conservación del orden génico que se observa entre genomas de *B. aphidicola*.

Estas conclusiones no son exclusivas de *B. aphidicola*, sino que se pueden considerar como un fenómeno común en los diferentes linajes de bacterias endosimbiontes con los que trabajamos. En el caso del endosimbionte primario de la hormiga carpintero *B. floridanus*, el segundo genoma secuenciado correspondiente a *B. pensylvannicus* muestra una conservación absoluta del orden génico con *B. floridanus*, de tal forma que la tasa de inversiones relativa de 1.04 que resulta de comparar *B. floridanus* con las bacterias entéricas de vida libre se correspondería con los mismos dos periodos de evolución que describimos en *B. aphidicola*, con los últimos 16-20 millones de años correspondientes a la divergencia entre *B. floridanus* y *B. pensylvannicus* con una tasa de reordenaciones nula mientras que las inversiones se acumularían en las etapas iniciales de la adaptación al modo de vida endosimbionte a una tasa aproximadamente el doble de los que hemos inferido (2.08) dado que la edad del ancestro de todas las hormigas que presentan un endosimbionte primario del género *Blochmannia* se ha estimado en 29,3-35,9 millones de años (Degnan et al., 2004; Degnan et al., 2005). Esta aceleración en etapas iniciales de la adaptación al modo de vida dependiente de hospedador queda demostrada en la evolución de *S. glossinidius*, el endosimbionte secundario de la mosca tsé-tsé, el cual ha establecido una transición muy reciente al modo de vida dependiente de huésped que se refleja en su tamaño genómico más próximo a

bacterias de vida libre como *E. coli* así como en la ausencia prácticamente total de divergencia nucleotídica entre cepas de *S. glossinidius* de diferentes especies de mosca tsé-tsé (Chen et al., 1999; Toh et al., 2006), y el cual vemos que presenta unas tasas de inversiones aproximadamente el doble respecto a bacterias de vida libre (2.00 veces superior). La relajación en la presión selectiva sobre un gran número de genes que no serían esenciales en nicho ecológico asociado al hospedador y la proliferación de diferentes tipos de elementos genéticos móviles en estas etapas iniciales de la transición al modo de vida dependiente de hospedador favorecería esta aceleración en las tasas de reordenaciones mediante recombinación homóloga entre secuencias repetidas a lo largo del genoma (Moran and Plague, 2004), mientras que en etapas avanzadas de la reducción genómica, la ausencia de secuencias repetidas conjuntamente con la pérdida de genes implicados en la maquinaria de recombinación explicaría la estabilidad genómica observada entre cepas de *B. aphidicola* (Silva et al., 2003).

Reconstrucción filogenética basada en inversiones y puntos de rotura: Filogenias de orden génico contra filogenias basadas en datos de secuencia

Las matrices de distancias de inversiones y distancias de puntos de rotura se utilizaron para reconstruir la filogenia de γ-proteobacterias con el objetivo de detectar periodos de evolución rápida o lenta a nivel de orden génico a partir de la longitud de las ramas del árbol así como para determinar la relación entre los diferentes linajes de bacterias endosimbiontes y su posición dentro de la filogenia de γ-proteobacterias. Dado que con los datos de orden génico únicamente tenemos una matriz de distancias (inversiones y puntos de rotura) entre genomas, utilizamos los métodos de Neighbor Joining (NJ) (Saitou and Nei, 1987) y Fitch-Margoliash (FM) (Fitch and Margoliash, 1967) para reconstruir las filogenias de orden génico, evaluando la fiabilidad de la reconstrucción filogenética mediante un test de Jacknife que consiste en muestrear al azar 122 de los 244 genes con los que trabajamos, calcular la matriz de inversiones y puntos de rotura, y reconstruir la filogenia con los mismos métodos (FM y NJ). Este proceso se repite 100 veces y se calcula el número de veces que cada nodo del árbol original aparece presente en los 100 árboles del muestreo Jacknife. La filogenia basada en datos de secuencia es inferida por máxima verosimilitud a partir del alineamiento concatenado de 3670 posiciones aminoacídicas utilizando el método de cuartetos implementado en el programa TREEPUZZLE 5.2 (Strimmer and von, 1996; Schmidt et al., 2002).

Tanto FM como NJ producen resultados similares, con altos valores de Jacknife apoyando la mayoría de nodos, y las filogenias inferidas a partir de distancias de inversiones y distancias de puntos de rotura son bastante concordantes, difiriendo únicamente en la posición del grupo de las Shigellas, que en las filogenias de

inversiones aparecen juntamente con el grupo de *E. coli* en un grupo monofilético con las Salmonellas como grupo externo mientras que en las filogenias de puntos de rotura aparecen en posición basal dentro de las entéricas, con *E. coli* y las Salmonellas como grupo monofilético (Ver Figuras 3.5 y 3.6). Las principales diferencias aparecen al comparar las filogenias de orden génico con la filogenia obtenida en base secuencias aminoacídicas a nivel de la posición del grupo de las Pasteurellas y del monofiletismo de los endosimbiontes primarios de insectos. En filogenias de orden génico, las Pasteurellas adquieren una posición basal próxima al grupo externo de las Xanthomonadales, mientras que en la filogenia basada en secuencias aminoacídicas aparece en posición interna formando un grupo monofilético con los endosimbiontes bacterianos de insectos (Ver Figura 3.2). Esto se explica por la aceleración en las tasas de reordenaciones frente al resto de γ-proteobacterias cuando se comparan las distancias de orden génico con las distancias en base a datos de secuencia, que puede producir un fenómeno de atracción de ramas largas (Nei, 1996) que arrastraría a este linaje a una posición basal próxima al grupo externo constituido por las Xanthomonadales que utilizamos para enraizar el árbol. Respecto a los tres linajes de endosimbiontes primarios, éstos aparecen como grupo monofilético en las filogenias basadas en secuencia ocupando una posición cercana al grupo de bacterias entéricas (Ver Figura 3.2), mientras que en filogenias de orden génico este monofiletismo se rompe, con *B. floridanus* que aparece próxima al grupo de *E. coli-Salmonella-Shigella* mientras que *B. aphidicola* y *W. glossinidia* aparecen en posición basal dentro del grupo de bacterias entéricas pero como linajes independientes en lugar de como grupo monofilético (Ver Figuras 3.5 y 3.6). El fenómeno de atracción de ramas largas descrito anteriormente podría explicar la agrupación de los endosimbiontes primarios como grupo monofilético dada la aceleración en las tasas de sustitución característica de estos linajes. De hecho, en la filogenia basada en datos de secuencia muestra el grupo monofilético de los endosimbiontes primarios asociado con ramas significativamente más largas que el resto de linajes de γ-proteobacterias. Se ha demostrado que incluso el uso de métodos de máxima verosimilitud pueden producir filogenias incorrectas siendo la atracción de ramas largas una de las principales fuentes de inconsistencia (Susko et al., 2004; Kolaczkowski and Thornton, 2004), y se ha visto que utilizando modelos evolutivos que asuman tasas de sustitución variables entre linajes se rompe el monofiletismo entre *B. aphidicola* y el resto de endosimbiontes ricos en AT (Herbeck et al., 2005), lo que indica que el monofiletismo de los endosimbiontes primarios continua siendo una cuestión de debate.

Por último, la longitud de las ramas en filogenias de orden génico muestra ejemplos adicionales de la heterogeneidad existente entre linajes a la hora de la fijación de sucesos de reordenación cromosómica, combinando periodos de estabilidad con periodos de rápida fijación de reordenaciones. Este es el caso del elevado número de inversiones que se observan en el linaje de *Shigella flexneri*, con 4 inversiones separando los genomas de *S. flexneri* 2a str. 301 y *S. flexneri* 2a str. 2457T y 9 y 11

inversiones respectivamente respecto a *E. coli* que contrasta con la única inversión entre los genomas de *E. coli* y *S. typhimurium* a pesar de que estos dos linajes se encuentran divergiendo desde hace 100 millones de años (Lawrence and Ochman, 1998), o las 14 inversiones que separan los dos genomas de *Y. pestis* que contrastan con la muy baja distancia de sustitución aminoacídica que se observa entre ambos (0.0003) que concuerda con su muy reciente divergencia (Achtman et al., 1999). Este es el caso también de la rama extremadamente larga de *S. glossinidius* en filogenias de orden génico en comparación con la filogenia basada en datos de secuencia, que se explica por la aceleración en las tasas de reordenaciones que está teniendo lugar en este linaje como consecuencia de la transición al modo de vida dependiente de hospedador.

CAPÍTULO 4: Proliferación de elementos genéticos móviles e inactivación génica en el genoma de *Sodalis glossinidius*, el endosimbionte secundario de la mosca tsé-tsé

En este capítulo se va a analizar de manera más específica la evolución de *S. glossinidius*, el endosimbionte secundario de la mosca tsé-tsé que se encuentra en etapas muy tempranas de la asociación con el insecto hospedador y como hemos visto en el capítulo anterior presenta una aceleración significativa en la tasa de reordenaciones genómicas respecto a bacterias de vida libre, característica típica de etapas iniciales de la asociación. La simbiosis se puede definir como la asociación entre dos o más organismos diferentes durante al menos una parte de su ciclo vital. Bajo este concepto, es posible diferenciar diferentes tipos de relaciones simbióticas en función de los beneficios o posibles desventajas que la relación tiene sobre las especies implicadas, de tal forma que si ambas especies se benefician de la asociación, la simbiosis se define como mutualista, mientras que si una de las especies se beneficia mientras que la otra se ve perjudicada la relación se define como parasitismo. Finalmente, el comensalismo se define como una asociación simbiótica donde una de las especies se beneficia mientras que la otra no se ve perjudicada pero tampoco recibe ningún beneficio (Moran, 2006; Silva et al., 2007; Moya et al., 2008). En la mayoría de casos la asociación se establece entre un eucariota multicelular y un microorganismo como procariotas u hongos unicelulares. Dentro de los simbiontes bacterianos se distinguen dos grandes grupos en base al grado de dependencia de la asociación simbiótica para su supervivencia, pudiendo encontrar simbiontes facultativos que son aquellos capaces de retornar a un estado de vida libre y simbiontes obligados que son aquellos que únicamente pueden sobrevivir en asociación con su correspondiente hospedador eucariota (Wernegreen, 2005). Finalmente, dependiendo de la localización del simbionte microbiano dentro

del hospedador se pueden identificar ectosimbiontes, que son aquellos que viven en diferentes superficies del hospedador como pueden ser bacterias que habitan el tubo digestivo, o endosimbiontes, que son aquellos que habitan en el interior de tejidos del hospedador tanto intracelularmente como extracelularmente (Nardon and Nardon, 1998).

La importancia de las asociaciones simbióticas en la evolución y origen de la vida se demuestra con el origen simbionte de cloroplastos y mitocondrias a partir de una asociación simbiótica ancestral de un eucarionte primitivo con una cianobacteria, que evolucionó hacia los cloroplastos actuales, y una α-proteobacteria, que evolucionó hacia las mitocondrias actuales, de tal forma que la adquisición de caracteres esenciales para el desarrollo de la vida como son la respiración aerobia a través de mitocondrias o la capacidad de fijar dióxido de carbono a oxígeno por fotosíntesis a través de cloroplastos son consecuencia de asociaciones simbióticas (Margulis, 1970; Margulis, 1981). Las asociaciones simbióticas han tenido lugar en múltiples ocasiones a lo largo de la evolución afectando a la mayor parte de filums eucariotas siendo clave para su evolución en términos de especiación y capacidad de colonización de nuevos hábitats (Gortz and Brigge, 1998; Friedrich et al., 1999; Steinert et al., 2000; Newton et al., 2007). Sin embargo, si existe un linaje donde las asociaciones simbióticas han sido especialmente importantes en su evolución es el de los artrópodos, donde más del 10% de especies de insectos en diversos órdenes taxonómicos dependen de asociaciones simbióticas mutualistas obligadas para su viabilidad y reproducción (Wernegreen, 2002; Silva et al., 2007). Como se ha descrito en la introducción general, la evolución de estos endosimbiontes bacterianos de insectos se caracteriza por un proceso de reducción génica masiva como consecuencia de la transición desde un modo de vida libre a un modo de vida dependiente de hospedador, con diferencias significativas en las características y dinámicas de evolución genómica en diferentes momentos de este proceso de reducción genómica (Ver Tabla 4.1 de endosimbiontes bacterianos de insectos con genomas completamente secuenciados). En todos los casos, los endosimbiontes bacterianos de insectos se pueden considerar como bacterias simbiontes obligadas ya que carecen de fase replicativa fuera del hospedador, aunque existen variaciones en el grado de dependencia del hospedador respecto a la asociación endosimbionte para su normal desarrollo y reproducción que en la mayoría de casos se correlaciona con la edad de la asociación. En este contexto ecológico, se definen tres grandes tipos mayoritarios de endosimbiontes bacterianos de insectos que son los endosimbiontes primarios, secundarios, y los parásitos reproductivos (Moya et al., 2008).

Los endosimbiontes primarios se corresponden con endosimbiontes mutualistas obligados esenciales para el normal desarrollo del hospedador, como es el caso de *B. aphidicola* en pulgones, *W. glossinidia* en mosca tsé-tsé, o *Blochmania spp*. en las hormigas carpintero. Estos endosimbiontes primarios representan asociaciones ancestrales altamente integradas en la biología del hospedador, donde el

endosimbionte reside intracelularmente en células especializadas del hospedador llamadas bacteriocitos que forman un órgano llamado bacterioma que se deriva de diferentes tejidos del hospedador dependiendo del linaje (Silva et al., 2007). Como se ha descrito también en la introducción, la mayoría de estas asociaciones tiene un propósito nutricional, de tal forma que la bacteria le proporciona al hospedador determinados metabolitos que el insecto es incapaz de producir por sí mismo o asimilar a partir de su dieta. Estos endosimbiontes primarios se transmiten verticalmente por línea materna desde las madres a sus descendientes, aunque existen diferentes rutas de transmisión dependiendo del insecto hospedador, como puede ser la infección de oocitos en endosimbiontes bacterianos que dejan el bacterioma y penetran a través del ovario en hormigas y pulgones partenogenéticos, la estricta localización en ovarios y células de la línea germinal en especies de gorgojo del género *Sitophilus*, o transmisión a través de secreciones de un órgano específico llamado "*milk gland*" utilizado por las hembras para alimentar a las larvas en desarrollo como en moscas tsé-tsé del género *Glossina* (Nardon and Nardon, 1998; Wernegreen, 2002; Baumann, 2005). Esta transmisión estrictamente vertical hace que sus poblaciones sufran una drástica reducción de su tamaño poblacional efectivo consecuencia del pequeño número de bacterias endosimbiontes que se transmiten a los descendientes en cada ciclo reproductivo, que en *B. aphidicola* se ha estimado en 850-8000 células en diferentes especies de pulgón (Mira and Moran, 2002), lo que hace que incremente el efecto de la deriva genética sobre la selección natural y que se fijen mutaciones ligeramente deletéreas simplemente por azar que no podrán ser revertidas a su estado inicial debido a la estricta localización intracelular de estos endosimbiontes primarios, un proceso conocido como trinquete de Muller (Moran, 1996). La transmisión estrictamente vertical de estos endosimbiontes primarios conjuntamente con la ausencia de transferencia horizontal entre linajes por su estricta localización intracelular genera un proceso característico de coevolución entre el endosimbionte primario y su correspondiente insecto hospedador que se ha utilizado para datar la edad de estas asociaciones a partir de datos del registro fósil del insecto hospedador (Moran et al., 1993; Ochman et al., 1999; Chen et al., 1999; Funk et al., 2000).

En contraste con los endosimbiontes primarios, los endosimbiontes secundarios de insectos presentan una asociación mucho más reciente con el insecto hospedador que hace que no sean igual de esenciales para la supervivencia del hospedador, presentando una asociación de tipo facultativo no mutualista. Muchos de estos endosimbiontes secundarios comparten características similares a bacterias patógenas en el sentido de que son capaces de invadir diferentes tipos celulares incluyendo tejidos reproductivos, son capaces de residir tanto intracelularmente como extracelularmente en la hemolinfa del insecto hospedador, y aunque presentan una herencia principalmente vertical, su distribución y patrones filogenéticos muestran que se pueden transmitir horizontalmente entre individuos y especies de hospedador (Fukatsu et al., 2000; Tsuchida et al., 2002; Haynes et al., 2003; Moran

et al., 2005b; Russell et al., 2003). La introducción experimental de estos endosimbiontes secundarios en insectos que no los albergan normalmente ha revelado que son capaces de establecer infecciones permanentes capaces de transmitirse verticalmente a los descendientes, lo que indica que su mantenimiento en el hospedador se produce por mecanismos específicos de la bacteria como la presencia de sistemas de secreción tipo III específicamente adaptados para la invasión de células del hospedador como en el caso de *S. glossinidius* en la mosca tsé-tsé (Dale et al., 2001; Dale et al., 2002; Dale et al., 2005). No obstante, muchos de estos endosimbiontes secundarios confieren características selectivamente ventajosas al insecto hospedador que explican su mantenimiento y persistencia a lo largo de la evolución del hospedador, como la protección del hospedador frente a invasiones de parásitos en el caso de *Hamiltonella defensa* en pulgones (Oliver et al., 2003; Scarborough et al., 2005) o el incremento de su tolerancia a altas temperaturas también por *H. defensa* y *Serratia symbiotica* en pulgones (Russell and Moran, 2006).

Por último, además de los endosimbiontes primarios y secundarios, muchas especies de insectos albergan bacterias parásitas que no les confieren ningún beneficio pero que manipulan la biología reproductiva del hospedador de diversas formas para favorecer su transmisión, como produciendo fenómenos de incompatibilidad citoplásmica donde machos infectados por el parásito esterilizan a hembras no infectadas, favoreciendo la transmisión a la descendencia del parásito, o induciendo partenogénesis, feminización de machos genéticos o la muerte de descendientes machos de hembras infectadas para favorecer la proporción de descendientes infectados con el parásito (Moran et al., 2008). El ejemplo más claro y predominante de este tipo de parásitos reproductivos son las α-proteobacterias del género *Wolbachia*, que es capaz de inducir todos los mecanismos descritos anteriormente y que se ha demostrado que infectan a más del 65% de especies de insectos, de tal forma que representan la bacteria intracelular más abundante que se conoce (Dobson et al., 1999; Stouthamer et al., 1999; Brownlie and O'Neill, 2005; Werren et al., 2008).

Estos diferentes tipos de endosimbiontes bacterianos de insectos (primarios, secundarios y parásitos reproductivos) no son excluyentes entre ellos, y es habitual la presencia de diferentes linajes de endosimbiontes dentro de un mismo hospedador con diferentes características ecológicas y evolutivas que constituyen asociaciones estables que favorecen la eficacia del hospedador, como en el caso de pulgones donde, además del endosimbionte primario *B. aphidicola* se encuentran presentes diferentes tipos de endosimbiontes secundarios como *Serratia symbiotica*, *Hamiltonella defensa* y *Regiella insecticola* (Sandstrom et al., 2001; Russell and Moran, 2006) conjuntamente con *Wolbachia* (Gomez-Valero et al., 2004b), y donde en el caso particular del pulgón del cedro *Cinara cedri*, se ha demostrado que el endosimbionte secundario *Serratia symbiotica* es necesario conjuntamente con *B.*

aphidicola para producir triptófano (Gosalbes et al., 2008). Otro ejemplo de complementación metabólica lo encontramos en la chicharrita de alas cristalinas *Homalodisca coagulata*, donde sus dos endosimbiontes primarios *Sulcia mulleri* y *Baumannia cicadellinicola* presentan actividades metabólicas complementarias esenciales para la supervivencia del hospedador Wu, 2006 18 /id;McCutcheon, 2007 6 /id}, lo que demuestra que diferentes tipos de interacciones más o menos complejas se pueden establecer entre las diferentes bacterias endosimbiontes que habitan un determinado hospedador y que es posible caracterizar de manera funcional a partir de la información de los genomas completamente secuenciados. Este es el caso del sistema endosimbionte que se establece entre moscas tsé-tsé del género *Glossina* (*Diptera: Glossinidae*), donde encontramos un endosimbionte primario ancestral que es *W. glossinidia* conjuntamente con un endosimbionte secundario de asociación mucho más reciente que es *S. glossinidius*, ambos con los genomas completamente secuenciados (Akman et al., 2002; Toh et al., 2006) y cuyas dinámicas evolutivas a nivel de reordenaciones cromosómicas se han analizado en el capítulo tercero de esta tesis. Las moscas tsé-tsé comprenden 31 especies y subespecies diferentes dentro del género *Glossina* que se clasifican en tres grupos (Morsitans, Palpalis y Fusca) en base a diferencias morfológicas en la estructura del aparato reproductor (Gooding and Krafsur, 2005), son insectos hematófagos estrictos, alimentándose exclusivamente de sangre de vertebrados, y son los vectores de transmisión mayoritarios de diferentes especies de tripanosomas patogénicos que son los agentes causales de la enfermedad del sueño en humanos y el nagana en ganado, siendo transmitidos a través de la mordedura de la mosca (Aksoy, 2000; Aksoy et al., 2003; Aksoy et al., 2005). El endosimbionte primario *W. glossinidia* es un endosimbionte mutualista obligado perteneciente a las γ-proteobacterias con una asociación ancestral con la mosca tsé-tsé, donde reside de en el interior de bacteriocitos que forman el órgano del bacterioma en la parte anterior del intestino (Aksoy, 1995). Su genoma se secuenció en 2002 por Akman y colaboradores, y muestra todas las características típicas de un endosimbionte primario ancestral, con un tamaño genómico reducido de 697 kilobases y un marcado sesgo composicional hacia AT (78% AT), con únicamente 621 genes codificantes de proteínas consecuencia de una pérdida de genes masiva desde la divergencia del ancestro de vida libre que incluye la mayoría de genes implicados en biosíntesis de aminoácidos, aunque retiene una gran fracción de genes implicados en biosíntesis de cofactores (Akman et al., 2002). En este contexto, la mosca tsé-tsé se alimenta de manera exclusiva de sangre de vertebrados, la cual es deficiente en vitaminas, de tal forma que la asociación con *W. glossinidia* tendría este propósito nutricional, el suplemento de vitaminas a la mosca tsé-tsé (Nogge, 1976; Nogge, 1981). La asociación con *W. glossinidia* es esencial para la supervivencia de la mosca tsé-tsé, ya que moscas aposimbióticas muestran un retraso significativo de crecimiento y una pérdida de capacidad reproductora (Aksoy, 1995; Dale and Welburn, 2001). El carácter ancestral de la asociación entre *W. glossinidia* y la

mosca tsé-tsé lo demuestra la estricta coevolución entre *W. glossinidia* y sus correspondientes especies hospedadoras, lo que hace que se haya podido datar la edad de esta asociación simbiótica entre 50-100 millones de años, donde se produciría una infección del hospedador ancestral que iría seguida de divergencia de los diferentes linajes de *Glossina* y sus correspondientes endosimbiontes primarios sin sucesos de transferencia genética horizontal entre linajes (Chen et al., 1999; Akman et al., 2002). Además de *W. glossinidia*, todas las moscas tsé-tsé de poblaciones de campo albergan un endosimbionte secundario que es *S. glossinidius*, también perteneciente a las γ-proteobacterias, con una asociación de tipo facultativo con la mosca tsé-tsé, pudiendo residir tanto intracelularmente como extracelularmente en diferentes tejidos del hospedador como intestino, músculos, cuerpo graso, glándulas salivales y "*milk gland*", variando significativamente la localización y la densidad poblacional entre diferentes linajes de mosca tsé-tsé, aunque su abundancia es tres órdenes de magnitud menor que *W. glossinidia* (Dale and Maudlin, 1999; Cheng and Aksoy, 1999; Rio et al., 2006). Análisis filogenéticos basados en el gen 16S ribosomal evidencian una transición muy reciente al modo de vida dependiente de hospedador ya que se observa una divergencia mínima a nivel de secuencia entre diferentes cepas de *S. glossinidius* que hace que no se produzca coevolución entre *S. glossinidius* y sus correspondientes especies hospedadoras (Chen et al., 1999), aunque si que se ha detectado diferencias significativas entre poblaciones de *S. glossinidius* de diferentes especies de *Glossina* a nivel de otros marcadores moleculares basados en la PCR como AFLP´s (Geiger et al., 2005; Geiger et al., 2007). Esta transición muy reciente al modo de vida dependiente del hospedador también se refleja en el hecho de que *S. glossinidius* depende de la actividad específica de dos tipos de sistemas de secreción tipo III para invadir las células del hospedador y proliferar en su interior (Dale et al., 2001; Dale et al., 2002; Dale et al., 2005), así como en el hecho de que es uno de las pocas bacterias endosimbiontes capaces de cultivarse in-vitro tanto en células de insecto como en cultivo sólido (Welburn et al., 1987; Beard et al., 1993), siendo posible introducir *S. glossinidius* en moscas tsé-tsé libres de endosimbionte secundario sin afectar a su eficacia biológica y establecer infecciones permanentes transmisibles a la descendencia al mismo nivel que en poblaciones naturales (Weiss et al., 2006). Además, la capacidad de tener cultivos celulares puros de *S. glossinidius* ha sido utilizada para desarrollar sistemas de transformación genética que permiten modificar genéticamente a *S. glossinidius* con un gen de interés, el cual se expresará en la mosca tse-tsé al infectarla con *S. glossinidius* recombinante, una estrategia que recibe el nombre de paratransgénesis y que se ha utilizado para expresar en la mosca tse-tsé productos con actividad anti-tripanosomas capaces de controlar su infección (Aksoy et al., 2003; Rio et al., 2004; Aksoy and Rio, 2005). Finalmente, la eliminación de *S. glossinidius* de la mosca tsé-tsé mediante tratamiento con estreptozoina tiene muy poco efecto sobre la fecundidad de la mosca tsé-tsé, aunque la longevidad de los descendientes aposimbioticos de moscas tsé-tsé tratadas con

estreptozoina se ve significativamente reducida (Dale and Welburn, 2001). En concordancia con su carácter facultativo y con su reciente asociación con la mosca tsé-tsé, el papel biológico de *S. glossinidius* en el contexto de la asociación endosimbionte es menos claro que el del endosimbionte primario *W. glossinidia*. Se ha visto que *S. glossinidius* potencia la susceptibilidad de las moscas tsé-tsé a infecciones por tripanosomas debido a que *S. glossinidius* segrega chitinasas esenciales para su mantenimiento en el hospedador que degradan la matriz peritrófica que envuelve el tubo digestivo de la mosca tsé-tsé, liberando monómeros del aminoazúcar N-acetilglucosamina el cual tiene un efecto inhibidor sobre lectinas segregadas por la mosca tsé-tsé con actividad anti-tripanosoma, aunque este es un fenómeno que afecta únicamente a moscas en estado anterior a su primera alimentación (Maudlin and Welburn, 1987; Mihok et al., 1992; Welburn and Maudlin, 1992; Welburn et al., 1993).

La evidencia final de la muy reciente asociación de *S. glossinidius* con la mosca tsé-tsé vino con la secuenciación completa de su genoma, que se inició con la secuenciación de sus elementos extracromosómicos, consistentes en tres plásmidos comunes en diferentes linajes y un elemento extracromosómico con características fágicas y con presencia y estructura variable entre linajes (Darby et al., 2005; Clark et al., 2007), y con la publicación en 2006 de su genoma completo consistente en un cromosoma circular de 4,3 megabases, mucho más cercano a una bacteria de vida libre como *E. coli* que a los genomas de endosimbiontes mutualistas obligados, sin sesgos composicionales hacia AT (54% GC frente al 22% de *W. glossinidia*), pero con una densidad codificante extremadamente baja (50% del genoma) consecuencia de un proceso masivo de inactivación génica que se refleja en la presencia de un total de 972 pseudogenes que fueron descritos pero no anotados en la anotación original del genoma disponible en las bases de datos (Toh et al., 2006). Desde el punto de vista evolutivo, *S. glossinidius* representa uno de los pocos genomas completamente secuenciados de una bacteria en etapas tan tempranas de la asociación con el hospedador de tal forma que tenemos la huella del genoma ancestral con los genes y pseudogenes, y a la luz de los resultados obtenidos en el capítulo anterior su genoma presenta unas dinámicas de reordenaciones cromosómicas significativamente aceleradas respecto a bacterias entéricas de vida libre, de tal forma que en este capítulo vamos a estudiar más en profundidad este genoma a través de la caracterización de diferentes tipo de elementos genéticos móviles y la evaluación de su impacto relativo sobre las capacidades funcionales de *S. glossinidius*, así como el impacto que el proceso de inactivación génica esta teniendo en *S. glossinidius* en el contexto de su interacción tanto con la mosca tsé-tsé como con el endosimbionte primario *W. glossinidia*, para lo cual se lleva a cabo una reanotación completa de su genoma incluyendo la caracterización *de-novo* de todos sus pseudogenes.

Resultados y discusión

Caracterización y ajuste del número de pseudogenes de *S. glossinidius*

Como se ha apuntado anteriormente, la anotación original de *S. glossinidius* postula la existencia de 972 pseudogenes (Toh et al., 2006). Sin embargo, ni sus coordenadas ni la función de los genes de los que se derivan aparecen anotados ni en el artículo del genoma ni en los archivos disponibles en las bases de datos, de tal forma que se ha llevado a cabo un completo análisis de las regiones intergénicas de *S. glossinidius* con el objetivo de caracterizar estos pseudogenes. Primero se llevaron a cabo BLASTX (Altschul et al., 1997) con las secuencias intergénicas contra el proteoma de todos los genomas completamente secuenciados hasta Diciembre de 2007, lo que nos permitió delimitar las coordenadas de un total de 1724 pseudogenes potenciales a partir de los cuales se utilizó el programa Genewise para predecir sus posibles pautas abiertas de lectura (Birney et al., 2004). En segundo lugar, la inspección manual de los resultados de BLASTP y FASTA sobre cada CDS permitió identificar pseudogenes adyacentes correspondientes con diferentes pautas de un mismo gen ancestral, de tal forma que estos pseudogenes fueron unidos en un único CDS. En tercer lugar, la caracterización de elementos IS y el análisis de sus secuencias adyacentes permitió identificar pseudogenes producto de inserción de un IS, donde se detectan diferentes pautas de un mismo pseudogen a ambos lados del IS, de tal forma que estas pautas fueron unidas en un único CDS. En cuarto lugar, durante el proceso de re-anotación de todo el genoma (genes y pseudogenes) se detectaron 142 situaciones en las que un putativo gen funcional según la anotación original se había dividido en una pauta abierta de lectura que incluía el gen originalmente anotado y un pseudogen adyacente detectado en base a los análisis anteriores. Estos pseudogenes se eliminaron del conjunto final pero se añadió un marcador al correspondiente gen especificando la proporción del gen ancestral representada por el gen y el pseudogen adyacente. Estas 142 situaciones se describen en la Tabla suplementaria 4.2. Por último. Pseudogenes que no presentan similitudes significativas en búsquedas de BLASTP y FASTA sobre los CDS predichos por Genewise y sin ninguna posible asignación funcional mediante búsquedas de dominios proteicos en bases de datos PFAM (Finn et al., 2009) y Prosite (Sigrist et al., 2009) o mediante predicción de localización celular por predicción de dominios transmembrana con TMHMM (Chen et al., 2003) o por predicción con PSORT (Gardy et al., 2003) fueron eliminados del conjunto de pseudogenes final. El resultado de estos análisis es un total de 1501 pseudogenes caracterizados tanto a nivel nucleotídico como aminoacídico, lo que supone un incremento de 529 genes respecto a los 972 pseudogenes que se describieron pero no se anotaron en la anotación original. Esto se explica por las diferentes metodologías que se han seguido para caracterizar los pseudogenes, ya que en la anotación original los pseudogenes se identificaron a partir de los resultados de programas de predicción de genes ab-initio, considerando como pseudogen todo aquel CDS con menos de la mitad de la longitud de su correspondiente homólogo identificado por BASTP (Toh et al., 2006), lo cual limita la identificación de pseudogenes a sucesos de

inactivación lo suficientemente recientes como para que los programas de predicción puedan detectar pautas abiertas de lectura con codón de inicio y de parada, mientras que con nuestra aproximación basada en el análisis de regiones intergénicas directamente por BLASTX es posible identificar pseudogenes altamente degradados que no se identifican por la aproximación anterior así como pseudogenes inactivados por inserción de elementos IS. Este incremento significativo del número de pseudogenes respecto a la anotación original también se ha detectado en diversos estudios que utilizan aproximaciones similares con otros genomas bacterianos (Liu et al., 2004; Lerat and Ochman, 2004; Lerat and Ochman, 2005).

Re-anotación funcional del genoma de *S. glossinidius*

A partir de la secuencia de los 1501 pseudogenes caracterizados en el apartado anterior y de los 2431 genes anotados originalmente se llevó a cabo una re-anotación funcional completa del genoma combinando información obtenida de búsquedas por similitud con BASTP y FASTA, búsquedas de dominios proteicos (PFAM, PROSITE, TMHMM, y SIGNALP) y predicciones de localización celular (PSORT). En la tabla suplementaria 4.1 se describe el esquema de clasificación funcional que se ha utilizado durante el proceso de re-anotación, que ha sido adoptado de la Unidad de Secuenciación de Patógenos (PSU) del instituto Sanger en Cambridge (UK). Los resultados de la re-anotación se resumen en la Figura 4.3. La categoría funcional con mayor número de CDS (genes y pseudogenes) es la de elementos genéticos móviles (831 CDSs), seguida por las proteínas de cubierta celular (770 CDSs). Estas categorías funcionales representan el 21.1% y el 19.6% respectivamente del total de CDS de *S. glossinidius*. La clase funcional de proteínas de superficie celular incluye genes codificantes de proteínas integrales de membrana (interna y externa), proteínas de secreción, transportadores de membrana, así como los genes implicados en la síntesis de componentes de la superficie celular como el lipopolisacárido bacteriano o el peptidoglicano. La categoría funcional de elementos genéticos móviles incluye todos los CDSs correspondientes a transposasas y diferentes componentes de elementos fágicos como integrasas, transcriptasas inversas, o proteínas fágicas hipotéticas. Estas dos categorías funcionales mayoritarias son las que albergan un mayor número de pseudogenes, con 447 pseudogenes asociados a elementos genéticos móviles y 345 pseudogenes asociados a proteínas de superficie celular, siendo las clases funcionales más afectadas por la pseudogenización (53.8% y 44.8% del total del CDS respectivamente). En el lado opuesto, las categorías funcionales con el menor número de CDS se corresponden con degradación de pequeñas y grandes moléculas (91 y 106 CDSs respectivamente), de acuerdo con las observaciones originales de que *S. glossinidius* retiene mayoritariamente rutas biosintéticas en lugar de degradativas (Toh et al., 2006).

La anotación original de *S. glossinidius* incluye un total de 787 genes que aparecen anotados como proteínas hipotéticas. Después de la re-anotación, 190 genes aparecen anotados como proteínas hipotéticas conservadas presentes en otros genomas próximos pero para las cuales no es posible inferir ninguna función, mientras 115 genes aparecen anotados como proteínas hipotéticas o "unknown proteins", que se corresponden con genes que no presentan similitud significativa con ningún otro gen de las bases de datos. Este último resultado contrasta con los 221 genes descritos en la anotación original de *S. glossinidius* sin homología con ningún gen de las bases de datos (Toh et al., 2006), lo que se puede explicar por el incremento exponencial en el número de genomas disponibles en las bases de datos sobre los cuales se llevan a cabo las búsquedas de similitud así como la utilización de diferentes fuentes de información funcional además de las búsquedas por similitud por FASTA y BASTP. El resumen de los resultados de la re-anotación sobre los 787 genes anotados originalmente como proteínas hipotéticas se representa en la Figura 4.4.

Caracterización de elementos fágicos

Debido a la presencia masiva de CDSs de origen fágico y su agrupamiento en regiones concretas del genoma detectado durante el proceso de re-anotación, se llevó a cabo un análisis mas detallado de estos CDSs. La clase funcional 5.1.2 correspondiente a profagos y funciones de origen fágico contiene un total de 698 genes y pseudogenes que representan un 17.8% de los CDSs de *S. glossinidius*, siendo la clase funcional más afectada por la pseudogenización con 353 pseudogenes. El origen fágico de estos genes y pseudogenes fue corroborado con un BLASTP adicional de los 698 genes y pseudogenes anotados como 5.1.2 contra la base de datos ACLAME específica de elementos genéticos móviles (http://aclame.ulb.ac.be/). Con el objetivo de caracterizar completos o parciales se llevó a cabo un TBLASTX del genoma completo de *S. glossinidius* contra todos los genomas de fagos completamente secuenciados disponibles en GenBank y los resultados fueron visualizados con la herramienta Artemis Comparison Tool (ACT) (Carver et al., 2005) para detectar regiones en las que se conservara no solo homología a nivel de genes individuales sino homología a nivel de genoma fágico completo. Este análisis revela la presencia de dos profagos completos diferentes pertenecientes a la subdivisión Mu de bacteriófagos de doble cadena en el genoma de *S. glossinidius* que se representan en la Figura 4.4. La primera región se encuentra comprendida entre los pseudogenes ps_SGL0195 y ps_SGL0213, y muestra sintenia a escala genómica con el bacteriófago enterobacteriano Mu (NC_000929.1), aunque la pseudogenización ha afectado a 19 de los 28 CDSs que incluye esta región incluyendo los genes responsables de la transposición, integración, regulación del ciclo lítico y lisogénico, así como la mayoría de genes implicados en el ensamblaje de la cápside viral (Morgan et al., 2002), indicando la inactividad de este elemento, que se encontraría en proceso de desintegración. La

segunda región se incluye 47 CDSs comprendidos entre el gen SG0816 y el pseudogen ps_SGL0453, y muestra sinténia a escala genómica con el bacteriófago *BcepMu* (NC_005882) aislado en *Burkholderia cenocepacia* str. J2315, que se encuentra integrado también en los genomas de *Salmonella typhi* CT18 y Ty2, *Photorhabdus luminescens* y *Chromobacterium violaceum* conservando un elevado grado de homología entre a lo largo de toda la longitud del profago (Summer et al., 2004). En contraste con la región fágica anterior, en esta región únicamente 5 de los 47 CDSs se corresponden con pseudogenes, reteniendo genes funcionales para la mayoría de actividades fágicas, aunque no se conserva la región de control esencial para el desarrollo del ciclo lítico, indicando que se trata de un profago críptico no-inducible (Summer et al., 2004). Además de estos dos profagos completos, las comparaciones por BLASTX revelan la presencia de 11 regiones con homología significativa de diferentes dominios fágicos. Dos de estas regiones fágicas parciales presentan homología con el profago Epsilon 15 (NC_004775.1), el cual se ha postulado como uno de los precursores del elemento extracromosómica de origen fágico pSG3 de *S. glossinidius* str. *Morsitans* (Clark et al., 2007), indicando un posible flujo de material genético entre el cromosoma y este elemento extracromosómica pSG3. Este elemento presenta una elevada plasticidad entre diferentes cepas de *S. glossinidius*, con diferencias de tamaño y organización en *S. glossinidius* de *Glossina morsitans morsitans* y *Glossina palpalis palpalis*, estando ausente en *S. glossinidius* de *Glossina austeni* (Darby et al., 2005; Clark et al., 2007), con lo cual es de esperar un contenido variable de elementos fágicos entre los genomas de diferentes cepas de *S. glossinidius*. Las características principales de las 13 regiones fágicas completas y parciales caracterizadas aparecen representadas en la Tabla 4.2.

Caracterización de Secuencias de Inserción (IS)

En la anotación original de *S. glossinidius*, 29 genes aparecen anotados como posibles transposasas aunque sin ninguna información adicional acerca del tipo de transposasa o del IS al que pertenece. La proliferación de elementos IS se considera uno de los fenómenos característicos de las etapas iniciales de la transición al modo de vida dependiente de hospedador, asociado a la disminución de la presión selectiva sobre un gran número de genes del genoma que pasan a ser no esenciales en el entorno más estable asociado al hospedador eucariota (Moran and Plague, 2004; Silva et al., 2007; Touchon and Rocha, 2007). Además, la proliferación de elementos IS ha sido extremadamente elevada en los genomas de los endosimbiontes primarios de las especies de gorgojo *Sitophilus oryzae* (SOPE) y *Sitophilus zeamays* (SZPE), que constituyen los linajes más próximos a *S. glossinidius* en reconstrucciones filogenéticas y también se encuentran en etapas iniciales de la asociación con el hospedador aunque son endosimbiontes primarios obligados en contra del carácter facultativo de *S. glossinidius* en el contexto de su asociación con la mosca tsé-tsé (Plague et al., 2008; Dougherty and Plague, 2008; Gil et al., 2008).

Con el objetivo de evaluar el impacto relativo de la proliferación de elementos IS sobre el genoma de *S. glossinidius* se ha llevado a cabo una caracterización del los diferentes tipos de IS que se encuentran presentes en este genoma. Cinco diferentes tipos o familias de elementos IS se han caracterizado en el genoma de *S. glossinidius* que representan un 2,52% del genoma. Estas cinco familias muestran homología significativa con elementos IS de γ-proteobacterias conocidos. Las principales características estructurales y evolutivas de estas cinco familias de elementos IS mayoritarias se describen en la Tabla 4.3.

Existen diferencias significativas entre los diferentes tipos de elementos IS a nivel de su funcionalidad y de los niveles de divergencia nucleotídica entre copias que indican diferentes estados de degradación de elementos IS. De las cinco familias de elementos IS caracterizadas, únicamente dos de ellas (IS_Sgl1 y IS_Sgl2) se pueden considerar como funcionales en el sentido de que contienen genes codificantes de transposasas plenamente funcionales y conservados entre copias de un mismo IS y con dominios proteicos PFAM completos correspondientes a transposasas (ver Figura 4.5). El elemento IS más abundante en *S. glossinidius* es el IS_Sgl1, que pertenece a la familia IS5 según el esquema de clasificación de la base de datos ISfinder (Siguier et al., 2006b), y que contiene una transposasa plenamente funcional de 307 aminoácidos con un dominio DDE completo característico de la mayoría de transposasas necesario para su correcto funcionamiento (Mahillon and Chandler, 1998). De los 16 genes originales que codifican para una putativa transposasa perteneciente a IS_sgl1, sólo 6 contienen un dominio DDE plenamente funcional y pueden considerarse como elementos IS potencialmente funcionales, mientras que el resto no contienen este dominio o lo contienen truncado o interrumpido, lo que indicaría que no son funcionales. Este elemento IS_Sgl1 presenta un 82% de identidad a nivel nucleotídico con el elemento ISsope1 de SOPE, siendo la única similitud entre los elementos IS de *S. glossinidius* y las cuatro familias de elementos IS descritas en SOPE por Gil y colaboradores (Gil et al., 2008), indicando un posible origen común de IS_Sgl1 y ISsope1 previo a la divergencia de ambos linajes mientras que el resto de elementos IS de cada genoma representarían adquisiciones posteriores en cada linaje. El otro tipo de elemento IS que contiene transposasas funcionales es el IS_Sgl2, perteneciente a la familia IS110 de elementos IS, que se caracteriza por la ausencia de repeticiones invertidas flanqueando al gen de la transposasa en la mayoría de sus miembros así como por la ausencia de repeticiones directas flanqueando al elemento IS completo después de su transposición (Mahillon and Chandler, 1998). Esto también se observa en las copias de IS_Sgl2, sin repeticiones invertidas ni directas flanqueando los 9 elementos IS_Sgl2 completos presentes en el genoma de *S. glossinidius*, aunque si que se detectan repeticiones invertidas de 11 nucleótidos internas solapantes con el gen de la transposasa (Figura 4.5). Los 6 genes originales codificantes de putativas transposasas pertenecientes a elementos IS_Sgl2 presentan el mismo perfil de dominios proteicos, con un dominio completo característico de elementos IS110

(PF02731 en base de datos PFAM) y un dominio transposasa adicional parcial (PF01548), siendo todos potencialmente funcionales excepto en uno de ellos, perteneciente a un IS_Sgl2 parcial, que muestra un codón de parada prematuro en comparación con el resto de genes de transposasas de IS_Sgl2 que indicaría que se encuentra en proceso de degeneración. Para el resto de elementos IS caracterizados (IS_Sgl3, IS_Sgl4, IS_Sgl5) no es posible identificar una transposasa funcional común a todas las copias, lo que indica que se tratarían de elementos no funcionales en vías de desaparición. La familia IS_Sgl5 es la más divergente de todas las familias de elementos IS caracterizadas, con unas diferencias nucleotídicas entre copias (posiciones diferentes/posiciones totales en alineamientos) que son un orden de magnitud superiores a las que se observan para el resto de familias. Esta familia incluye 8 genes originalmente anotados como proteínas hipotéticas que presentan homología significativa con dominio transposasa 31 de PFAM (PF04754), aunque no es posible identificar repeticiones invertidas ni directas flanqueantes ni tan siquiera definir una secuencia consenso ya que no existe similitud a nivel de secuencia entre copias fuera del gen de la transposasa. Análisis por BLASTX contra la base de datos del ISfinder únicamente detecta similitud significativa con un ISplu15 de Photorhabdus luminescens perteneciente a la familia ISNCY donde se incluyen elementos IS pobremente caracterizados (Mahillon and Chandler, 1998). Sin embargo, las transposasas de IS_Sgl5 presentan un 75% de identidad a nivel aminoacídico con una transposasa codificada en el elemento extracromosómica pSG3 de origen fágico, lo cual conjuntamente con la presencia de dos dominios fágicos épsilon 15 anteriormente descritos refuerza la idea de un posible flujo de elementos genéticos móviles entre el cromosoma y el elemento extracromosómica pSG3.

Además de estas 5 familias de elementos IS mayoritarias, durante el proceso de re-anotación se caracterizaron 3 elementos IS completos y parciales presentes en copia única, lo que incrementa la fracción del genoma de *S. glossinidius* correspondiente a elementos IS a un 2.71%. BLASTX con las secuencias flanqueantes de todos los elementos IS identificados contra bases de datos de proteínas no redundantes permitió identificar pseudogenes generados como consecuencia de la inserción de un elemento IS, donde detectamos el mismo perfil de resultados del BLASTX a ambos lados del elemento IS. Este análisis revela que de los 1501 pseudogenes caracterizados, solo 18 de ellos se originaron por inserción de un elemento IS (, afectando principalmente a proteínas de superficie celular (7 pseudogenes), elementos genéticos móviles (3 pseudogenes), degradación de moléculas pequeñas (3 pseudogenes) y metabolismo central e intermediario (2 pseudogenes), lo que indica que la proliferación de elementos IS no ha sido un factor excesivamente importante en el proceso de inactivación génica en *S. glossinidius*.

Reconstrucción metabólica

A partir de los resultados de la re-anotación de genes y pseudogenes, se ha llevado a cabo una reconstrucción de las capacidades metabólicas de *S. glossinidius* mediante la reconstrucción de rutas metabólicas producto del análisis con los programas KAAS y Blast2GO (Moriya et al., 2007; Gotz et al., 2008) y su posterior refinamiento manual en base a búsquedas bibliográficas y en bases de datos específicas de metabolismo como ECOCYC y METACYC, lo que permitió identificar diferentes características funcionales no descritas en la anotación original. El perfil metabólico de *S. glossinidius* se corresponde con una bacteria aerobia heterotrófica típica, capaz de producir energía a partir de diferentes fuentes de carbono como Glucosamina y N-acetilglucosamina, Manosa, y Manitol a través de sistemas PTS funcionales para todos estos azúcares, aunque la pseudogenización a afectado a los sistemas PTS de glucosa, fructosa, maltosa, sucrosa, celobiosa, N-acetilgalactosamina, galactitol y N-ascorbato. *S. glossinidius* presenta un metabolismo central completamente funcional, aunque se han perdido la mayoría de actividades isoenzimáticas ya que la pseudogenización afecta a genes codificantes de diferentes isozimas, reteniéndose en la mayoría de casos el gen funcional codificante de la isozima metabólicamente más eficaz, como es el caso de la fosfoglicerato mutasa glicolítica, donde la pseudogenización afecta a la isozima menor II (*gpmB*) mientras se retiene la isozima mayoritaria I (*gpmA*), o la pseudogenización de los genes codificantes de la transketolasa A (*tktA*) y la transaldolasa B (*talB*) de la ruta de las pentosas fosfato mientras se retienen genes funcionales codificantes de las isozimas mayoritarias transketolasa B (*tktB*) y transaldolasa A (*talA*). A nivel biosintético, *S. glossinidius* es capaz de sintetizar la mayoría de componentes metabólicos esenciales (aminoácidos, cofactores, nucleótidos, lípidos y fosfolípidos) así como las principales estructuras de envuelta celular como el peptidoglicano, diferentes polisacáridos extracelulares como el ácido colánico y el antígeno común de enterobacterias (ECA) o el lipopolisacárido bacteriano completo con la excepción del antígeno O extracelular que se relaciona con la adaptación de *S. glossinidius* al sistema inmune del hospedador (Toh et al., 2006).

La diferencia principal respecto a la anotación original que ha revelado la re-anotación del genoma de *S. glossinidius* es la inactivación de la ruta de biosíntesis del aminoácido esencial L-arginina (Figura 4.10). En la anotación original se postula que *S. glossinidius* es capaz de sintetizar todos los aminoácidos con la excepción de L-alanina (Toh et al., 2006). Por el contrario, la re-anotación revela que *S. glossinidius* es capaz de sintetizar L-alanina tanto a partir de L-valina y piruvato a través de una piruvato-alanina aminotransferasa codificada por el gen *avtA* o a partir de cisteína a través de una cisteína desulfurasa codificada por el gen *iscS*. Sin embargo, *S. glossinidius* es incapaz de sintetizar L-arginina; la L-arginina se sintetiza a partir de L-glutamato en ocho reacciones enzimáticas, con cinco de ellas necesarias para producir L-ornitina desde L-glutamato y los últimos tres pasos necesarios para producir L-arginina desde L-ornitina y carbamoil fosfato, un

intermediario metabólico esencial necesario para la biosíntesis de pirimidinas (Cunin et al., 1986). En *S. glossinidius*, la pseudogenización ha afectado al primer (*argA*), tercer (*argC*), cuarto (*argD*) y séptimo (*argG*) pasos de la ruta (Figura 4.10), lo que indica una incapacidad para sintetizar L-arginina de-novo. Esto indica que *S. glossinidius* necesita incorporar L-arginina a partir de su entorno, lo cual se puede llevar a cabo a través de un sistema ABC de transporte de L-arginina completamente funcional (*artM*, *artQ*, *artI*, *artP*) (Wissenbach et al., 1995), a pesar de que la pseudogenización ha afectado al sistema ABC mayoritario de transporte de Lisina, Arginina y Ornitina (LAO) a nivel de los genes *hisM* y *hisQ* codificantes de los componentes integrales de membrana. La inactivación de este sistema de transporte también afecta a otras rutas de biosíntesis como la ruta de biosíntesis de putrescina, una poliamina implicada en desarrollo celular y en la función ribosomal conjuntamente con su derivado espermidina. En concordancia con la incapacidad de sintetizar ni incorporar del entorno L-ornitina debido a la inactivación de la ruta de biosíntesis de L-arginina y del sistema de transporte ABC-LAO, se ha inactivado también el gen *speE* codificante de una ornitina descarboxilasa responsable de la biosíntesis de putrescina a partir de la descarboxilación de L-ornitina, reteniendo genes funcionales *speA* (arginina decarboxilasa) y *speB* (agmatinasa) capaces de sintetizar putrescina a partir de L-arginina.

El gen *argD* se ha postulado que presenta actividad succinildiaminopimelato aminotransferasa adicional que sería necesaria para la biosíntesis de mesodiaminopimelato, un intermediario de la ruta de biosíntesis de L-lisina que también es necesario para la biosíntesis de peptidoglicano (Ledwidge and Blanchard, 1999). Sin embargo, esta evidencia procede de ensayos in-vitro de actividad de *ArgD* con diferentes sustratos, y muestran que la actividad con acetilornitina es mayor que con succinildiaminopimelato. Además, otros estudios postulan la existencia de una gen con actividad succinildiaminopimelato aminotransferasa adicional a *argD* en base a la presencia de aminotransferasas de función desconocida en *E. coli* como b2290, que presenta un ortólogo funcional en *S. glossinidius* (SG1602) no asignado a ninguna ruta, de tal forma que podría estar implicado en la biosíntesis de mesodiaminopimelato (Cox and Wang, 2001). Alternativamente también se ha descrito que el mesodiaminopimelato puede ser reemplazado por el intermediario de la cisteína cistationina en *E. coli* (Mengin-Lecreulx et al., 1994), que también se podría producir en *S. glossinidius* ya que presenta una ruta de biosíntesis de cisteína completamente funcional.

Las capacidades metabólicas de *S. glossinidius* y *W. glossinidia* también se han analizado en su conjunto para comparar las capacidades metabólicas de cada endosimbionte en el contexto de su asociación con la mosca tsé-tsé, específicamente a nivel de las rutas de biosíntesis de cofactores que se ha postulado como el principal motivo de esta asociación endosimbiótica (Rio et al., 2004; Aksoy and Rio, 2005). El resultado de esta comparativa revela que todas las capacidades funcionales de *W.*

glossinidia se encuentran funcionales en *S. glossinidius* con la excepción de la ruta de biosíntesis del cofactor tiamina. En *S. glossinidius*, esta ruta se encuentra severamente inactivada, lo que sugiere en un principio que no solo la mosca tsé-tsé sino también *S. glossinidius* dependen de *W. glossinidia* para incorporar tiamina. Sin embargo, el análisis detallado de esta ruta en *W. glossinidia*, aunque permite caracterizar genes previamente no-anotados en este genoma, revela que esta ruta se encuentra también incompleta en *W. glossinidia* debido a la ausencia del gen *thiI* y a la pseudogenización del gen *thiF*. La ruta de biosíntesis de tiamina. El perfil de esta ruta en ambos endosimbiontes de la mosca tsé-tsé se representa en la figura 4.19A. *S. glossinidius* ha inactivado los *thiS*, *thiF*, *thiG* y *thiE*, mientras que el gen *thiH* aparece como un gen descrito en la anotación original pero con un codón de parada prematuro al compararlo con sus homólogos funcionales, detectando el fragmento final como un pseudogen adyacente, probablemente reflejando un proceso de inactivación reciente. No obstante presenta genes *dxs*, *thiL*, *thiI* y *iscS* funcionales debido probablemente a la implicación de estos genes en otras rutas metabólicas funcionales en *S. glossinidius*. Por el contrario, *S. glossinidius* retiene un sistema de transporte de tiamina completamente funcional (*thiP, thiQ, tbpA*) conjuntamente con una tiamina kinasa (*thiK*) y tiamina fosfato kinasa (*thiL*) funcionales capaces de producir la forma activa del coenzima, tiamina difosfato, a partir de tiamina exógena. Sin embargo, *S. glossinidius* es capaz de sintetizar el precursor tiazol fosfato carboxilato a partir de una ruta secundaria (Ver Figura 4.19A). El mismo análisis sobre el genoma de *W. glossinidia* revela que aunque retiene la mayoría de genes de biosíntesis de tiamina, no presenta el gen *thiI* y los genes *thiS* y *thiF* aparecen no anotados pero presentes en la región intergénica entre *thiG* y *thiE* (ver Figura 4.19B). TBLASTN con los genes *thiS* y *thiF* de *E. coli* contra el genoma de *W. glossinidia* confirma la presencia de los dos genes, aunque mientras que el gen *thiS* aparece funcional, el gen *thiF* en *W. glossinidia* presenta un codón de parada interno que interrumpe la traducción del gen en el aminoácido 165, antes del residuo de cisteína en posición 184 de la proteína responsable de la interacción entre *ThiS-ThiF* esencial para la biosíntesis de tiazol fosfato (Xi et al., 2001). La ausencia de *thiF* y *thiI* limita la capacidad de *W. glossinidia* a sintetizar únicamente la subunidad hidroximetilpirimidina pirofosfato.

Debido a que ambos endosimbiontes parecen incapaces de sintetizar por si mismos la forma activa del coenzima tiamina difosfato, buscamos un posible fenómeno de complementación entre ambos endosimbiontes, que aparece al detectar en *S. glossinidius* la presencia de un gen funcional *ybjQ*, el cual se ha demostrado recientemente que codifica una actividad tiamina fosfato sintasa capaz de evitar la auxotrofia de tiamina en mutantes *thiE* (Morett et al., 2008). Esta actividad postula un escenario en el que *W. glossinidia* sería capaz de sintetizar la subunidad hidroximetilpirimidina pirofosfato a través de la ruta de-novo mientras que *S. glossinidius* sería capaz de sintetizar la subunidad hidroxitiazol fosfato a través de la ruta de reciclaje de tiamina. Ambos intermediarios serían compartidos entre ambos

endosimbiontes, de tal forma que *W. glossinidia* sería capaz de sintetizar tiamina difosfato a través de *ThiE* y *ThiL* mientras que *S. glossinidius* la sintetizaría a través de *YbjQ* y *ThiL* (Ver Figura 4.20). A favor de este escenario encontramos también que la proteína *YbjQ* de *S. glossinidius* presenta una alanina en la posición 89. En *E. coli* se ha demostrado que esta mutación incrementa la actividad tiamina difosfato de *YbjQ* (Morett et al., 2008).

Además de esta posible complementación a nivel de la biosíntesis de tiamina, encontramos también que las rutas de biosíntesis de los cofactores folato y coenzima A se encuentran severamente incompletas en W. glossinidius mientras que ambas están completamente funcionales en *S. glossinidius*. En el caso de la ruta de biosíntesis de folato, *W. glossinidia* es incapaz de sintetizar el precursor p-aminobenzoato a partir de corismato, el cual tampoco es capaz de sintetizarlo endógenamente, mientras que *S. glossinidius* presenta la ruta completa incluyendo la ruta de biosíntesis de corismato a partir del intermediario de la ruta de pentosas fosfato eritrosa 4 fosfato (Ver Figura 4.17). El corismato exógeno no puede ser incoporado exógenamente, y la única forma de sintetizarlo es a través de la ruta del siquimato a partir de eritrosa 4 fosfato, mientras que el p-aminobenzoato puede difundir libremente entre células y se ha demostrado que un suplemento exógeno de p-aminobenzoate es necesario para la viabilidad de células incapaces de sintetizarlo endógenamente (Green et al., 1992; Green et al., 1996), de tal forma que *S. glossinidius* sería capaz de actuar como una fuente de p-aminobenzoato para la biosíntesis de folato en *W. glossinidia*. De hecho, mutantes de *pabC* en *E. coli* necesitan un suplemento de p-aminobenzoato para sobrevivir (Green et al., 1992). En el caso de la ruta de biosíntesis de Coenzima A, *W. glossinidia* es incapaz de sintetizar 2-cetoisovalerato, el metabolito inicial de la ruta, en concordancia con la ausencia de genes implicados en la biosíntesis de valina, careciendo también del gen *panD* necesario para producir el metabolito intermediario β-alanina a partir de aspartato, mientras que *S. glossinidius* presenta la ruta completamente funcional (Ver Figura 4.18). Diversos estudios de inactivación génica revelan que la ausencia del gen *panD* en *E. coli* y Salmonella supone un requerimiento esencial de β-alanina o pantotenato exógeno para la supervivencia celular (Cronan, Jr. et al., 1982; Kennedy and Kealey, 2004). Asimismo, mutantes de *E. coli* en los genes de biosíntesis de cetoisovalerato presentan un requerimiento esencial de pantotenato y aminoácidos de cadena ramificada para su viabilidad (Whalen and Berg, 1982), de tal forma que *S. glossinidius* podría actuar como fuente de estos metabolitos para *W. glossinidia*. En este contexto, se ha comprobado que aproximadamente el 90% de pantotenato sintetizado por *E. coli* es secretado en lugar de ser utilizado para la biosíntesis de Coenzima A (Jackowski and Rock, 1981), de tal forma que *S. glossinidius* podría ser secretado por *S. glossinidius* y incorporado por *W. glossinidia* para sintetizar Coenzima A.

En este contexto, solo unos pocos casos de complementación metabólica entre diferentes endosimbiontes de un mismo hospedador se han descrito, siendo posible distinguir dos tipos de complementación diferentes. Por un lado, la complementación se puede establecer en el sentido de que cada endosimbionte es capaz de sintetizar diferentes productos finales del metabolismo, como en el caso de los dos endosimbiontes primarios de la chicharrita de alas cristalinas *Sulcia mulleri* y *Baumannia cicadellinicola*, donde la mayoría de aminoácidos son sintetizados por *S. mulleri* y suministrados tanto al hospedador como a *B. cicadellinicola*, con la excepción de metionina y cisteína, que son sintetizadas por este último y proporcionadas al hospedador y a *S. mulleri* (Wu et al., 2006; McCutcheon and Moran, 2007). Este sería el caso también del folato y coenzima A, los cuales pueden ser completamente sintetizados *de-novo* por *S. glossinidius* pero no por *W. glossinidia*, que necesita como mínimo el suplemento externo de diferentes precursores que no es capaz de producir por si misma. Por otro lado, la complementación también se puede producir en el sentido que ambos endosimbiontes sean necesarios para producir el metabolito final de una ruta de síntesis, que no puede ser sintetizado completamente por ninguno de los endosimbiontes de forma individual. Este es el caso de la ruta de biosíntesis de metionina en el sistema endosimbionte de la chicharrita de alas cristalinas descrito anteriormente, donde *S. mulleri* produce homoserina que es incorporada por *B. cicadellinicola* para producir metionina (Wu et al., 2006; McCutcheon and Moran, 2007), así como la ruta de biosíntesis de triptófano en los endosimbiontes del pulgón *Cinara cedri*, donde *B. aphidicola* es capaz de sintetizar el intermediario ácido antranilinico, que es incorporado por *Serratia symbiotica*, donde tiene lugar la síntesis final de triptófano que será suministrado al pulgón y a *B. aphidicola* (Gosalbes et al., 2008). Éste sería el caso de la posible complementación entre *S. glossinidius* y *W. glossinidia* en la ruta de biosíntesis de tiamina.

Además, la complementación a nivel de la ruta de biosíntesis de Coenzima A también se produce en los endosimbiontes de la chicharrita de alas cristalinas, donde B. cicadellinicola es capaz de sintetizar Coenzima A a partir de 2-cetoisovalerato pero no es capaz de producir este precursor por si mismo, el cual es producido por S. mulleri en el contexto de la biosíntesis de valina (Wu et al., 2006; McCutcheon and Moran, 2007). No obstante, el nivel de integración entre ambos endosimbiontes que se observa en este caso, donde tanto *S. mulleri* como *B. cicadellinicola* se encuentran muy próximas dentro del bacterioma, es mucho menor que en el caso de *S. glossinidius* y *W. glossinidia* en la mosca tsé-tsé, ya que la asociación de *S. glossinidius* es mucho más reciente y su tropismo a nivel de localización dentro de la mosca tsé-tsé es mucho mayor que en el caso de *W. glossinidia*, que se localiza en el interior de bacteriocitos, aunque recientemente se ha descrito la existencia de poblaciones extracelulares de *W. glossinidia* en el lumen del órgano "*milk-gland*" conjuntamente con *S. glossinidius*, desde donde ambos endosimbiontes son transmitidos a la descendencia (Attardo et al., 2008; Pais et al., 2008). Además, la

integración de *S. glossinidius* en la fisiología de la mosca tsé-tsé es mucho menor que en el caso de *W. glossinidia*, lo que queda reflejado en el mayor efecto deletéreo que la eliminación de *W. glossinidia* presenta sobre la eficacia biológica de la mosca tsé-tsé, aunque la eliminación de *S. glossinidius* también tiene efectos significativos sobre la longevidad del hospedador (Dale and Welburn, 2001). También es importante tener en cuenta que ambos genomas de *S. glossinidius* y *W. glossinidia* pertenecen a especies de hospedador diferentes, ya que el genoma de *S. glossinidius* pertenece a *Glossina morsitans morsitans* (Grupo Morsitans), mientras que *W. glossinidia* pertenece a *Glossina brevipalpis* (Grupo Fusca), lo que hace que estas inferencias metabólicas se tengan que tomar con precaución, especialmente dado la muy reciente asociación de *S. glossinidius* con el hospedador y la gran plasticidad de su genoma. En este contexto, sería de gran valor la secuenciación de los endosimbiontes complementarios (*W. glossinidia* de *G. morsitans morsitans* y *S. glossinidius* de *G. brevipalpis*) para tener una visión más real de la asociación de ambos endosimbiontes con la mosca tsé-tsé y determinar si los posibles fenómenos de complementación descritos se reproducen al comparar los endosimbiontes de una misma especie de hospedador.

CAPÍTULO 5: Reconstrucción y análisis funcional de la red metabólica de *Sodalis glossinidius*, el endosimbionte secundario de la mosca tsé-tsé**: Una aproximación a la evolución reductiva basada en la biología de sistemas**

En el capítulo anterior se ha llevado a cabo un análisis cualitativo de las capacidades funcionales de *S. glossinidius* en base a los resultados de la re-anotación funcional completa de su genoma, una aproximación que podríamos calificar como descriptiva y que permite inferir como se comportaría el organismo bajo diferentes condiciones de entorno y cual serían sus capacidades biosintéticas y degradativas. Sin embargo, la complejidad inherente a la mayoría de sistemas biológicos y la naturaleza multigénica de la mayoría de funciones celulares son conceptos que necesariamente tienen que ser considerados a la hora de extraer información fenotípica a nivel de respuestas celulares a partir de la información contenida en las secuencias genómicas. En el contexto del metabolismo celular, esto se puede aproximar mediante la reconstrucción de redes metabólicas que describen las capacidades funcionales potenciales de un determinado organismo y el análisis de sus propiedades sistémicas mediante el análisis de balance de flujos FBA o el análisis de rutas. Una red metabólica se puede definir como un conjunto de reacciones enzimáticas y de transporte que describen la circulación y producción de diferentes metabolitos en un determinado organismo (Schilling et al., 2000b). Este conjunto de reacciones se pueden inferir a partir de la anotación del genoma así como a partir de

información de estudios experimentales y de búsquedas en diferentes bases de datos (Durot et al., 2009). El objetivo último de la reconstrucción es obtener un modelo que permita inferir de la forma más fidedigna posible el comportamiento y las capacidades metabólicas del organismo que estudiamos, para lo cual sería necesario considerar la variable temporal a través de la incorporación de las cinéticas de reacción para las reacciones de la red. Sin embargo, en reconstrucciones a escala genómica se carece de información cinética para un gran número de reacciones de la red. Esta limitación se puede superar analizando el comportamiento del sistema en estado estacionario, donde únicamente se tiene en cuenta la estequiometria de las reacciones de la red, que a diferencia de las propiedades cinéticas, constituye una propiedad estructuralmente invariable de las redes metabólicas (Varma and Palsson, 1994a; Schilling et al., 2000a; Lee et al., 2006). La estequiometria se define como las relaciones molares de conversión de sustratos a productos en las reacciones de la red, y que a diferencia de las cinéticas de reacción no dependen ni de las condiciones externas ni del estado del sistema (Reder, 1988). La estequiometria de las reacciones de la red metabólica se representa matemáticamente en forma de una matriz estequiométrica S donde las filas (m) representan todos los metabolitos del sistema y las columnas (n) representan las diferentes reacciones de la red, de tal forma que cada índice de la matriz representa la estequiometria de un determinado metabolito en una determinada reacción. Un ejemplo sencillo de una red metabólica y su correspondiente matriz estequiométrica se representa en la Figura 5.1.

A partir de la estequiometria de las reacciones, se pueden formular ecuaciones dinámicas para cada metabolito que describan la variación de su concentración a lo largo del tiempo, y que vendrán definidas por la diferencia neta entre el conjunto de flujos metabólicos que consumen y producen cada metabolito (Figura 5.1). En ausencia de información cinética que permita tratar o considerar la variante temporal, el análisis estequiométrico de redes metabólicas asume que las redes se encuentran en estado estacionario dado que los flujos metabólicos internos de interconversión de metabolitos son generalmente más rápidos que las respuestas celulares a nivel de crecimiento o respuesta a cambios externos (Varma and Palsson, 1994a; Schilling et al., 2000a). Como consecuencia de este estado estacionario, los flujos metabólicos que producen y que consumen un determinado metabolito tienen que estar equilibrados, es decir, que el balance neto entre producción y consumo de cada metabolito de la red sea igual a cero (Figura 5.1). A nivel matemático, la asunción de estado estacionario supone que el producto de la matriz estequiométrica por el vector de flujos metabólicos sea igual a cero, transformando el sistema de ecuaciones diferenciales formuladas para cada metabolito en un sistema de ecuaciones lineales, de tal forma que conociendo la estequiometria de las reacciones podemos inferir el vector de flujos metabólicos del sistema que define sus propiedades funcionales (Covert et al., 2001). Sin embargo, este sistema de ecuaciones lineales es indeterminado debido a que el número de reacciones de la red sobrepasa al número de metabolitos, lo que supone que no se pueda inferir una única

solución o un único vector de flujos metabólicos debido a que múltiples distribuciones de flujos metabólicos satisfacen la condición del estado estacionario, generando un espacio de soluciones posibles (distribuciones de flujos metabólicos) que puede ser explorado matemáticamente mediante análisis convexo (Figura 5.1) (Rockafellar, 1970). El conjunto de soluciones del sistema en un espacio n-dimensional donde n representa el número de reacciones de la red genera geométricamente lo que se denomina un espacio convexo o "*convex set*" que contiene en su interior todas las distribuciones de flujos metabólicos que se ajustan a la asunción de estado estacionario, de tal forma que cada punto dentro de este espacio de soluciones representa una de estas posible distribuciones. Desde un punto de vista biológico, el espacio convexo contiene todas las capacidades funcionales o potenciales fenotipos que el sistema representado por la red metabólica puede exhibir en estado estacionario (Varma and Palsson, 1994a; Schilling et al., 2000a; Edwards et al., 2002; Famili and Palsson, 2003). El espacio convexo determina lo que el sistema es capaz de hacer bajo unas determinadas condiciones de entorno, que componentes de biomasa celular es capaz de producir, como de eficaz es el sistema bajo diferentes fuentes de carbono externas, o cuales son los posibles puntos críticos o genes esenciales del sistema en base al efecto que su eliminación tiene sobre la estructura de este espacio convexo (Schilling and Palsson, 2000; Reed et al., 2003; Forster et al., 2003b; Blank et al., 2005; Kuepfer et al., 2005; Borodina et al., 2005).

El espacio convexo se puede explorar desde dos aproximaciones matemáticas diferentes. Por un lado se puede explorar el conjunto de rutas metabólicas que pueden operar en el sistema mediante la caracterización computacional del conjunto de Modos de Flujo Elemental (EFM) y de Rutas Extremas (Schuster et al., 1999; Papin et al., 2004; Planes and Beasley, 2008). Por otro lado, se puede determinar de todas las posibles distribuciones de flujos metabólicos en estado estacionario cual es la que optimiza una determinada función de objetivo de interés biológico mediante Análisis de Balance de Flujos o FBA (Schilling et al., 2000a; Schilling et al., 2000b).

Análisis estequiométrico de rutas metabólicas

Las rutas metabólicas se pueden definir como conjuntos de enzimas que actúan de forma coordinada para la producción o degradación de determinados metabolitos o moléculas. Estas rutas se pueden caracterizar de forma cualitativa mediante el mapeo de funciones génicas en rutas metabólicas descritas en la literatura y representadas en diferentes bases de datos (Overbeek et al., 2000; Kanehisa et al., 2004; Keseler et al., 2009; Caspi et al., 2009) o a través de la identificación computacional de módulos en redes metabólicas a partir del espacio convexo de soluciones de la red en estado estacionario (Schilling et al., 2000a; Schilling et al., 2000b; Schuster et al., 2000; Papin et al., 2004). En este contexto computacional, las rutas metabólicas se definen como conjuntos de reacciones conectadas o acopladas

entre si en la red metabólica, y su identificación se ha llevado a cabo a través de dos aproximaciones matemáticas similares, los Modos Elementales de Flujo (EFM) y las Rutas Extremas (EP). Los EFM se definen como el conjunto de módulos no descomponibles que operan en estado pseudo-estacionario a partir de las restricciones estequiométricas de la red (Schuster et al., 1999; Pfeiffer et al., 1999). Las EP representan un subconjunto dentro de los EFM que cumplen el requisito adicional de la independencia sistémica, por la cual ninguna EP se puede representar a partir de combinaciones lineales de otras EP (Schilling et al., 2000b; Planes and Beasley, 2008). El conjunto de EP´s de un determinado sistema se define como sus bases convexas, es único de ese sistema, y representa el mínimo conjunto de rutas o módulos a partir de los cuales se describen todas las posibles distribuciones de flujos del sistema (Schilling et al., 2000a; Klamt and Stelling, 2003). Desde el punto de vista de la geometría del espacio convexo, las EP´s se corresponden con los bordes del espacio convexo, y cualquier punto del espacio convexo, equivalente a una de las posibles distribuciones de flujos metabólicos en estado estacionario, se puede representar a través de una combinación lineal de EP´s (Schilling et al., 2000a).

Desde un punto de vista funcional, el número de módulos o rutas alternativas para una misma función es un indicativo de la robustez estructural del sistema, que se puede evaluar a partir del efecto que mutaciones en genes de la red tienen sobre el número de EFM y EP (Stelling et al., 2002). Además, el conjunto de EFM´s permite identificar subconjuntos enzimáticos en el sentido de reacciones que siempre operan de forma coordinada bajo diferentes condiciones de entorno y que serían susceptibles de compartir mecanismos reguladores comunes, así como reacciones excluyentes que nunca operan conjuntamente en ningún EFM (Klamt and Stelling, 2003). Sin embargo, estas aproximaciones basadas en EFM y EP presentan severos problemas computacionales al tratar con redes metabólicas complejas como son las redes a escala genómica, ya que se produce un incremento exponencial en el número de EFM´s y EP´s a medida que se incrementa el tamaño de la red, un fenómeno que se denomina explosión combinacional y que hace que la inferencia de EFM´s y EP´s sea impracticable para redes complejas (Planes and Beasley, 2008). Como ejemplo, en una red metabólica de 110 reacciones que comprende el metabolismo central de *E. coli* presenta un total de 27099 EFM´s con glucosa como fuente de carbono externa (Stelling et al., 2002).

Análisis de Balance de Flujos (FBA)

El espacio convexo de distribuciones de flujos metabólicos a través de reacciones de la red en estado estacionario se puede restringir con la imposición de restricciones adicionales al sistema como restricciones termodinámicas definidas en base a la reversibilidad-irreversibilidad de las reacciones, restricciones topológicas a nivel de la restricción de metabolitos y reacciones a determinados compartimentos celulares, o restricciones ambientales en base a la definición de metabolitos de entrada al

sistema (Price et al., 2003; Price et al., 2004; Becker et al., 2007). Del mismo modo, se pueden determinar experimentalmente los flujos o rangos de flujo metabólico a través de reacciones de la red mediante Análisis Metabólico de Flujos (MFA), que permite obtener distribuciones de flujos metabólicos más próximas a las distribuciones reales (Christensen and Nielsen, 2000a; Christensen and Nielsen, 2000b; Fischer and Sauer, 2005). En este contexto, FBA es una aproximación matemática que utiliza técnicas de optimización linear para determinar la distribución de flujos metabólicos dentro del espacio convexo definido por el conjunto de restricciones aplicadas al sistema que optimizan una determinada función objetivo de interés biológico, lo que permite evaluar y explorar los fenotipos metabólicos posibles de forma cuantitativa (Ver Figura 5.1) (Schilling et al., 2000a; Edwards et al., 2002; Durot et al., 2009). Por tanto, el punto crítico en FBA es la definición de una función objetivo que defina un determinado comportamiento celular de manera realista. Para ello, la aproximación predominante consiste en formular una ecuación de biomasa capaz de describir cuantitativamente el crecimiento celular del sistema en estudio, lo cual es uno de los objetivos prioritarios del FBA sobre redes metabólicas genómicas (Price et al., 2003; Price et al., 2004). Para ello, se incorpora una ecuación adicional a la red metabólica que representa los requerimientos de energía y componentes celulares necesarios para la supervivencia del sistema y que se define como ecuación de biomasa, la cual constituye la función objetivo a maximizar por optimización lineal. Esto hace que la composición de la ecuación de biomasa sea determinante a la hora de obtener predicciones realistas de crecimiento celular por FBA. La ecuación de biomasa se define como la proporción de metabolitos necesarios para producir un gramo de peso seco en cultivos celulares (Joyce and Palsson, 2008). Para ello se puede recurrir a medidas experimentales propias o obtenidas de la bibliografía en las cuales se mide la proporción de metabolitos internos a partir de cultivos celulares (Reed et al., 2003; Forster et al., 2003a). Alternativamente, si no es posible obtener esta información de manera específica para el organismo en estudio, se puede adoptar la ecuación de biomasa de otro organismo próximo (Durot et al., 2009; Zhang et al., 2009).

Las predicciones de crecimiento celular por FBA sobre redes metabólicas utilizando biomasa como función objetivo se han demostrado coincidentes con medidas experimentales de crecimiento celular en diferentes organismos (Edwards and Palsson, 2000d; Schilling et al., 2002; Covert and Palsson, 2002). Además, FBA se ha utilizado para estudiar la fisiología de producción de ATP durante la síntesis de ácidos grasos en adipocitos (Fell and Small, 1986), para predecir la secreción de etanol en levaduras (Sonnleitner and Kappeli, 1986), o para predecir la secreción de acetato en *E. coli*, donde se ha demostrado que las predicciones de crecimiento celular, secreción de acetato y tasas de consumo de glucosa utilizando biomasa como función objetivo coinciden con medidas experimentales en cultivos celulares (Varma et al., 1993b; Varma and Palsson, 1994b). Sin embargo, además de la predicción de crecimiento celular, otra de las aplicaciones predominantes del análisis

por FBA sobre redes metabólicas es predecir el efecto que deleciones génicas tienen sobre la capacidad de crecimiento celular mediante la simulación de sucesos de deleción génica restringiendo a cero el flujo metabólico a través de las reacciones catalizadas por estos genes. Esto es posible gracias a la adición de lo que se llama un nivel de asociación Gen-Proteína-Reacción (*GPR layer*) a la red metabólica (Reed et al., 2003). Este tipo de análisis sobre las redes metabólicas de *E. coli*, *Saccharomyces cerevisiae* o *Pseudomonas putida* reveló que las predicciones de esencialidad en base a simulaciones sobre la red son altamente coincidentes con los resultados de *knockouts* experimentales (Edwards and Palsson, 2000d; Duarte et al., 2004; Oberhardt et al., 2008). Esta aproximación se ha aplicado con éxito a para caracterizar dianas metabólicas potenciales en *Mycobacterium tuberculosis* sobre un modelo metabólico de producción de ácidos micólicos (Raman et al., 2005), así como para identificar un core de reacciones esenciales compartidas por *E. coli*, *S. cerevisiae* y *H. pylori* que pudieran corresponderse con potenciales dianas de antibióticos que interfirieran con el metabolismo celular (Almaas et al., 2005). Además de FBA, se han desarrollado dos aproximaciones alternativas para evaluar el efecto de inactivaciones génicas sobre la red metabólica que se basan en el concepto de minimización del ajuste metabólico, donde se asume que el sistema después de la inactivación génica trata de minimizar el cambio metabólico respecto del estado metabólico inicial en lugar de maximizar la función de biomasa, que es la asunción que sigue FBA (Segre et al., 2002; Shlomi et al., 2005). Por un lado, el MOMA (minimización del ajuste metabólico) identifica la distribución de flujos metabólicos sobre la red inactivada que minimice la distancia euclídea de la distribución de flujos óptima sobre la red original (Segre et al., 2002), mientras que la aproximación conocida como ROOM (*Regulatory ON/OFF minimization*) minimiza el número de cambios de flujo significativos respecto a la red original (Shlomi et al., 2005). Estas aproximaciones mejoran ligeramente las predicciones de crecimiento celular en momentos inmediatamente posteriores a la inactivación génica, pero experimentos de evolución experimental sobre *E. coli* midiendo crecimiento celular en medio mínimo con glicerol demostraron que aunque en etapas iniciales de la inactivación el crecimiento era subóptimo, el sistema evolucionaba hacia la maximización de crecimiento celular de acuerdo con las predicciones de FBA maximizando producción de biomasa (Ibarra et al., 2002).

Desde el punto de vista evolutivo, simulaciones de evolución reductiva con FBA sobre la red metabólica de *E. coli* en condiciones de entorno que simulan el interior de pulgones y de la mosca tsé-tsé predicen el contenido génico de *B. aphidicola* y *W. glossinidia* respectivamente con elevada precisión (Pal et al., 2006b). Del mismo modo, la combinación de FBA sobre la red de *E. coli* con análisis de genómica comparada y evolución de contenido génico reveló el papel predominante de la transferencia genética horizontal sobre la duplicación génica en la evolución de la red metabólica de *E. coli* así como la integración preferencial de genes transferidos horizontalmente en nodos periféricos de la red con actividad específica bajo

determinadas condiciones externas (Pal et al., 2005a; Pal et al., 2005b). En este capítulo vamos a llevar a cabo la reconstrucción y análisis por FBA de la red metabólica de *S. glossinidius* a diferentes etapas del proceso de evolución reductiva, lo que nos permitirá tener una visión global de como el proceso de inactivación génica esta afectando al crecimiento y viabilidad de *S. glossinidius*, y cual podría ser su posible evolución futura en el contexto de la reducción genómica mediante la identificación de genes esenciales y delecionables en base a su esencialidad para el funcionamiento del sistema, evaluando también si estos genes presentan algún patrón de evolución diferencial a nivel de secuencia.

Resultados y discusión

La red metabólica de *S. glossinidius* a diferentes etapas del proceso de evolución reductiva

Es este capítulo, la red metabólica ancestral y funcional de *S. glossinidius* ha sido reconstruida en base a la re-anotación funcional de genes y pseudogenes llevada a cabo en el capítulo anterior. Para ello la red metabólica de *E. coli* K12 JR904 (Reed et al., 2003) es tomada como red de referencia de una bacteria entérica de vida libre, lo cual nos facilita el proceso de reconstrucción así como la interpretación de los resultados de FBA con los de una bacteria de referencia como *E. coli*. El proceso de reconstrucción de las redes de *S. glossinidius* se resume en: (1) Un mapeo inicial de genes ortólogos entre el genoma ancestral de *S. glossinidius* (genes y pseudogenes) y el genoma de la red de *E. coli* K12 JR904, lo cual nos proporciona un primer borrador de la red ancestral de *S. glossinidius*, (2) Identificación de sucesos de deleción en el genoma de *S. glossinidius* desde su divergencia del ancestro de vida libre mediante comparaciones de sinténia con *E. coli* K12 y mapeo de genes ortólogos en genomas evolutivamente próximos, (3) Fase de validación y refinamiento del modelo donde se incorporan genes y pseudogenes de *S. glossinidius* relacionados con funciones metabólicas sin relación de ortología con *E. coli* K12 y se valida la funcionalidad de las redes metabólicas obtenidas mediante FBA comparando con la red metabólica de *E. coli* K12 JR904 utilizando como función objetivo la ecuación de biomasa de *E. coli* K12 JR904. La composición de la ecuación de biomasa se representa en la Tabla 5.1.

La red metabólica ancestral de *S. glossinidius* se encuentra compuesta por 668 productos génicos, 741 reacciones internas incluyendo reacciones metabólicas y de transporte celular así como la ecuación de biomasa, y 690 metabolitos de los cuales 590 son citoplásmicos y 143 son extracelulares (Ver Tabla 5.3). De los 668 productos génicos, 479 se corresponden con genes, 148 se corresponden con pseudogenes, y 41 se corresponden con genes delecionados durante la evolución de *S. glossinidius* desde su ancestro de vida libre. Los 627 CDSs representados por genes y pseudogenes constituyen un 16% del total de CDSs del genoma de *S.*

glossinidius, lo que se encuentra en el rango de otras redes metabólicas disponibles (Ver Tabla 5.3). De las 741 reacciones internas, 683 (78.27%) están asociadas a al menos un gen, pseudogen, o gen delecionado, mientras que 58 reacciones que en la red de *E. coli* JR904 aparecen sin gen asociado fueron incorporadas a la red ancestral y funcional de *S. glossinidius*. La red ancestral incluye 27 pseudogenes que no presentan similitud a nivel de secuencia con *E. coli* K12 pero que están relacionados con funciones metabólicas concretas con estequiometria bien definida en base a los resultados de la re-anotación. Estos 27 pseudogenes y sus correspondientes reacciones se encuentran representados en la Tabla suplementaria 5.3, y representan funciones metabólicas ancestrales de *S. glossinidius* como la capacidad de crecer con formato como fuente de carbono a través de la presencia de dos pseudogenes correspondientes a formato deshidrogenasas(ps_SGL0061c y ps_SGL0873), la capacidad de crecer con glicerol como fuente de carbono a través de la presencia de un pseudogen (ps_SGL1175) codificante de una glicerol fosfotranferasa ausente en *E. coli*, o la capacidad de producir metionina a partir de acetil coenzima A y homoserina a través de la presencia de un pseudogen (ps_SGL1185c) codificante de una homoserina-O-acetiltransferasa equivalente desde el punto de vista funcional a la homoserina-O-succiniltransferasa funcional codificada por el gen *metA* que utiliza succinil coenzima A como dador del grupo succinato (Mirza et al., 2005).

En la transición a la red funcional de *S. glossinidius*, todas las reacciones catalizadas por pseudogenes y genes delecionados fueron eliminadas de la red ancestral, dando lugar a una red funcional compuesta por 458 productos génicos, 560 reacciones internas incluyendo reacciones metabólicas y de transporte celular así como la ecuación de biomasa, y 624 metabolitos de los cuales 481 son citoplásmicos y 143 son extracelulares (Ver Tabla 5.3). De las 560 reacciones internas, 502 se encuentran asociadas al menos con un gen, conservándose las 58 reacciones sin gen asociado de la red de *E. coli* JR904. Los 458 genes de la red funcional representan un 18.84% del total de genes de *S. glossinidius*, en el rango de otras reconstrucciones de redes metabólicas a escala genómica (ver Tabla 5.3). En la transición a la red metabólica funcional, 21 genes funcionales fueron eliminados de la red debido a que codifican subunidades de complejos enzimáticos no funcionales donde alguna de las subunidades se encuentra inactivada o delecionada. Además, 13 genes funcionales sin relación de ortología con *E. coli* por similitud a nivel de secuencia fueron incorporados tanto a la red ancestral como a la red funcional. Estos genes y sus reacciones correspondientes se representan en la Tabla suplementaria 5.4. La mayoría de reacciones asociadas con estos genes se encuentran presentes en la red de *E. coli* K12 JR904 aunque no se conserve similitud a nivel de secuencia entre los correspondientes genes en ambos genomas, como es el caso de la presencia en *S. glossinidius* de una Isocitrato deshidrogenasa monomérica codificada por el gen SG0700 en contraste con la presencia de isocitato deshidrogenasas homodiméricas en la mayoría de enterobacterias incluyendo *E. coli* K12 completamente diferentes a

nivel de secuencia (Eikmanns et al., 1995; Sahara et al., 2002), o la presencia de genes *rfa* codificantes de glicosil hidrolasas diferentes en *S. glossinidius* y *E. coli* K12 responsables de la adición de diferentes subunidades de azúcares al core de oligosacáridos del lipopolisacárido bacteriano, conservándose el orden génico entre ambos genomas, lo que indica una composición diferencial de esta estructura de cubierta celular, algo que también se ha observado entre linajes de *E. coli* y Salmonella (Heinrichs et al., 1998).

Análisis de las redes metabólicas de *S. glossinidius* por FBA: Transición al modo de vida dependiente de hospedador

Con el objetivo de comparar las capacidades funcionales de las redes de *S. glossinidius* ancestral y funcional con la red de *E. coli* JR904 se llevaron a cabo FBA sobre las tres redes metabólicas utilizando como función objetivo la ecuación de biomasa formulada para la red metabólica de *E. coli* JR904 (Reed et al., 2003) y descrita en la Tabla 5.1, debido a la ausencia de información acerca de la composición celular interna de *S. glossinidius*. Esta misma ecuación de biomasa de *E. coli* JR904 se ha utilizado en muchas otras reconstrucciones metabólicas y se considera una buena aproximación en aquellos casos en los que no es posible formular una ecuación de biomasa específica (Puchalka et al., 2008; Oberhardt et al., 2008; Zhang et al., 2009). De hecho, comparaciones de las tasas de producción de biomasa predichas sobre la red metabólica de *Pseudomonas putida* utilizando la ecuación de biomasa de *E. coli* K12 JR904 son altamente coincidentes con medidas experimentales de crecimiento celular en cultivo de *P. putida*, donde también se ha comprobado que variaciones en las proporciones de los diferentes metabolitos incluidos en la ecuación de biomasa tiene un efecto mínimo sobre el resultado de las simulaciones (Puchalka et al., 2008). En el caso de *S. glossinidius*, la proximidad evolutiva con *E. coli* y la ausencia de información detallada acerca de su composición interna justifica la utilización de la ecuación de biomasa de *E. coli*, lo cual también nos permite comparar los resultados del FBA sobre las redes de *S. glossinidius* y *E. coli* en base a una ecuación de biomasa que describe el crecimiento celular de una bacteria de vida libre típica. Los resultados del FBA sobre las tres redes metabólicas maximizando la producción de biomasa según la ecuación de *E. coli* K12 JR904 se representan en la Figura 5.7. Inicialmente llevamos a cabo las simulaciones en unas condiciones de entorno aerobias mínimas con únicamente glucosa como fuente de carbono, a partir de la cual el sistema tiene que ser capaz de sintetizar todos los componentes de la ecuación de biomasa. Para simular estas condiciones de entorno se incluyen en la red metabólica 143 reacciones de intercambio o pseudoreacciones que nos permiten definir las condiciones de entorno del sistema o, dicho de otro modo, los metabolitos que el sistema tiene a su disposición para producir la biomasa (Schilling et al., 2000b). Alterando los límites superiores e inferiores de estas pseudoreacciones definimos qué metabolitos pueden entrar y salir del sistema, de tal forma que si el flujo a través de las reacciones de

intercambio adquiere un valor negativo, el metabolito puede entrar en el sistema y ser utilizado para producir biomasa mientras que si el flujo adquiere un valor positivo el metabolito sale del sistema y no puede ser utilizado para producir biomasa. En la Figura 5.6 se representan las condiciones de entorno mínimas iniciales con solo glucosa como fuente de carbono en base a los límites superiores e inferiores de las correspondientes reacciones de intercambio o pseudoreacciones. Bajo estas condiciones, la red ancestral de *S. glossinidius* es completamente funcional, con una tasa de producción de biomasa de 0.545 gr. Peso Seco (mmol Glucosa)$^{-1}$ muy similar a la que se observa para la red de *E. coli* K12 JR904 (0.5391 gr. Peso Seco (mmol Glucosa)$^{-1}$), lo que indica que la red ancestral de *S. glossinidius* es completamente funcional al mismo nivel que una bacteria de vida libre en un medio mínimo con únicamente glucosa como fuente de carbono a partir de la cual sintetizar todos los componentes de biomasa. Esto supone una evidencia adicional a favor de la muy reciente transición de *S. glossinidius* al modo de vida dependiente de hospedador que se añade a la ausencia de coevolución entre las filogenias de *S. glossinidius* y sus correspondientes hospedadores (Chen et al., 1999), su capacidad única entre los endosimbiontes bacterianos de insectos de ser cultivable in-vitro (Welburn et al., 1987), y sus características genómicas más próximas a bacterias de vida libre que a endosimbiontes obligados (Toh et al., 2006).

Por el contrario, la red metabólica funcional de *S. glossinidius* considerando únicamente los genes no es viable en estas condiciones de entorno mínimas, produciendo un fenotipo que podríamos considerar como letal en términos de producción de biomasa (0 gr. Peso Seco (mmol Glucosa)$^{-1}$). Esto se debe a tres sucesos de pseudogenización principales. Por un lado la inactivación de los genes *glgA* y *glgB* responsables de la biosíntesis de glicógeno (ps_SGL1448 y ps_SGL1450 respectivamente) y la inactivación de los genes *argA*, *argG*, *argD*, y *argC* (ps_SGL1220, ps_SGL0928, ps_SGL1434, y ps_SGL1382c respectivamente) implicados en la síntesis de L-arginina, ya que ambos metabolitos están incluidos en la ecuación de biomasa, con lo cual estos sucesos de inactivación producen un fenotipo letal al hacer FBA sobre la red funcional de *S. glossinidius*. Es necesario eliminar el glicógeno de la ecuación de biomasa y añadir una fuente externa de arginina conjuntamente con la glucosa para obtener un fenotipo viable en términos de producción de biomasa en la red funcional de *S. glossinidius* (Ver Figura 5.6). L-arginina es necesaria no solo como componente de biomasa sino también como precursor para la síntesis de las poliaminas putrescina y espermidina, ambas también componentes de la ecuación de biomasa y que tienen que ser producidas por el sistema para tener un fenotipo viable en términos de producción de biomasa. Por otro lado, además de estos dos sucesos de inactivación (genes de biosíntesis de glicógeno y genes de biosíntesis de arginina), la inactivación del gen *ppc* codificante del enzima anaplerótico fosfoenolpiruvato carboxilasa tiene consecuencias particularmente deletéreas ya que este es una actividad enzimática esencial para suplementar oxalacetato al ciclo del ácido cítrico a partir del cual se sintetizan un

gran número de metabolitos esenciales (March et al., 2002) y cuya eliminación incluso sobre la red ancestral de *S. glossinidius* tiene consecuencias letales en términos de producción de biomasa (ver Figura 5.8) debido a que en *S. glossinidius* representa la única actividad anaplerótica capaz de incrementar el flujo metabólico a través del ciclo del ácido cítrico ya que no presenta un ciclo del glioxilato funcional como sí tiene *E. coli* K12, donde como se demuestra en los resultados de la Figura 5.8, la eliminación del gen *ppc* activa dicha ruta, unos resultados que son coincidentes con observaciones experimentales (Peng et al., 2004; Peng and Shimizu, 2004). El fenotipo funcional que observamos en la red funcional con L-arginina como metabolito externo se debe a que L-arginina puede ser incorporado como componente de biomasa y como precursor de las poliaminas putrescina y espermidina y a que la putrescina es degradada a el intermediario del ciclo del ácido cítrico succinato, de tal forma que se puede complementar la inactivación de la fosfoenolpiruvato carboxilasa anaplerótica (ver Figura 5.8), lo que indica que unos pocos sucesos de inactivación génica producen cambios drásticos en las capacidades funcionales de *S. glossinidius* que explican la transición desde un ancestro completamente funcional en un medio mínimo a un estado actual donde se necesita de forma esencial un suplemento externo de arginina para sobrevivir y que se asociaría al modo de vida dependiente de hospedador.

FBA también se utilizó para evaluar la funcionalidad de las redes metabólicas de *S. glossinidius* y *E. coli* K12 JR904 bajo diferentes fuentes de carbono en condiciones aerobias, con el objetivo de comparar los resultados de estas simulaciones con datos experimentales de cultivos celulares de *S. glossinidius* (Dale and Maudlin, 1999). Los resultados de este análisis se representan en la Figura 5.9, donde se representa la diferencia entre la producción de biomasa con arginina y la fuente de carbono y la producción de biomasa únicamente con arginina como metabolito externo, ya que con arginina únicamente el sistema es capaz de producir biomasa aunque a niveles que son aproximadamente la mitad que con glucosa como fuente de carbono debido a la utilización de arginina para producir putrescina que será degradada a succinato descrito anteriormente como consecuencia de la inactivación del gen *ppc*. Los resultados de este análisis muestran que de las 27 fuentes de carbono diferentes evaluadas, en 19 de ellas (70.37%) se muestran resultados coincidentes entre las simulaciones por FBA y los resultados experimentales descritos por Dale y colaboradores (Dale and Maudlin, 1999) incluyendo la capacidad de crecimiento con glucosa, N-acetilglucosamina y manitol, con 8 falsos positivos que representan fuentes de carbono bajo las cuales FBA predice un fenotipo viable en términos de producción de biomasa mientras que los datos experimentales no reportan crecimiento significativo y ningún falso negativo. Aunque dos de los falsos positivos (etanol y piruvato) son consecuencia de reacciones de difusión no asociadas a gen en la red de *E. coli* K12 JR904 que han sido asumidas en las redes de *S. glossinidius*, en otros casos el fenotipo observado experimentalmente no se ajusta con el comportamiento esperado dada la información del genoma, la cual se reproduce en

las simulaciones de FBA. Este es el caso de la metabolización de glucosa pero no manosa ni fructosa descrita experimentalmente por Dale y colaboradores (Dale and Maudlin, 1999) a pesar de que, como describimos en el capítulo anterior, el genoma de *S. glossinidius* presenta un sistema PTS completamente funcional para la manosa pero ha inactivado los sistemas PTS y transportadores adicionales de glucosa y fructosa. De acuerdo con la amplia especificidad de sustrato del sistema PTS de la manosa, que es capaz de transportar con similar eficacia manosa, fructosa, glucosa (Curtis and Epstein, 1975; Postma et al., 1993; Kornberg, 2001), esta amplia especificidad de sustrato es incorporada a la red metabólica ancestral y funcional de *S. glossinidius* asociando el transporte de manosa, glucosa y fructosa al transportador PTS de la manosa, dando lugar al crecimiento positivo esperado al hacer FBA con la red ancestral y funcional de *S. glossinidius*. A este respecto, existen evidencias de cambios mutacionales en componentes de los sistemas PTS capaces de cambiar la especificidad de sustrato bajo determinadas condiciones de entorno (Oh et al., 1999; Notley-McRobb and Ferenci, 2000), lo cual explicaría un posible cambio de especificidad del sistema PTS de manosa de *S. glossinidius* para asimilar glucosa con mayor eficacia.

Análisis de robustez en redes metabólicas de *S. glossinidius* y *E. coli* K12 JR904

La robustez genética de las redes metabólicas ancestral y funcional de *S. glossinidius* y de la red metabólica de *E. coli* K12 JR904 se ha evaluado desde dos perspectivas diferentes. Por un lado se ha evaluado la robustez de las redes frente a sucesos de deleción de genes individuales, para lo cual en cada suceso de deleción se restringe a cero el flujo a través de las reacciones asociadas a cada gen y la producción de biomasa de esta red delecionada se evalúa por FBA. Por otro lado se ha estudiado como varía la producción de biomasa en respuesta a variaciones en los flujos metabólicos a través de determinadas reacciones de la red. Los resultados del análisis de robustez frente a sucesos de deleción se representan en la Figura 5.10, y muestran que el comportamiento de la red ancestral de *S. glossinidius* y la red de *E. coli* K12 JR904 son similares a nivel del número de genes esenciales cuya deleción da un fenotipo letal en términos de producción de biomasa (160 genes en *E. coli* K12 JR904 y 166 genes en la red ancestral de *S. glossinidius*), difiriendo en el número de genes cuya deleción no afecta a la funcionalidad del sistema (675 genes en *E. coli* K12 JR904 y 396 genes en la red ancestral de *S. glossinidius*). Las diferencias aparecen al analizar la red funcional de *S. glossinidius*, donde 204 genes aparecen como esenciales en base a las simulaciones de deleción. Esto supone una disminución significativa de la robustez de las redes metabólicas como consecuencia del proceso de pseudogenización en comparación con sus ancestros de vida libre cuyo extremo lo encontramos al analizar los resultados de estas mismas simulaciones sobre la red metabólica de *B. aphidicola* del pulgón *A. pisum*, la única red metabólica a escala genómica disponible para una bacteria endosimbionte (ver Figura 5.11) (Thomas et al., 2009), y que contrasta con la comprobada robustez

topológica de las diferentes redes biológicas incluyendo las metabólicas consecuencia de su organización estructural libre de escala (Jeong et al., 2000; Albert et al., 2000; Podani et al., 2001; Barabasi and Oltvai, 2004). Resultados similares se han observado al analizar la robustez topológica y funcional de un red metabólica mínima teórica (Gabaldon et al., 2007). Por otro lado, los análisis de robustez frente a cambios en los valores de flujo metabólico a través de diferentes reacciones de la red también muestran una disminución significativa en el rango de flujos capaces de producir un fenotipo viable en términos de producción de biomasa (Ver Figura 5.12), aunque estos análisis se han llevado a cabo únicamente con las reacciones de la glicólisis, que por otra parte es una ruta metabólica central de la cual dependen gran parte de las capacidades biosintéticas celulares.

Estos resultados muestran que el proceso de inactivación génica asociado a la reducción genómica de bacterias endosimbiontes produce una marcada disminución de la robustez de los sistemas metabólicos, no solo a sucesos de deleción sino también a variaciones en las actividades enzimáticas en comparación con bacterias de vida libre, mostrando un perfil metabólico más frágil que explicaría la transición a un modo de vida dependiente de hospedador, que constituye un ambiente más estable y con menos fluctuaciones que el ambiente típico de una bacteria de vida libre como *E. coli* (Moran and Plague, 2004; Silva et al., 2007). Este efecto es más pronunciado en etapas más avanzadas del proceso de reducción genómica, ejemplificadas en la red metabólica de *B. aphidicola*, probablemente consecuencia de la pérdida de genes implicados en reparación de DNA (Sharples, 2009) y que explicaría el mayor grado de integración de estos endosimbiontes primarios en el interior de bacteriocitos del hospedador en comparación con la mayor variabilidad a nivel de localización en diferentes tejidos del hospedador, tanto intracelularmente como extracelularmente, que se observa en bacterias endosimbiontes como *S. glossinidius* en etapas iniciales de la transición al modo de vida dependiente de hospedador.

Simulaciones de evolución reductiva sobre la red funcional de *S. glossinidius*

Con el objetivo de tratar de predecir cual podría ser la evolución futura de *S. glossinidius* en el contexto del proceso de reducción genómica se llevaron a cabo simulaciones de evolución reductiva sobre la red metabólica funcional de *S. glossinidius* que consisten en la eliminación secuenciar de todas las reacciones de la red evaluando a cada paso la funcionalidad de la red resultante mediante FBA, de tal forma que si la tasa de producción de biomasa en la red delecionada se encuentra por encima de un determinado valor de corte, consideramos la reacción como no esencial y la eliminamos conjuntamente con sus correspondientes genes, mientras que si está por debajo del valor de corte consideramos la reacción como esencial, de tal forma que lo retenemos en la red conjuntamente con sus genes. Procedemos de esta forma hasta que todas las reacciones de la red han sido evaluadas. Se utilizan

tres diferentes valores de corte para definir las reacciones como esenciales, y para cada valor de corte se llevan a cabo 500 simulaciones de este tipo donde en cada simulación se permuta al azar el vector de reacciones de la red, de tal forma que al final tenemos 1500 redes mínimas sobre las cuales es posible identificar genes esenciales que se corresponderían con genes presentes en todas las redes mínimas, necesarios para la viabilidad del sistema en cualquier condición, y genes delecionables que serían aquellos ausentes en todas las redes mínimas y por tanto que podrían perderse sin afectar a la funcionalidad del sistema. Estas simulaciones se llevan a cabo además bajo dos condiciones de entorno diferentes. Por un lado en condiciones de entorno mínimas con solo glucosa y arginina, y por otro lado en condiciones ricas en nutrientes que tratan de simular las condiciones que *S. glossinidius* encuentra en el interior de la mosca tsé-tsé que consisten en permitir la entrada de 41 metabolitos diferentes adaptados de un estudio de Pal y colaboradores donde se utiliza una estrategia similar para modelizar la evolución de *W. glossinidia* a partir de la red funcional de *E. coli* JR904 (Pal et al., 2006b).

Los resultados de estas simulaciones aparecen representados en la Tabla 5.4. Los tres valores de corte dan resultados similares tanto en condiciones con glucosa y arginina como en condiciones ricas en nutrientes. Las 1500 redes mínimas en condiciones con glucosa y arginina como metabolitos externos presentan un tamaño promedio de 280,11 reacciones y 291,75 genes, con 237 reacciones y 255 genes comunes en todas las redes mínimas y 200 reacciones y 127 genes ausentes en todas las redes mínimas, lo que supone que el 84,61% de las reacciones y el 87,4% de los genes son comunes en las 1500 redes mínimas. Respecto a las simulaciones en condiciones ricas en nutrientes, las 1500 redes mínimas presentan un tamaño promedio de 231,9 reacciones y 237,25 genes, con 141 reacciones y 139 genes presentes en todas las redes mínimas y 185 reacciones y 111 genes ausentes en todas las redes mínimas, lo que supone que el 60,86% de las reacciones y el 58,69% de los genes son comunes en las 1500 redes mínimas, Esto supone un mayor grado de plasticidad evolutiva en condiciones ricas en nutrientes en comparación con las simulaciones con glucosa y arginina, donde el rango de posibles resultados en términos de contenido génico y de reacciones de las redes mínimas está más restringido (Ver Figura 5.13).

Al comparar las redes mínimas en ambas condiciones (Tabla 5.5), encontramos que 138 de los 139 genes esenciales en las redes mínimas en condiciones ricas en nutrientes también lo son en condiciones con glucosa y arginina, lo cual es de esperar dado que una función génica esencial en condiciones ricas en nutrientes también es de esperar que lo sea en condiciones de entorno mínimas más restringidas. El conjunto de 138 genes y 141 reacciones esenciales tanto en condiciones con glucosa y arginina como en condiciones ricas en nutrientes se representan en la Tabla suplementaria 5.5, y incluyen reacciones implicadas en la producción de componentes de biomasa que no pueden ser asimilados a partir del

entorno, como pueden ser las rutas de biosíntesis de espermidina y putrescina conjuntamente con los genes del sistema de transporte ABC específicico de arginina, las rutas de biosíntesis de ácidos grasos, fosfolípidos, peptidoglicano y lipopolisacárido bacteriano, o las rutas de biosíntesis de cofactores como tetrahidrofolato, coenzima A, NAD y FAD. Existen también 117 genes y 96 reacciones que aparecen como esenciales en redes mínimas únicamente en condiciones con glucosa y arginina pero no en condiciones ricas en nutrientes (ver Tabla suplementaria 5.6), que se corresponderían con genes que podrían perderse en el contexto de la asociación de *S. glossinidius* con la mosca tsé-tsé, y que de hecho incluyen un gran número de genes implicados en la biosíntesis de aminoácidos que se encuentran también ausentes en el genoma de *W. glossinidia*, el endosimbionte primario de la mosca tsé-tsé en etapas más avanzadas del proceso de reducción genómica (Akman et al., 2002; Zientz et al., 2004). De acuerdo con su inclusión en la ecuación de biomasa, estos genes son esenciales en condiciones con glucosa y arginina como únicos metabolitos externos, mientras que en condiciones ricas en nutrientes su carácter esencial depende de que el aminoácido se encuentre presente en el entorno y de que exista un sistema de transporte en la red funcional de *S. glossinidius*. Esto hace que los genes implicados en la biosíntesis de asparagina, aspartato, alanina, cisteina, serina, glicina, histidina, treonina, lisina, tirosina, triptófano, fenilalanina, valina, leucina y isoleucina aparezcan como esenciales únicamente en las condiciones con glucosa y arginina, ya que para todos ellos la red metabólica de *S. glossinidius* retiene un sistema de transporte funcional. En el caso de la ruta de biosíntesis de lisina, únicamente el último gen de la ruta, *lysA* (SG1988), que codifica para una diaminopimelato descarboxilasa que cataliza la descarboxilación de meso-diaminopimelato a L-lisina, aparece incluido en esta lista mientras que el resto de genes de la ruta de biosíntesis desde L-aspartato a meso-diaminopimelato aparecen como esenciales en todas las condiciones, ya que el meso-diaminopimelato es un precursor esencial para la biosíntesis de peptidoglicano. Este resultado también se ha observado experimentalmente en knockouts de *E. coli* generados en condiciones ricas en nutrientes, donde mutantes *lysA* son viables (Gerdes et al., 2003). Del mismo modo, encontramos casos como el de la ruta de biosíntesis de aminoácidos aromáticos, donde los genes iniciales de la ruta correspondientes a la biosíntesis de corismato desde eritrosa 4-fosfato aparecen como esenciales en todas las condiciones ya que el corismato es esencial para la biosíntesis de tetrahidrofolato, mientras que los genes finales implicados en la biosíntesis de los aminoácidos aromáticos a partir de corismato aparecen como esenciales únicamente en condiciones con glucosa y arginina debido a la presencia del transportador APC de aminoácidos aromáticos *AroP* codificado por el gen SG0465 (ver Figura 5.14).

Respecto a los genes susceptibles de ser delecionados, hay 109 genes y 172 reacciones que están ausentes en todas las redes mínimas tanto en condiciones con glucosa y arginina como en condiciones ricas en nutrientes (ver Tabla suplementaria

5.7). Estos genes incluyen los remanentes de rutas biosintéticas inactivas en *S. glossinidius* como los genes funcionales implicados en la biosíntesis de arginina o los genes implicados en la ruta de Entner-Doudoroff de degradación de hexurónidos (glucuronato y fructuronato) a los intermediarios glicolíticos gliceraldehido 3-fosfato y piruvato (Peekhaus and Conway, 1998), en concordancia con la inactivación de los genes *edd* y *eda* (ps_SGL1378 y ps_SGL0711 respectivamente) responsables de los pasos finales de la ruta y la ausencia de transportadores de hexurónidos. Sin embargo, esta lista también incluye rutas biosintéticas completas cuyos productos finales no se encuentran incluidos en la ecuación de biomasa y, como consecuencia, no serán producidos por el sistema bajo ninguna circunstancia. Entre ellos se incluyen los genes implicados en la biosíntesis de ubiquinona, piridoxina 5-fosfato, biotina, y grupos hemo, cofactores que se han postulado como esenciales en la asociación de la mosca tsé-tsé con sus hospedadores y que de hecho se han retenido en el genoma de *W. glossinidia* (Akman et al., 2002).

Al comparar el conjunto de genes esenciales y delecionables en cada una de las condiciones con el contenido génico real de *W. glossinidia* mediante la identificación de ortólogos entre *S. glossinidius* y *W. glossinidia* vemos que las simulaciones ricas en nutrientes mejoran la sensibilidad (fracción de genes esenciales presentes en *W. glossinidia*) y la especificidad (fracción de genes delecionables ausentes en *W. glossinidia*) respecto a las simulaciones en condiciones con glucosa y arginina, valores que aumentan si consideramos a ubiquinona, biotina, piridoxina 5-fosfato y grupos hemo como componentes de biomasa (Figura 5.15), lo que demuestra la capacidad de estas aproximaciones a la hora de predecir la evolución del contenido génico en este contexto de evolución reductiva (Pal et al., 2006b). Además, dentro de los genes esenciales en condiciones ricas en nutrientes encontramos 43 genes ausentes en el genoma de *W. glossinidia* que incluyen los genes de biosíntesis de corismato y p-aminobenzoato, esenciales en el contexto de la ruta de biosíntesis de tetrahidrofolate que como describimos en el capítulo anterior de esta tesis aparecen ausentes en *W. glossinidia*, indicando un posible suceso de complementación a nivel de suplemento de p-aminobenzoato por parte de *S. glossinidius* a *W. glossinidia* para poder sintetizar tetrahidrofolato. Una mayor predictividad de las simulaciones sería posible mediante un ajuste de la ecuación de biomasa que reflejara las capacidades biosintéticas de *W. glossinidia*, que no incluirían tetrahidrofolato o coenzima A a menos que exista un suplemento externo de p-aminobenzoato y β-alanina tal y como se describe en el capítulo anterior de esta tesis.

Análisis evolutivo de genes esenciales y delecionables en simulaciones de evolución reductiva

Con el objetivo de analizar si el carácter esencial o delecionable de los genes de la red metabólica de *S. glossinidius* definido en base a la importancia del gen para la

funcionalidad del sistema en términos de producción de biomasa se corresponde con patrones evolutivos y de expresión diferenciales entre ambos conjuntos de genes, se ha analizado los valores de Índice de Adaptación de Codones (CAI) de *S. glossinidius*, y el número de sustituciones sinónimas por sitio sinónimo (dS) y sustituciones no sinónimas por sitio no sinónimo (dN) entre *S. glossinidius* y *E. coli* K12. La hipótesis de partida es que los genes esenciales para la funcionalidad del sistema estarían más restringidos a nivel de cambios nucleotídicos que los genes cuya deleción no afectaría a la viabilidad del sistema. Esta hipótesis, propuesta por Alan Wilson y colaboradores en 1977, se basa en la idea de que dos proteínas sujetas a las mismas restricciones funcionales pero que difieren en el nivel de esencialidad evolucionarán a tasas diferentes (Wilson et al., 1977; Hurst and Smith, 1999). Los resultados es estos análisis sobre los genes esenciales y delecionables en redes mínimas en condiciones con glucosa y arginina como metabolitos externos muestran diferencias estadísticamente significativas para los tres parámetros que consideramos, de tal forma que los genes esenciales presentes en todas las redes mínimas presentan un valor de CAI promedio significativamente superior que los genes delecionables ausentes en todas las redes mínimas, que a su vez presentan unos valores promedio de dN y dS significativamente superiores a los de los genes esenciales (Figuras 5.15 y 5.16).Estos resultados se repiten cuando se analizan los genes esenciales y delecionables en condiciones ricas en nutrientes (Tabla 5.6), y se ajustan a lo que cabría esperar bajo la hipótesis de esencialidad de Wilson y colaboradores. Resultados análogos a los que observamos también fueron obtenidos por Jordan y colaboradores en tres linajes bacterianos (*E. coli*, *Helicobacter pylori* y *Neisseria meningitidis*) analizando los valores promedio de dN y dS a partir de genes esenciales definidos experimentalmente en *E. coli* (Jordan et al., 2002). Del mismo modo, se ha observado que genes esenciales definidos experimentalmente en *E. coli* y *B. subtilis* muestran valores promedio de CAI significativamente superiores a los de genes no esenciales (Fang et al., 2005), lo que indica una concordancia significativa entre nuestros resultados en base a simulaciones computacionales y resultados experimentales. En el contexto de la reducción genómica, se ha observado que genes con bajos valores de CAI y altos valores de dN se corresponden con genes selectivamente menos restringidos que se han perdido de forma predominante en linajes de endosimbiontes mutualistas ancestrales desde la divergencia del ancestro de vida libre (Delmotte et al., 2006), mientras que genes con niveles de expresión potencialmente elevados en base a los valores de CAI de sus correspondientes ortólogos en *E. coli* presentan una patrón de evolución más restringido tanto a nivel de sitios sinónimos como no sinónimos a pesar de la ausencia de uso de codones adaptativo como consecuencia de la aceleración en las tasas de sustitución característica de estos linajes (Herbeck et al., 2003; Schaber et al., 2005), lo que indica que los patrones evolutivos de genes esenciales y delecionables de *S. glossinidius* se corresponderían con genes susceptibles de ser eliminados a lo largo de su evolución futura.

Aunque este patrón evolutivo entre genes esenciales y delecionables se repite tanto con glucosa y ariginina como en condiciones ricas en nutrientes, las diferencias entre ambos conjuntos de genes (esenciales y delecionables) son menores en las condiciones ricas en nutrientes que en las condiciones con glucosa y arginina. La diferencia principal entre ambas condiciones reside en los 117 genes esenciales únicamente con glucosa y arginina, que como se ha descrito anteriormente incluye un gran número de genes ausentes en el genoma de *W. glossinidia* que se podrían perder en el contexto de la asociación de *S. glossinidius* con un ambiente rico en nutrientes como es el de los tejidos de la mosca tsé-tsé. Cuando se comparan los valores promedio de CAI, dN y dS entre este conjunto de genes condicionalmente esenciales y el resto de genes esenciales comunes en todas las condiciones vemos que, si bien no existen diferencias estadísticamente significativas a nivel de CAI, los genes condicionalmente esenciales presentan valores promedio de dN y dS significativamente menores que el resto de genes esenciales en todas las condiciones (Figura 5.18), lo cual va en contra de lo que cabría esperar si *S. glossinidius* estuviera siguiendo el mismo patrón de evolución reductiva que *W. glossinidia*. Esto podría indicar que estos genes condicionalmente esenciales continúan siendo importantes para *S. glossinidius* debido a su muy reciente asociación con el hospedador, que haría que las presiones selectivas que han gobernado la evolución de *W. glossinidia* no estuvieran actuando con la misma intensidad sobre *S. glossinidius*, una observación que se ajusta a lo que se observa al analizar los patrones evolutivos de los 48 genes de biosíntesis de cofactores ausentes en todas las redes mínimas en ambas condiciones, que aparecen más próximos al resto de genes no esenciales para la funcionalidad del sistema (Figura 5.7). Sin embargo, aunque los valores de CAI se calculan de forma específica para los genes de *S. glossinidius* en base al uso de codones de un subconjunto de genes con niveles de expresión probablemente elevados como genes de la maquinaria ribosomal (Sharp and Li, 1987a; Puigbo et al., 2008a), los valores de dN y dS están calculados entre *S. glossinidius* y *E. coli*, y por tanto reflejan los cambios que han tenido lugar en cada linaje desde su divergencia del ancestro común. En este contexto, la aceleración en las tasas de sustitución sinónimas y no sinónimas es una característica común en la evolución de los endosimbiontes bacterianos en comparación con bacterias de vida libre próximas asociada a la transición al modo de vida dependiente de hospedador (Moran, 1996; Brynnel et al., 1998; Clark et al., 1999).

Análisis de tasas de sustitución sinónimas y no sinónimas especificas de linaje

Para poder calcular los valores de dN y dS específicos de los linajes de *S. glossinidius* y *E. coli* K12 es necesario incluir en la comparación una tercera especie que actúe como grupo externo. Para ello elegimos *Vibrio cholerae* O1 en base a la reconstrucción filogenética de γ-proteobacterias obtenida en el capítulo 3 de esta tesis. Con las estimas de dN y dS entre *S. glossinidius*-*E. coli* K12, *S. glossinidius*-*V. cholerae* O1 y *E. coli* K12-*V. cholerae* O1, las dN y dS específicas de la rama de *S.*

glossinidius y *E. coli* K12 se pueden calcular a través de las ecuaciones representadas en la Figura 5.4, y permiten evaluar si las conclusiones inferidas a partir de las estimas entre *S. glossinidius* y *E. coli* se reproducen cuando la evolución de cada linaje se analiza de forma independiente. Asimismo, nos permitirá determinar si existe algún tipo de aceleración en las tasas de sustitución en el linaje de *S. glossinidius*.

Los valores promedio de dN y dS para cada linaje entre genes esenciales y delecionables en condiciones con glucosa y arginina están representados en la tabla 5.8, y muestran que en ambos linajes los genes delecionables presentan unos valores de dN promedio significativamente superiores que los valores dN promedio de los genes esenciales. Estos resultados se repiten al analizar los valores de dS para ambos linajes, aunque en este caso las diferencias no son estadísticamente significativas. Además, tanto en las dN como en las dS, los valores promedio de genes esenciales y delecionables son superiores en el linaje de *S. glossinidius* respecto al de *E. coli*, lo que apunta a una posible aceleración en las tasas de sustitución en el linaje de *S. glossinidius* ya que las estimas de dN y dS para cada linaje están calculadas para un mismo periodo evolutivo, que seria el periodo comprendido desde la divergencia de ambos linajes desde su ancestro común. La comparación de los valores de dN y dS para cada gen en ambos linajes mediante un test T de muestras emparejadas muestra diferencias estadísticamente significativas que indican que *S. glossinidius* esta acumulando un mayor número de sustituciones sinónimas y no sinónimas que *E. coli* K12, siendo esta diferencia mayor en los sitios sinónimos (media(*dSsgl-dSeco*) = 0.543; media(*dNsgl-dNeco*) = 0.01432). El hecho de que estas estimas estén calculadas para un mismo periodo evolutivo en ambos linajes, correspondiente al tiempo de divergencia desde su ancestro común, hace que estos resultados se puedan interpretar como una aceleración significativa en las tasas de sustitución sinónimas y no sinónimas en el linaje de *S. glossinidius*, aunque no sea posible calcular tasas de sustitución propiamente dichas ya que el tiempo de divergencia exacto entre *S. glossinidius* y *E. coli* es desconocido.

La aceleración en las tasas de sustitución en el linaje de *S. glossinidius*, especialmente a nivel de los sitios sinónimos, podría estar afectando a sus patrones de uso adaptativo de codones. Mientras que una correlación negativa entre CAI y dN es un indicativo de una mayor conservación de genes con elevados niveles de expresión (Herbeck et al., 2003; Rocha and Danchin, 2004), una correlación negativa entre CAI y dS se explica por acción de la selección purificadora en sitios sinónimos en genes con niveles de expresión elevados como consecuencia del uso adaptativo de codones (Sharp and Li, 1987b; Sharp, 1991). De hecho, en bacterias endosimbiontes con asociaciones ancestrales con el insecto hospedador como *B. aphidicola* se observa una distribución homogénea de sustituciones sinónimas entre genes, lo que indica una ausencia de uso adaptativo de codones consecuencia de la aceleración en las tasas de sustitución y del sesgo composicional hacia AT (Clark et

al., 1999). Cuando llevamos a cabo estos mismos análisis con el conjunto de genes esenciales y delecionables en condiciones con glucosa y arginina, vemos que aunque existe una correlación negativa significativa entre CAI y dN en ambos linajes, indicativo de una mayor conservación de genes altamente expresados, no existe correlación negativa significativa entre CAI y dS en el linaje de *S. glossinidius* (Ver Figura 5.19), lo cual se explicaría por la aceleración en las tasas de sustitución sinónimas que observamos en el linaje de *S. glossinidius*. Estos resultados apuntan más hacia una disminución en la intensidad del uso adaptativo de codones en el linaje de *S. glossinidius* en comparación con *E. coli* que a una ausencia de uso adaptativo de codones en el linaje de *S. glossinidius*. De hecho, la correlación negativa que se observa entre CAI y dS en *E. coli*, aunque significativa, es menor (r=-0.213) que la correlación que se observa cuando el análisis se lleva a cabo entre *E. coli* y *S. typhimurium* (r=-0.68), dos linajes evolutivamente mucho más próximos (Sharp and Li, 1987b; Sharp, 1991), que se explicaría por el hecho de que nuestras estimas se han llevado a cabo para un tiempo evolutivo mucho mayor, correspondiente la divergencia desde el ancestro común con *S. glossinidius*. En el caso del linaje de *S. glossinidius*, el uso adaptativo de codones se refleja por la correlación negativa entre CAI y dN, así como por el agrupamiento significativo de genes putativamente con altos niveles de expresión cuando se lleva a cabo un análisis de correspondencia del Uso de Codones Sinónimos Relativo (RCSU) de todos los genes del genoma (Puigbo et al., 2008b) (Ver Figura 5.20), de tal forma que la ausencia de correlación negativa significativa entre CAI y dS en *S. glossinidius* se puede explicar por la aceleración en las tasas de sustitución sinónimas y el tiempo de divergencia elevado sobre el cual llevamos a cabo las estimas, aunque la comparativa con *E. coli* sí que revela una disminución de la intensidad de este uso adaptativo de codones.

El efecto de la aceleración en las tasas de sustitución en el linaje de *S. glossinidius* es más evidente cuando se comparan el conjunto de genes condicionalmente esenciales con el resto de genes esenciales en todas las condiciones. Cuando se comparan los valores promedio de dN y dS de genes esenciales y delecionables en condiciones ricas en nutrientes se observa que, aunque los genes esenciales presentan valores promedio de dN y dS inferiores que los genes delecionables, estas diferencias no son estadísticamente significativas (Tabla 5.9). Esto puede ser consecuencia de un patrón de evolución más conservado de los genes condicionalmente esenciales, que no están presentes cuando analizamos los genes esenciales y delecionables en condiciones ricas en nutrientes. Como se ha apuntado anteriormente, estos genes condicionalmente esenciales representan genes que se podrían perder en un contexto rico en nutrientes similar al que encuentra *S. glossinidius* en el interior de la mosca tsé-tsé, y de hecho contiene un gran número de genes, especialmente implicados en la síntesis de aminoácidos, que están ausentes en *W. glossinidia*. Cuando se comparan los valores promedio de dN y dS de estos genes condicionalmente esenciales y el resto de genes esenciales en todas las condiciones en ambos linajes se

observa que en el linaje de *E. coli*, estos genes condicionalmente esenciales se encuentran significativamente más conservados tanto en sitios sinónimos como no sinónimos que el resto de genes esenciales en ambas condiciones (Tabla 5.10), de acuerdo con el papel esencial que estos genes desempeñan en el contexto de una bacteria de vida libre con unas condiciones de entorno mucho más restringidas. Sin embargo, en *S. glossinidius*, aunque estos genes condicionalmente esenciales se encuentran significativamente más conservados que el resto de genes esenciales en ambas condiciones a nivel de sitios no sinónimos, las diferencias entre ambos grupos son menores que en el linaje de *E. coli*, mientras que a nivel de los sitios sinónimos se observa la tendencia opuesta, con los genes condicionalmente esenciales que acumulan sustituciones sinónimas a una tasa significativamente superior que el resto de genes condicionalmente esenciales en ambas condiciones. Estos resultados apuntan a una relajación en la presión selectiva sobre estos genes condicionalmente esenciales en el linaje de *S. glossinidius* como consecuencia de su localización en un ambiente rico en nutrientes en comparación con su papel más esencial, y como consecuencia su mayor conservación a nivel de secuencia, que se observa en el linaje de una bacteria de vida libre como *E. coli* K12. Además, es importante tener en cuenta que las estimas de dN y dS en el linaje de *S. glossinidius* se calculan para todo el periodo de evolución desde la divergencia del ancestro de vida libre común con *E. coli* K12. Este periodo evolutivo, como hemos visto en el capítulo 3 de esta tesis, estaría dividido en un primer periodo donde *S. glossinidius* estaría evolucionando como una bacteria de vida libre hasta su transición al modo de vida dependiente de hospedador, y un segundo periodo desde esta transición al modo de vida dependiente de hospedador hasta el periodo actual. Durante el primer periodo, *S. glossinidius* estaría evolucionando a tasas similares a las de una bacteria de vida libre como *E. coli*, mientras que en el segundo periodo correspondiente a la evolución como bacteria endosimbionte es donde tendría lugar la aceleración en las tasas de sustitución que observamos tanto a nivel de sitios sinónimos como no sinónimos. Esto significa que la aceleración que observamos para toda la rama de *S. glossinidius* desde la divergencia del ancestro con *E. coli* será mayor durante el periodo evolutivo correspondiente a la evolución como bacteria endosimbionte, lo que indica que, a pesar de la muy reciente transición al modo de vida dependiente de hospedador, el análisis evolutivo de genes esenciales y delecionables en base a las simulaciones sobre la red metabólica muestra que *S. glossinidius* empieza a mostrar características asociadas a bacterias endosimbiontes con asociaciones mucho mas ancestrales.

DISCUSIÓN GENERAL

Evolución de la organización genómica en γ-proteobacterias

La disponibilidad de múltiples genomas completos de bacterias con diferentes niveles de proximidad evolutiva es un factor esencial para poder estudiar la

evolución de la organización genómica. Además ha permitido estudiar en el capítulo 3 de esta tesis como evoluciona la organización genómica en diferentes linajes bacterianos con características ecológicas variables. La comparación de las distancias de orden génico con las distancias basadas en secuencias, las cuales podemos considerar como proporcionales al tiempo de divergencia entre linajes, ha revelado que , si bien existe un incremento en el número de reordenaciones cromosómicas a medida que aumenta el tiempo de divergencia entre especies, este incremento presenta una marcada heterogeneidad en diferentes linajes dentro de las γ-proteobacterias, encontrando desde linajes con una estabilidad genómica total como los endosimbiontes primarios de insectos con asociaciones ancestrales con el hospedador, ejemplificados por los genomas de *B. aphidicola*, hasta linajes donde se acumulan un gran número de reordenaciones como son el linaje de las Pasteurellas. Esto hace que cualquier intento de utilizar estos caracteres para reconstruir la historia evolutiva de bacterias de manera fidedigna necesite considerar de manera específica esta heterogeneidad así como la prevalencia de los diferentes sucesos de reordenación que tienen lugar en los genomas bacterianos, ya que si bien en γ-proteobacterias encontramos que las inversiones son mayoritarias, es necesario considerar también el papel esencial que los sucesos de deleción están jugando en la evolución de los linajes de bacterias endosimbiontes de insectos. No obstante, la comparación de las filogenias obtenidas en base a datos de orden génico y las obtenidas en base a secuencia nos permite tener una visión más global de la evolución bacteriana, revelando que la acumulación de reordenaciones es un proceso común en bacterias que han sufrido una transición a un modo de vida dependiente de hospedador, no solo endosimbiontes, sino también bacterias patógenas como *Shigella flexneri* o *Yersinia pestis*, donde un periodo de tiempo relativamente corto (menos de 1 millón de años) se acumulan un número de reordenaciones significativamente superior al que se observa entre bacterias evolutivamente más alejadas como *E. coli* y Salmonella, aunque muchos de estos reordenaciones pueden reflejar polimorfismos transitorios que serán eliminados por la selección natural.

Sin embargo, a pesar de la ausencia de modelos evolutivos detallados que permitan modelizar la evolución del orden génico, la comparación de las filogenias basadas en estas distancias con filogenias basadas en datos de secuencia permiten tener una visión más global del proceso evolutivo que resulta particularmente interesante en el caso de la evolución de bacterias endosimbiontes, cuyo monofiletismo en base a los datos de secuencia contrasta con el parafiletismo que se observa en base a datos de orden génico. En este caso, si bien los datos de orden génico presentan el problema de la ausencia de modelos evolutivos que permitan inferir distancias evolutivas reales a partir de diferencias observadas, los modelos basados en datos de secuencia presentan problemas similares cuando existe heterogeneidad en las tasas de sustitución entre diferentes linajes, lo cual tiene lugar de manera específica en estos linajes de bacterias endosimbiontes ancestrales.

Breve resumen en castellano

El análisis de linajes endosimbiontes a diferentes estadios de la asociación con el hospedador ha revelado también una marcada aceleración en las tasas de reordenación genómica en comparación con bacterias entéricas de vida libre, de manera similar a lo que se observa con datos de secuencia. Además, estas reordenaciones se concentran en un periodo evolutivo muy concreto como son las etapas iniciales de la asociación con el hospedador, ejemplificado en el genoma de *S. glossinidius*, cuya tasa de evolución a nivel de inversiones cromosómicas es aproximadamente el doble que en bacterias entéricas de vida libre, aunque esta aceleración también se puede inferir en linajes endosimbiontes más ancestrales en base a la estricta conservación del orden génico entre cepas de un mismo linaje (genomas de *B. aphidicola*) que contrasta con la aceleración en las tasas de reordenaciones que se observa al comparar con bacterias entéricas de vida libre.

Dinámicas de evolución genómica en *S. glossinidius*, el endosimbiontes secundario de la mosca tsé-tsé

Las particulares características evolutivas de *S. glossinidius* dentro de los endosimbiontes bacterianos de insectos y la marcada aceleración en las tasas de reordenaciones genómicas observada en el capítulo 3 de esta tesis nos llevó a estudiar más en profundidad la estructura y características funcionales de este genoma, uno de los pocos disponibles de una bacteria en etapas muy tempranas de la asociación con el hospedador. Además, el hecho de que *S. glossinidius* forme parte de un sistema endosimbiótico conjuntamente con *W. glossinidia*, un mutualista obligado ancestral de la mosca tsé-tsé responsable de suplementar cofactores al hospedador, permite estudiar conjuntamente el perfil metabólico de ambos endosimbiontes en el contexto de la asociación con el hospedador. Para ello tuvimos que llevar a cabo una reanotación funcional completa del genoma de *S. glossinidius*, incluyendo la caracterización y anotación funcional de un total de 1501 pseudogenes que contrastan con los 972 pseudogenes que fueron descritos pero no anotados en la anotación original del genoma, y que refleja la importancia del análisis de secuencia a la hora de identificar y caracterizar pseudogenes, especialmente en este tipo de genomas donde la transición a un modo de vida dependiente de hospedador, con relajación en la presión selectiva sobre muchos genes ancestrales y un incremento del efecto de la deriva genética consecuencia de la reducción del tamaño efectivo poblacional, produce la acumulación de un gran número de pseudogenes, muchos de los cuales se encuentran en etapas avanzadas del proceso de degradación, lo cual dificulta su identificación utilizando programas de predicción génica.

La reanotación funcional de todo el genoma incluyendo genes y pseudogenes ha revelado la presencia masiva de elementos genéticos móviles, principalmente de origen fágico, siendo la clase funcional más afectada por la pseudogenización probablemente como consecuencia de la estricta localización de *S. glossinidius* en el interior de la mosca tsé-tsé, donde se limita el intercambio de este tipo de elementos

genéticos móviles por transferencia genética horizontal. A través de la genómica comparada ha sido posible caracterizar 2 inserciones fágicas completas y 11 parciales a lo largo del genoma de *S. glossinidius*, así como un posible flujo de material genético entre el cromosoma y el elemento extracromosómica de origen fágico pSG3. La caracterización de los diferentes elementos IS presentes en el genoma revela que estos representan un 2.71% del genoma, con una densidad de elementos IS similar a las de otras bacterias simbióticas y patógenas con reciente transición al modo de vida dependiente de hospedador, aunque su papel en el proceso de inactivación génica es mínimo si tenemos en cuenta que solo 18 pseudogenes se producen como consecuencia de la inserción de un elemento IS. Sin embargo, estos elementos IS pueden jugar un papel clave en la observada aceleración de las tasas de reordenaciones cromosómicas que esta teniendo lugar en *S. glossinidius*, ya que su dispersión a lo largo del genoma y el elevado grado de identidad entre copias hacen que puedan actuar como puntos de recombinación homóloga, la cual se podría llevar a cabo gracias a la presencia de todos los genes implicados en la maquinaria recombinacional en *S. glossinidius*.

La reconstrucción del metabolismo completo de *S. glossinidius* en base a los resultados de la reanotación funcional reflejan un perfil metabólico más próximo al de una bacteria de vida libre que al de un endosimbionte clásico, reteniendo la mayoría de rutas de biosíntesis de componentes celulares esenciales con la excepción de la ruta de biosíntesis de L-arginina y de tiamina. En el caso de la L-arginina, supone un cambio significativo respecto a las conclusiones de la anotación original, donde se postulaba que *S. glossinidius* era capaz de sintetizar todos los aminoácidos esenciales excepto L-alanina. La re-anotación ha revelado que la ruta de biosíntesis de alanina se encuentra completamente funcional mientras que la ruta de biosíntesis de L-arginina se encuentra inactivada por cuatro sucesos de pseudogenización, lo que implica la necesidad de un suplemento externo de arginina para la viabilidad celular. Respecto a la ruta de biosíntesis de tiamina, es la única ruta metabólica que aparece inactiva en *S. glossinidius* y aparentemente funcional en *W. glossinidia*. Sin embargo el análisis del genoma de *W. glossinidia* revela la inactivación del gen *thiF* y la ausencia del gen *thiI*, ambos esenciales para la biosíntesis de tiamina, revelando un escenario de posible complementación metabólica entre ambos endosimbiontes como única forma de sintetizar la forma activa del coenzima, la tiamina difosfato, y que podría explicar la coexistencia de ambos endosimbiontes en el interior de la mosca tsé-tsé, a través de una inactivación reciente del gen *thiF* de *W. glossinidia* que sería compensada por la presencia de *S. glossinidius*. Este escenario también facilitaría la síntesis de folato y coenzima A, otros dos cofactores cuyas rutas de biosíntesis se encuentran significativamente incompletas en *W. glossinidia* y que necesitarían del suplemento de β-alanina (para sintetizar coenzima A) y p-aminobenzoato (para sintetizar tetrahidrofolato) para poder producir estos dos cofactores. No obstante, es importante tener en cuenta que los genomas de *S. glossinidius* y *W. glossinidia* pertenecen a especies de hospedador

diferentes, con lo cual es posible que lo que estamos observando no se ajuste a la situación real dentro del correspondiente hospedador, para lo cual sería necesario disponer de los genomas de *W. glossinidia* de *Glossina morsitans morsitans* y *S. glossinidius* de *Glossina brevipalpis*.

Reconstrucción y análisis de la red metabólica de *S. glossinidius* en diferentes etapas del proceso de evolución reductiva

El estudio de redes metabólicas a través del análisis de balance de flujos (FBA) permite una aproximación sistémica a las capacidades funcionales de un organismo a partir de la información de su genoma. Si bien con la información de la anotación de genes y pseudogenes hemos podido reconstruir el metabolismo completo de *S. glossinidius* de manera cualitativa, viendo qué funciones génicas se encuentran funcionales o inactivas y como estos sucesos de inactivación afectarían a rutas metabólicas concretas, la reconstrucción de la red metabólica ancestral y funcional de *S. glossinidius* proporciona una visión global de como todas estas funciones celulares se integran en su conjunto para asegurar la funcionalidad del sistema a diferentes etapas del proceso reductivo. Este análisis revela que la red metabólica ancestral de *S. glossinidius* es completamente funcional en un entorno aerobio mínimo con únicamente glucosa como fuente de carbono a partir de la cual producir todos los componentes de biomasa que describen el crecimiento celular de una bacteria de vida libre como *E. coli* K12, y que la inactivación de los genes de biosíntesis de glicógeno, arginina, y especialmente la inactivación del gen anaplerótico *ppc* producen un cambio drástico en las capacidades funcionales del sistema que hace que dependa de forma esencial de un suplemento externo de arginina para sobrevivir, lo que se ajusta con la transición desde un modo de vida libre a un modo de vida dependiente de hospedador que asegure este suplemento adicional de arginina.

El análisis comparado de las redes metabólicas de *S. glossinidius* y *E. coli* K12 revela también que el proceso de inactivación génica se encuentra asociado con una disminución de la robustez de los sistemas metabólicos tanto a sucesos de pérdida génica como a cambios en el rango de flujo metabólico a través de diferentes reacciones de la red que contrasta con la robustez estructural de las redes biológicas consecuencia de su organización libre de escala. Esto refleja diferencias en el patrón evolutivo funcional y estructural de los sistemas biológicos, de tal forma que si bien la integridad estructural del sistema se mantiene a pesar de la pérdida de genes, esta pérdida si que afecta de forma drástica a la funcionalidad del sistema, lo que determinaría la necesidad obligada de este tipo de bacterias endosimbiontes de un ambiente estable con pocas fluctuaciones como el que encuentran el interior del hospedador. En este contexto, *S. glossinidius*, con un 56% de genes esenciales para la funcionalidad de la red metabólica funcional, se encontraría en una situación intermedia entre una bacteria de vida libre como *E. coli* (24% de genes esenciales) y

un endosimbionte mutualista obligado como *B. aphidicola* (84% de genes esenciales), que se correlacionaría con el tropismo más amplio de *S. glossinidius* a nivel de localización tanto intracelular como extracelular en diferentes tejidos del hospedador frente a la estricta localización en el interior de bacteriocitos de un endosimbionte primario mutualista obligado como *B. aphidicola* donde las condiciones de entorno son mucho mas estables.

Por último, las simulaciones de evolución reductiva sobre la red funcional de *S. glossinidius* han permitido reducir este sistema hasta el mínimo número de funciones metabólicas capaces de asegurar la supervivencia celular en términos de producción de biomasa bajo diferentes condiciones de entorno, lo cual supone una aproximación a la posible evolución futura de *S. glossinidius* en el contexto de la reducción genómica. El contenido génico de estas redes metabólicas mínimas es altamente dependiente de las condiciones de entorno del sistema, de tal forma que en unas condiciones ricas en nutrientes que simulan el entorno que *S. glossinidius* encuentra en el interior de la mosca tsé-tsé vemos que las redes mínimas presentan una mayor variabilidad en el número de reacciones y genes que en condiciones de entorno mínimas. Del mismo modo, la combinación de estas condiciones de entorno próximas a las del interior de la mosca tsé-tsé con ajustes precisos de la ecuación de biomasa producen redes mínimas cuyo contenido génico reproduce con una elevada especificidad y sensibilidad el contenido génico real de un genoma mínimo como es *W. glossinidia*. Del mismo modo, el análisis de los patrones de evolución de genes esenciales y delecionables en *S. glossinidius* definidos en base a su presencia o ausencia en las redes metabólicas mínimas revelan que los genes esenciales comunes en todas las redes mínimas se corresponden con genes significativamente más expresados y con menores tasas de evolución tanto a nivel de sustituciones sinónimas como no sinónimas, una observación que coincide con diferentes estudios de esencialidad en base a inactivaciones experimentales. Sin embargo, el linaje de *S. glossinidius* muestra una aceleración significativa en las tasas de sustitución tanto a nivel de sitios no sinónimos como especialmente a nivel de sitios sinónimos, una característica común de las bacterias endosimbiontes, que parece afectar a su uso adaptativo de codones en comparación con una bacteria de vida libre como *E. coli*, y que aparece evidente al analizar los genes condicionalmente esenciales que se podrían perder en un entorno rico en nutrientes y que se corresponden con un gran número de genes ausentes en el genoma de *W. glossinidia*, cuyos patrones evolutivos en *S. glossinidius* aparecen mucho menos restringidos que en *E. coli*, una bacteria de vida libre donde estos genes continúan siendo esenciales para la supervivencia celular. Esto apunta a que *S. glossinidius* empieza a mostrar características propias de un endosimbionte bacteriano ancestral.

10 Supplementary files

Supplementary Table4.1: Functional classification outline: Scheme of the functional classification used during the re-annotation process.

Supplementary Table 4.2:Putative CDSs including an originally annotated gene and a re-annotated pseudogene. The coordinates corresponds to the combined limits of the gene and pseudogene corresponding to each putative CDS. "Coverage" columns show the percentage of the putative CDS represented by the originally annotated gene and the adjacent pseudogene together with the name of the originally annotated gene. "Relative position" columns represents the relative position of gene and pseudogene in the putative CDS (5P representing the 5' end of the putative CDS and 3P representing the 3' end of the putative CDS). Putative CDS with coordinates 2812192-2814586 includes one gene and two pseudogenes.

Supplementary Table 5.1: Reactions and genes included in *S. glossinidius* ancestral metabolic network.

Supplementary Table 5.2: Reactions and genes included in *S. glossinidius* functional metabolic network.

Supplementary Table 5.3: Reactions of *S. glossinidius* ancestral network that includes pseudogenes without sequence similarity with E. coli K12. These 27 pseudogenes are highlighted in red.

Supplementary Table 5.4: Reactions of *S. glossinidius* ancestral network that includes S. glossinidius specific genes without sequence similarity with E. coli K12. These 13 pseudogenes are highlighted in red.

Supplementary Table 5.5: Common essential genes and reactions in all minimal networks under nutrient-limited and nutrient-rich conditions.

Supplementary Table 5.6: Conditionally essential genes and reactions in minimal networks only under nutrient-limited conditions.

Supplementary Table 5.7: Common disposable genes and reactions absents in all minimal networks under nutrient-limited and nutrient-rich conditions.

Supplementary Table 5.8: Cofactor biosynthesis genes and reactions absents in all minimal networks under nutrient-limited and nutrient-rich conditions