Selenoproteins across the tree of life

Methods and applications

Autor: Didac Santesmasses Ruiz

TESI DOCTORAL UPF / ANY L'any de la tesi: 2016

DIRECTOR DE LA TESI
Roderic Guigó i Serra - Bionformatics and Genomics, Centre de Regulació Genòmica (CRG)

upf. Universitat Pompeu Fabra Barcelona

Per a la Maria

Abstract

Selenocysteine is known as the 21st amino acid. Selenoproteins incorporate selenocysteine in response to specific UGA codons through a recoding mechanism, which present in the three domains of life, but not in all organisms. Standard gene prediction programs consider UGA only as stop, and selenoproteins are normally misannotated. We have developed computational methods for prediction of selenoproteins. By applying these and other tools, we have characterized selenoproteins across the Tree of Life, which shows a dynamic evolution of the utilization of selenocysteine in the different lineages. We have characterized the abundance and distribution of selenoproteins in the human microbiota. We characterized the selenoproteins in *Lokiarchaeota*, which have some eukaryotic-like features. Finally we gave special attention to insects, in which a progressive reduction in the number of selenoproteins culminated in multiple independent selenoprotein extinctions.

Resum

La selenocïsteina és coneguda com a l'aminoàcid 21. Les selenoproteïnes incorporen selenocïsteina en resposta a codons UGA específics mitjançant un mecanisme de recodificació, el qual és present en els tres dominis de la vida, però no en tots els organismes. Els programes estàndard per a la predicció de gens consideren UGA només com a codó stop, per aquesta raó l'anotació de selenoproteínes és, generalment, incorrecte. Hem desenvolupat mètodes computacionals per a la predicció de selenoproteïnes. Mitjançant l'aplicació d'aquestes i altres eines, hem caracteritzat selenoproteïnes a través de l'Arbre de la Vida, on hem observat una evolució dinàmica en la utilització de selenocïsteina en els diferents llinatges. Hem caracteritzat l'abundància i distribució de selenoproteïnes en el microbioma humà. Hem caracteritzat les selenoproteïnes presents a *Lokiarchaeota*, les quals presenten trets eucariòtics. Finalment hem dedicat especial atenció als insectes, en els quals una progressiva reducció en el nombre de selenoproteïnes culminà en múltiples extincions de selenoproteïnes en esdeveniments evolutius independents.

Preface

Selenoproteins are a diverse class of proteins that contain selenium in the form of the non-canonical amino acid selenocysteine (Sec). Sec is known as the 21st amino acid, and is considered as an expansion to the genetic code. The codon for Sec insertion is a UGA codon, normally stop codon, which is recoded as Sec through a complex molecular mechanism.

There are very few selenoproteins in a genome. Humans have 25 genes, *Drosophila* have 3, and *C. elegans* just 1. Sec is the main biological form of the trace element selenium, and selenoproteins are essential in mouse, and are tightly conserved in vertebrates. Although they are present in the three domains of life, many lineages do not have selenoproteins and do not use Sec. Fungi and plants are known to completely lack selenoproteins, and they were also lost in some insects and nematodes.

It is not trivial to identify selenoproteins. The vast majority of UGA codons in a genome correspond to stop codons, and standard gene prediction programs do no consider Sec. Specific methods are required to identify the specific features of selenoprotein genes and correctly predict Sec codons. During my PhD I worked on developing tools for selenoprotein predictions, and the Sec-specific $tRNA^{Sec}$.

In this thesis, we have analyzed genomes from the Tree of Life. We have characterized for the first time the abundance and distribution the selenoproteins present in the human microbiota. These are important to understand the utilization of selenium by the microbes we host in our body. We analyzed also the closest relative to eukaryotes known, *Lokiarchaeota*. This study shed some light into the evolution of the system for the synthesis of selenoproteins. We gave special emphasis to insects, in which a relaxation for the constrains on selenoproteins had previously been described. We now have a much detailed picture of the evolution of selenoproteins in this lineage.

Thanks to the growing number of sequenced genomes, and the sequencing efforts of our lab, we have been able to follow the fate of selenoproteins at genome level, and at large evolutionary scale.

Contents

Chapter 1

INTRODUCTION

1.1 Selenocysteine, the 21st amino acid

The standard amino acid alphabet is composed of 20 members. A notable natural expansion of the genetic code is selenocysteine (Sec), known as the 21st amino acid. Sec is a selenium-containing structural analogue of cysteine (Cys), with selenium (Se) in place of sulphur. Proteins carrying Sec are called selenoproteins. Sec is inserted into selenoproteins, during translation, in response to an in-frame UGA codon through a recoding mechanism. The canonical use of UGA is to terminate translation, but in selenoprotein mRNAs, UGA is recoded as a sense codon for Sec insertion. The mechanism by which UGA specifies Sec challenged the dogma that one codon can just have a single meaning in a given organism. Selenoproteins are found in the three domains of life, although no in all of organisms. The meaning of the UGA codon is ambiguous in selenoprotein-containing bacteria, archaea and eukaryotes.

1.2 Sec synthesis and insertion

The single selenium atom in Sec is costly for the organism. A set of dedicated factors is required for Sec biosynthesis and insertion into selenoproteins. Sec is formed on its own tRNA by the tRNA-dependent modification of serine (Ser). The process is reminiscent of the synthesis of glutamine and asparagine in some prokaryotes [Sheppard et al., 2008], and it is the only known amino acid in eu-

karyotes whose synthesis occurs on its tRNA. The canonical elongation factor, responsible for the delivery of the 20 standard amino acids to the ribosome, does not recognize $tRNA^{Sec}$. Instead, a Sec-specific elongation factors is required for Sec insertion.

1.2.1 $tRNA^{Sec}$

The Sec-specific tRNA ($tRNA^{Sec}$) is a key molecule and central component of the selenoprotein synthesis. The anticodon in $tRNA^{Sec}$ is UCA, complementary to the UGA codon. It is the longest tRNA with 90-100 nucleotides (nt) and has an unusual structure, different than that of canonical tRNAs. The 3D structure of $tRNA^{Sec}$ has been solved in the three domains of life: eukaryotes [Palioura et al., 2009,Itoh et al., 2009], archaea [Chiba et al., 2010] and bacteria [Itoh et al., 2013]. The $tRNA^{Sec}$ structure in bacteria has a 8 base pairs (bp) stem in the acceptor arm and a 5 bp stem in the T arm. The 8+5 structure not only differs from the 7+5 structure in canonical tRNAs, but also from the 9+4 structure in archaea and eukaryotes. The variable arm in $tRNA^{Sec}$ is remarkably long (with 5-9 bp stem), and the D arm has a 6 bp stem and a 4 nt loop, in contrast to the 3-4 bp stem and 7-12 nt loop in the D arm of other tRNAs. In archaea, the D stem has 7 bp, with the only exception of *Methanopyrus kandleri* that has a 6 bp D-stem [Sherrer et al., 2011] (see figure 1.1).

Two major isoforms of $tRNA^{Sec}$ are expressed in mammalian cells. They are distinguished by a post-transcriptional modification (mcm^5U or mcm^5Um) in position 34, the wobble position in the anticodon. Their relative distribution is influenced by the levels of Se. Under conditions of Se deficiency, the level of mcm^5U is greater than mcm^5Um, and during Se supplementation, the ratio of the two isoform is reversed. Interestingly, housekeeping selenoproteins are synthesized by the mcm^5U isoform, while stress-related selenoproteins are synthesized by mcm^5Um. Some selenoproteins appear to be synthesized by both isoforms ([Labunskyy et al., 2014] and references therein).

Remarkably, $tRNA^{Sec}$ genes in certain bacteria had recently reported to have an anticodon different that UCA. In those organisms, the selenoprotein transcripts support Sec insertion in codons different than UGA [Mukai et al., 2016].

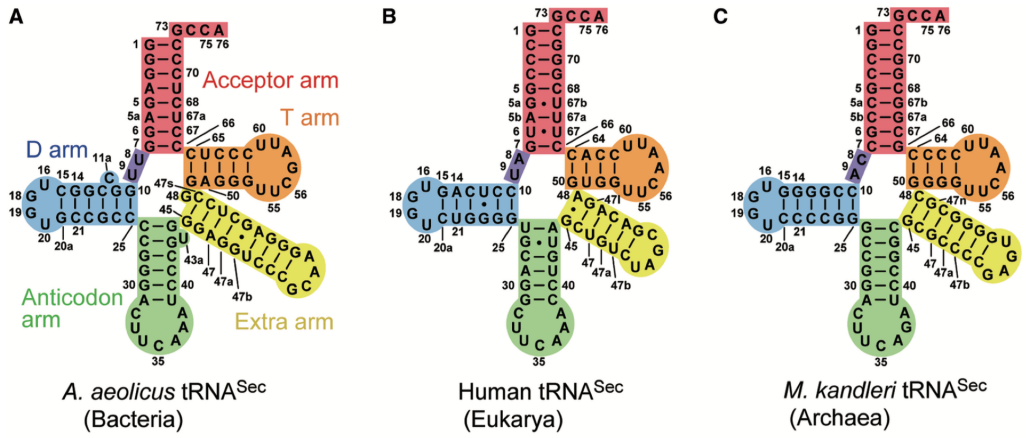


Figure 1.1: Cloverleaf secondary structure of $tRNA^{Sec}$ in the three domains of life. Adapted from [Itoh et al., 2013].

1.2.2 Sec synthesis pathway

Sec synthesis occurs on its own tRNA. $tRNA^{Sec}$ lacks an aminoacyl-tRNA synthetase; it is, instead, misacylated with serine (Ser) by the conventional seryl-tRNA synthetase (SerRS), followed by the conversion of $Ser\text{-}tRNA^{Sec}$ to $Sec\text{-}tRNA^{Sec}$. In bacteria, the selenocysteine synthase (SelA) directly converts Ser to Sec. Archaea and eukaryotes use an intermediate step in which $Ser\text{-}tRNA^{Sec}$ is phosphorylated by the O-phosphoseryl-$tRNA^{Sec}$ kinase (PSTK) to give $Sep\text{-}tRNA^{Sec}$, the substrate of the eukaryotic/archaeal selenocysteine synthase (SecS or SepSecS) in the final enzymatic step (figure 1.2). SelA and SecS are type-I pyridoxal phosphate (PLP)-dependent enzymes that use selenophosphate as the activated selenium donor, which is, in turn, synthesized by selenophosphate synthetase (SPS, SelD in bacteria) from selenite and ATP [Glass et al., 1993].

Since both $tRNA^{Sec}$ and $tRNA^{Ser}$ are charged with serine, Sec synthesis systems must strictly discriminate $Ser\text{-}tRNA^{Sec}$ from $Ser\text{-}tRNA^{Ser}$. In archaea, the specific interaction between the unique $tRNA^{Sec}$ D arm and the PSTK C-terminal domain accounts for the strict tRNASec selectivity by the PSTK [Chiba et al., 2010]. Human SecS specifically recognises $tRNA^{Sec}$ through interaction with the residue 73, the discriminator base, but $Sep\text{-}tRNA^{Sec}$ is the obligate substrate of SecS, and not $Ser\text{-}tRNA^{Ser}$, because the phosphoryl is required for proper posi-

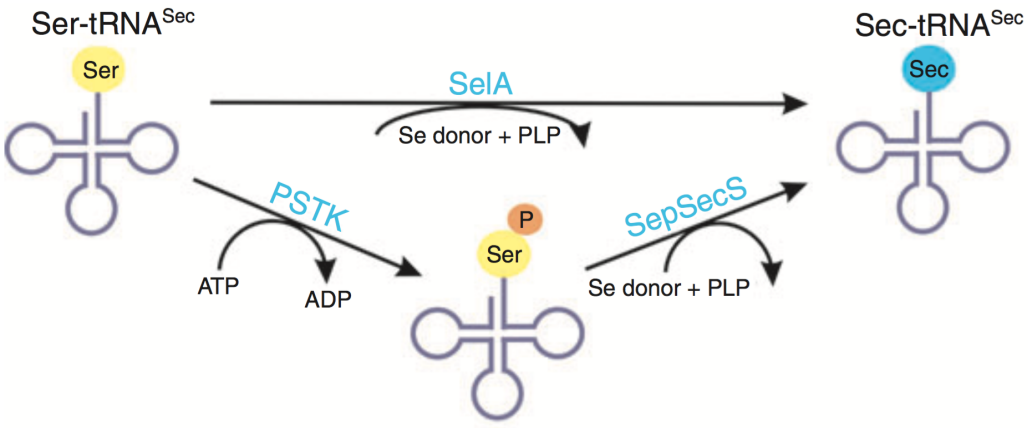


Figure 1.2: tRNA-dependent transformations leading to selenocysteine in bacteria (top) and eukaryotes/archaea (bottom). Adapted from [Ambrogelly et al., 2007].

tioning [Palioura et al., 2009]. SelA in bacteria, instead, has a different substrate than SecS (Ser-tRNA^{Sec} vs. O-phosphoseryl-tRNA^{Sec}) and discriminates Ser-tRNA^{Sec} from Ser-tRNA^{Ser} by the tRNA^{Sec} specific D arm through its N-terminal domain [Itoh et al., 2013]. Hence SecS and SelA functional forms are different, and the two enzymes were regarded as the result of convergent evolution of two independent Sec synthesis systems [Itoh et al., 2013].

1.2.3 Sec insertion: stop making sense

Insertion of Sec occurs during translation in response to an in-frame UGA codon. UGA normally terminates translation, but in selenoprotein mRNAs, a complex molecular mechanism prevents premature termination and dictates recoding of UGA as Sec. The recoding mechanism requires protein *trans*-acting factors, Sec-tRNA^{Sec}, and a *cis*-acting RNA stem loop called the Selenocysteine insertion sequence (SECIS) element, present in all selenoprotein mRNAs. The protein factors include SBP2 and eEFSec in eukaryotes, and SelB in bacteria. In selenoprotein mRNAs, the Sec-tRNA^{Sec} translates the UGA codon as Sec in response to the SECIS element.

In eukaryotes, the SECIS binding protein 2 (SBP2) was identified as a protein that cross-linked the selenoprotein GPx4 3'UTR [Lesoon et al., 1997]. Sub-

sequently, it was shown that SBP2 was required for Sec insertion and that it bound the SECIS element with high affinity and specificity [Copeland et al., 2000]. Its C-terminal region contains an RNA-binding domain (RBD), which belongs to the L7Ae RNA-binding protein family, known to interact with kink-turn motifs [Fletcher et al., 2001]. The eukaryotic SECIS has a characteristic kink-turn motif (see section 1.2.4). SBP2 is also stably associated with the ribosome, and interacts with the eukaryotic Sec-specific elongation factor (eEFSec). eEFSec protein sequence has strong similarity to the canonical elongation factor eEF1A [Fagegaltier et al., 2000, Tujebajeva et al., 2000], but it contains a unique C-terminal domain that is proposed to be involved in the interaction with SBP2 and $tRNA^{Sec}$ [Gonzalez-Flores et al., 2012]. It binds to Sec-$tRNA^{Sec}$ but not its precursor Ser-$tRNA^{Sec}$ [Fagegaltier et al., 2000, Tujebajeva et al., 2000]. eEFSec delivers Sec-$tRNA^{Sec}$ to the A site of the ribosome and facilitates Sec incorporation (figure 1.3). Other SECIS-binding proteins are proposed to play regulatory roles in selenoprotein synthesis (ribosomal protein L30, eIF4a3 and nucleolin), but the core factors ($tRNA^{Sec}$, SECIS, SBP2 and eEFSec) are known to be required and sufficient for Sec incorporation in vitro [Gupta et al., 2013].

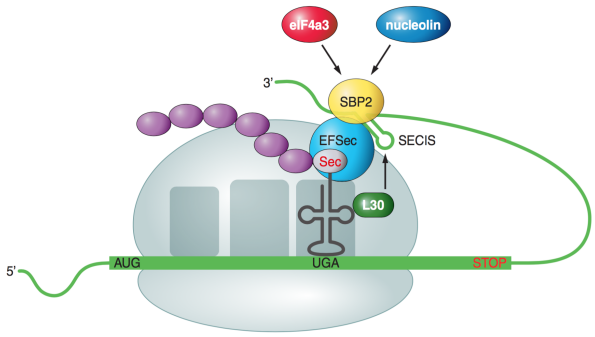


Figure 1.3: Mechanism for Sec insertion in eukaryotes. From [Labunskyy et al., 2014].

In bacteria, the SECIS binding activity and the Sec-specific elongation are carried out by a the same protein, SelB [Forchhammer et al., 1990] and [Baron et al., 1993] (figure 1.4). SelB is homologous to the canonical elongation factor EF-tu in its N-terminal part, and it contains a C-terminal extension responsible for

SECIS binding activity [Kromayer et al., 1996].

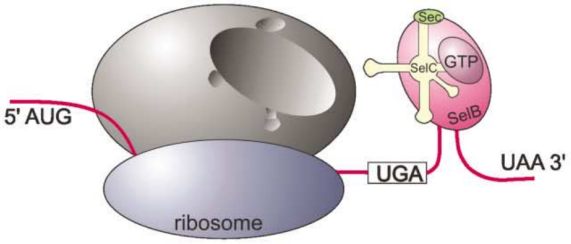


Figure 1.4: Mechanism for Sec insertion in bacteria. Adapted from [Birringer et al., 2002].

The homologous of SelB in archaea was identified and characterized [Rother et al., 2000]. In contrast to SelB, no binding to SECIS was found. The protein lacks the C-terminal domain, which is responsible for SECIS binding in bacteria. It was speculated that in archaea the functions of bacterial SelB are distributed over at least two different proteins, like in eukaryotes. To date, the SECIS binding protein in archaea is not known.

1.2.4 SECIS elements

The main signal for Sec specification is a *cis*-acting RNA stem-loop structure present in selenoprotein mRNAs, the SECIS element. The term SECIS was first used in [Berry et al., 1991] and stands for SElenoCysteine Insertion Sequence. All selenoprotein transcripts have a SECIS, but its sequence, structure and location within the mRNA is not conserved across the three domains of life (figure 1.5). Here we use the terms eSECIS, bSECIS and aSECIS to designate eukaryotic, bacterial and archaeal SECIS respectively.

The eukaryotic SECIS is a hairpin-loop structure formed by two stems separated by an internal loop, a GA Quartet structure, and an apical loop. The GA Quartet is located at the base of stem II and is composed of four non-Watson-Crick pairs, including two tandem GA/AG base pairs, which are characteristic of kink-turn motifs [Latrèche et al., 2009]. The GA Quartet is the main functional element of the eSECIS and is required for interaction with SBP2. The apical region

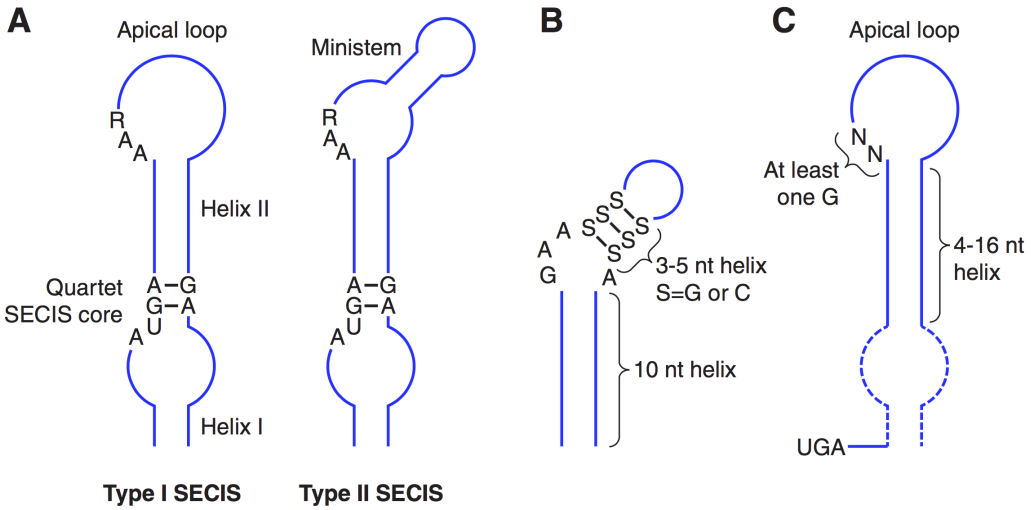


Figure 1.5: Scheme of the consensus secondary structures of SECIS elements in eukaryotes (A), archaea (B) and bacteria (C). From [Labunskyy et al., 2014].

contains a conserved AAR motif of unknown function. Two types of eSECIS exist, type 1 and 2, distinguished by the presence of an additional stem in the apical loop, which is present in type 2 and absent in the type 1 (figure 1.5A).

The bacterial SECIS [Hüttenhofer and Böck, 1998, Krol, 2002] element resides within the coding sequence, located immediately downstream the UGA-Sec codon, thus having both coding potential and base-pairing ability. They were characterized in *Escherichia coli*. Analysis of *E. coli* Sec-containing formate dehydrogenases established that the UGA codon itself was not enough to discriminate between Sec insertion and translation termination [Baron et al., 1990]. Structure based studies proposed two domains for an essential RNA stem-loop structure: the first domain comprises the UGA codon and would prevent the release factor 2 from binding to UGA; the apical loop constitutes the second domain, and includes critical elements for SelB recognition [Krol, 2002]. SelB interacts with the bSECIS through its unique C-terminal region [Baron et al., 1993]. Looking across species, bSECIS exhibit very poor sequence identity, and also high amount of structural variation. No primary sequence conservation was observed, other than a guanosine (G) residue in either of the first two positions of the apical loop [Zhang and Gladyshev, 2005] (figure 1.5B).

The archaeal SECIS [Krol, 2002] is located, like in eukaryotes, in the 3' UTR. In a single documented exception, it was found in the 5' UTR [Wilting et al., 1997]. The aSECIS are characterized by two stems separated by an invariant internal loop. The lower stem is GC rich and encompasses ten pairs; the internal loop consists of a GAA trinucleotide at the 5' and a single adenine at the 3'; and the upper stem is composed of three GC pairs. The apical loop has variable length and may contain additional pairings (figure 1.5C). No protein is known to interact with the archaeal SECIS. The SBP2 counterpart has never been observed in archaea, while the archaeal EFsec was shown not bind archaeal SECIS elements [Stock and Rother, 2009]. The question of how the SECIS and the Sec UGA site communicate remains open.

1.3 Selenoprotein families

The key common feature of all selenoproteins is the presence of a Sec residue. With few exceptions, Sec is located in the catalytic site. All functionally characterized selenoproteins perform redox reactions to serve diverse biological roles. Humans have 25 selenoproteins classified in 17 protein families [Kryukov et al., 2003].

Many selenoproteins are involved in the antioxidant defence. Glutathione peroxidases (GPx) reduce hydrogen peroxide (H2O2) using glutathione (GSH) as electron donor. Some GPxs, instead, have specificity for thioredoxin or other thiol oxidoreductases [Labunskyy et al., 2014]. Selenoprotein GPxs are widespread in the three domains of life [Toppo et al., 2008]. Thioredoxin reductases (TR) are large flavoprotein oxidoreductases that reduce thioredoxin at expenses of NADPH. The thioredoxin system is the major disulfide reduction system [Arnér and Holmgren, 2000]. Methionine sulfoxide reductases (Msr) catalyse the reduction of methionine sulfoxides to methionine. Two distinct enzyme families have evolved as a repair mechanism to reverse oxidative damage by ROS (reactive oxygen species). MsrA is specific for the S-form of methionine sulfoxide, whereas MsrB can only reduce the R-form [Kim and Gladyshev, 2007]. MsrA was found as a selenoprotein in bacteria, algae, and invertebrate animals, but not in vertebrates, while Sec-containing MsrB was found only in eukaryotes, including mammals, some invertebrate animals, and *Aureococcus anophagefferens* [Kim, 2013]. Peroxire-

doxins (Prx) are thiol/selenol peroxidases, like GPxs. Selenoprotein Prx were found in green algae [Dayer et al., 2008] and prokaryotes [Zhang and Gladyshev, 2008]. Glutaredoxins (Grx) are GSH-dependent reductases with thioredoxin-like fold that catalyze the reversible reduction of protein disulfides at expenses of NADHP [Lillig et al., 2008]. Grx were found as selenoproteins in prokaryotes [Zhang and Gladyshev, 2008], and as a conserved domain in Sec-containing TGR (Trx and GSSG reductase).

Related to the oxidative stress, some selenoproteins are involved in the protein folding control. The 15-kDA selenoprotein (Sel15) and Selenoprotein M (SelM) are ER (endoplasmic reticulum)-resident selenoproteins with the thioredoxin-like fold proposed to have a role in protein folding control [Gromer et al., 2005]. Fish Sel15-like (Fep15) is a distant homologue of Sel15 found only in fishes. Selenoproteins K (SelK) and S (SelS) have no sequence similarity but are considered related proteins based on their topology [Shchedrina et al., 2011]. Both have been implicated in ER-associated degradation (ERAD) of misfolded proteins [Labunskyy et al., 2014].

Selenoproteins involved in electron transport and energy-yielding pathways are also commonly found in bacteria. Formate dehydrogenases (FDH) catalyze the reversible oxidation of CO2 to formate and is involved in energy metabolism [Stock and Rother, 2009]. The alfa subunit of FDH is the most widely distributed selenoprotein in bacteria. Glycine reductase complex selenoproteins (GrdB and GrdA) are part of the Glycine Reductase system, key in the acetate formation via glycine [Stock and Rother, 2009].

Other functions are also known to be performed by Sec-containing proteins. Selenoprotein P (SelP) is a secreted selenoprotein abundantly found in plasma. A unique feature of SelP is the presence of multiple Sec residues (10 in human). Two SECIS elements at the 3'UTR of the SelP gene direct the readthrough of the multiple UGA codons. The protein was proposed to function as a Se supplier to peripheral tissues [Saito and Takahashi, 2002]. Iodothyronine deiodinases (DI) are involved in the regulation of thyroid hormone activity by reductive deodination in mammals. Homologues were also found in single cell eukaryotes and bacteria but their function is not known. Selenophosphate synthetase 2 (SPS2, SelD in bacteria) catalyze the synthesis of the Se donor selenophosphate, necessary for Sec synthesis [Xu et al., 2007]. Sec-containing SPS2 are found widespread in

the three domains of life [Mariotti et al., 2015]. Selenoprotein N (SelN) is a ER-resident transmembrane glycoprotein. Mutations in human SelN have been associated with SEPN1-related myopathies [Arbogast and Ferreiro, 2010], although the contribution of SelN to normal muscle is not known.

The function of several vertebrate selenoproteins, including some human proteins, is still unknown. Selenoprotein I (SelI) is a transmembrane protein only found in vertebrates. It contains a CDP-alcohol phosphatidyltransferase known to be involved in the synthesis of phospholipids. The Sec residue is found in a unique C-terminal domain with unknown function. The Rdx family of selenoproteins comprise SelW, SelT, SelH and SelV. They possess a thioredoxin-like fold, with a conserved motif CXXU, and are predicted to be thiol-based oxidoreductases, but the exact function is not known [Dikiy et al., 2007]. An the less studied SelO, SelJ [Castellano et al., 2005], SelL, SelU.

Most of the currently known selenoproteins are listed in figure 1.6.

1.4 Selenoprotein identification methods

The first selenoproteins were identified in the 70's as Se-containing proteins [Flohe et al., 1973, Andreesen and Ljungdahl, 1973], and it was shown that the Se moiety was essential for the protein activity and that it corresponded to a Sec residue [Cone et al., 1976]. Identification of selenoproteins was initially based on experimental approaches. The presence of Se, incorporated in the form of Sec, was identified through analysis of proteins by mass spectrometry and detection of radioactive ^{75}Se [Behne et al., 1990, Ballihaut et al., 2007]. Yet, the codon for Sec insertion was not identified until ten years later, when genetic analysis of the *fdhF* locus in E. coli revealed an open reading frame (ORF) with an in-frame UGA codon. The UGA codon, until then only considered to terminate translation, corresponded to the position of the Sec residue in the protein [Zinoni et al., 1986]. The mechanism for Sec insertion was elucidated first in bacteria, showing that recoding of the UGA-Sec codon required a *cis*-acting SECIS element in the mRNA [Baron et al., 1990, Berg et al., 1991].

More recently, with the advances in sequencing technologies and the adoption of bioinformatics, the genome-wide identification of selenoproteins could be accomplished by means of computational and comparative genomics.

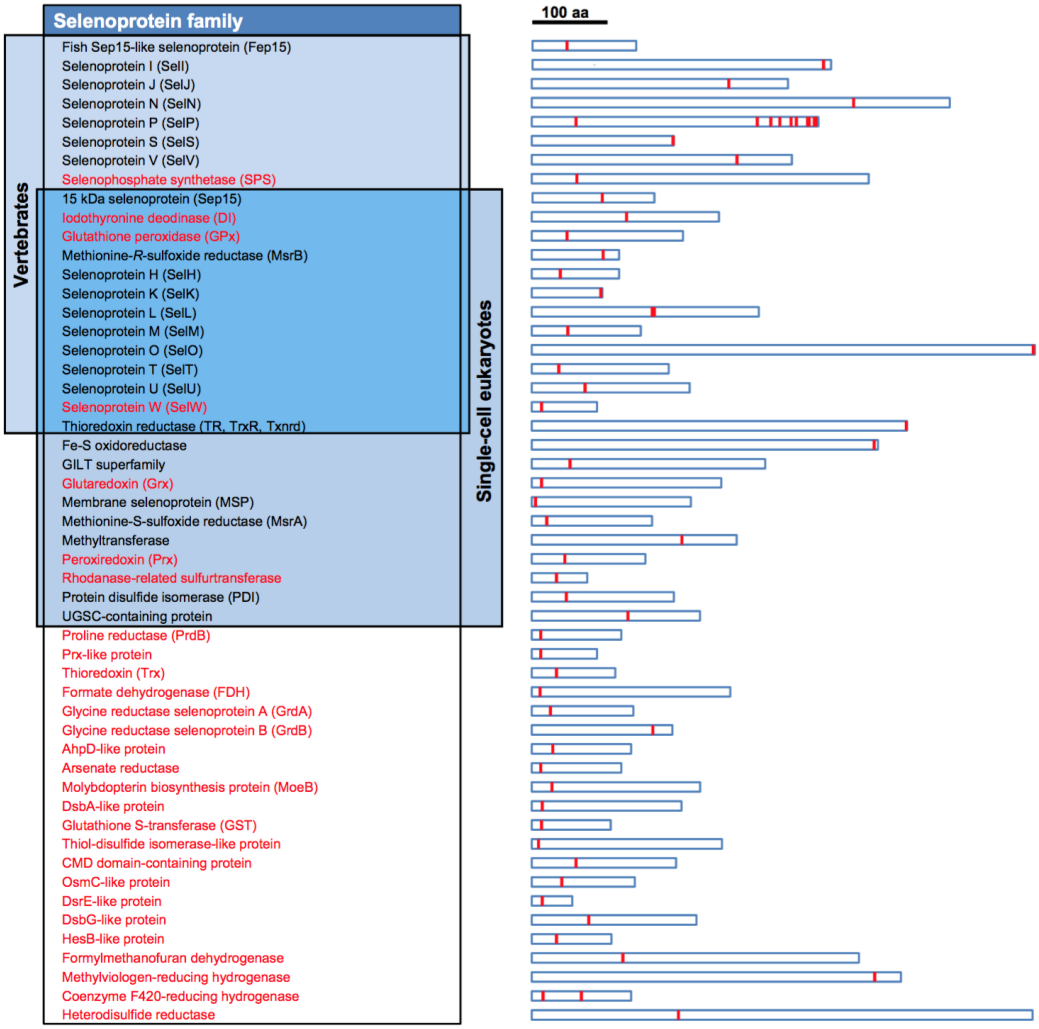


Figure 1.6: Selenoprotein families. The shaded boxes group selenoprotein families found in the indicated lineages. Prokaryotic selenoproteins are highlighted in red. The relative size of each selenoprotein is shown on the right, with the Sec residue in red. From [Labunskyy et al., 2014].

Because of the non-canonical usage of the UGA codon–normally a stop codon–in selenoproteins, their identification in genomes poses a challenge. Only a tiny fraction of the UGA codons in the coding sequences of an organism are translated as Sec residues, and all standard gene annotation programs simply ignore Sec. Consequently, selenoproteins are usually misannotated in protein databases and

genome projects. The identification of selenoprotein genes rely on their unique genomic features. Essentially, all methods are based on the occurrence of an in-frame UGA codon in the gene ORF, and in the presence of a properly located SECIS element.

A first approach was the identification of eukaryotic SECIS associated with a UGA-containing ORF in the upstream region. This strategy was first applied to human expressed sequence tag (EST) database by two different groups. Collaborators from the group of Vadim Gladyshev developed the first method to identify eukaryotic SECIS, called SECISearch [Kryukov et al., 1999]. The method was based on patterns and RNAfold [Lorenz et al., 2011]. An analogous strategy was developed by collaborators from the group of Alain Krol [Lescure et al., 1999]. In this case the search for SECIS elements was based on the pattern-based RNAMOT program. In both works new selenoproteins were successfully identified and validated experimentally. Another strategy was developed in our group, that allowed scanning complete genomes. Unlike the previous methods, that were applied on ESTs, which contain spliced coding DNA (cDNA) only, this new strategy was applied to the then-newly published complete *Drosophila melanogaster* genome [Castellano et al., 2001]. The method was based on the genome-wide prediction of SECIS and UGA-containing ORFs. By crossing the two sets of predictions, the very high false positive rate of SECISearch could be alleviated. In order to identify UGA-containig ORFs, a modified version of the program geneid [Guigó et al., 1992], a de novo gene predictor, was used. In practice, geneid was modified to allow genes with good scoring potential, to include a single in-frame UGA codon. The approach succeeded for the characterization of the complete selenoproteome of *D. melanogaster*, composed by three selenoproteins.

Selenoprotein families comprise both Sec- and Cys-containing protein members. In Cys-containing homologues, a Cys residue (UGU or UGC codon) replaces Sec. Thus the alignment of members from the same protein family display Sec/Sec or Sec/Cys pairs. Using this criteria, a SECIS-independent approach based on comparative genomics was able to identify new selenoproteins. The strategy was based on the correlated analysis of the genome of two closely related species. The method was used in [Castellano et al., 2004] and [Castellano et al., 2005] to obtain predictions of standard genes (with geneid [Guigó et al., 1992])

and selenoprotein candidates (with the modified version of geneid [Castellano et al., 2001]) in two vertebrates. Then, using blastp [Altschul et al., 1997] to obtain the pairwise alignment of all predictions both intra- and inter-species, to identify putative homologues. From the blast alignments, Sec/Sec (putative selenoproteins in both species) and Sec/Cys (putative selenoprotein in one species and a Cys homologue in the other) pairs were selected. Finally, after further filtering steps, the strongest candidates were searched for a SECIS element in the region downstream. New vertebrate selenoproteins were identified and experimentally validated in both studies [Castellano et al., 2004, Castellano et al., 2005]. Similar strategies to were applied to identify eukaryotic selenoproteins [Novoselov et al., 2002, Taskov et al., 2005, Novoselov et al., 2006, Lobanov et al., 2006, Shchedrina et al., 2007].

The same key concepts used for eukaryotic selenoprotein gene identification were also applied to prokaryotes, correcting for the different structure and location of SECIS elements. A program was developed using the same structure as the original SECISearch: fixed patterns (built from known archaeal selenoproteins in this case) were used to scan nucleotide sequences, filtering by thermodynamic stability and structural criteria to generate aSECIS candidates. This program, together with a SECIS-independent method based on the Sec/Sec and Sec/Cys criteria were used to characterize prokaryotic selenoproteomes [Kryukov and Gladyshev, 2004]. An analogous method called bSECISearch was developed for the identification of selenoproteins in bacteria [Zhang and Gladyshev, 2005]. First a structural consensus model for bSECIS was build from previously known bacterial selenoproteins. Based on that model, a three-module program using RNAfold (v1.4), the segment-based alignment program DIALIGN [Morgenstern et al., 1996], and position specific scoring matrices (PSSM) matrices, were used to scan the nucleotide sequence and generate a set of candidate ORFs that contained a TGA-bSECIS. Tblastn and blastx [Altschul et al., 1997] were applied to identify sequence conservation in the TGA-flanking region. The method was applied to characterize prokaryotic selenoproteomes [Zhang and Gladyshev, 2005] and to validate bSECIS in novel selenoproteins identified in environmental metagenomic sequences from the Global Ocean Sampling (GOS) project [Zhang and Gladyshev, 2008].

Programs for a fully automated identification of selenoproteins have recently

been developed in our group. Selenoprofiles [Mariotti and Guigó, 2010] is a profile-based annotation pipeline that and produces accurate predictions of selenoprotein genes, and cysteine homologues, with null or very little human intervention. It is a fast method based on sequence homology that relies on a set of manually curated profiles of known selenoproteins. Seblastian [Mariotti et al., 2013] is a pipeline that couples the SECIS search method SECISearch3, with the analysis of the upstream sequence for selenoprotein coding potential. It uses blastx [Altschul et al., 1997] against a protein database and is able to identify known and novel eukaryotic selenoproteins. Seblastian is available through a web server.

1.5 Distribution of selenoproteins

Selenoproteins are present in the three domains of life, but not in all organisms. The number of selenoproteins encoded in a genome ranges from 1 to 56 in *Aureococcus anophagefferens* [Gobler et al., 2011]. Many organisms, however, do not posses selenoproteins nor the ability to synthesize Sec. In these organisms, Cys-based non-selenoprotein orthologues are often found.

Among eukaryotes, selenoproteins show a scattered distribution. Fungi and land plants lack selenoproteins completely [Lobanov et al., 2009], while in metazoans they are widely present. Selenoproteins in vertebrates are tightly conserved. The ancestral vertebrate selenoproteome is composed of 28 selenoproteins [Mariotti et al., 2012], the 25 selenoproteins present in human and other mammals [Kryukov et al., 2003], plus other selenoproteins identified in bony fishes: SelU [Castellano et al., 2004], SelJ [Castellano et al., 2005], Fep15 [Novoselov et al., 2006] and SelL [Shchedrina et al., 2007]. Bony fishes have larger selenoproteomes than other vertebrates [Mariotti et al., 2012]. Nematodes, instead, have a minimal selenoproteome. *Caenorhabditis elegans* has a single selenoprotein gene [Taskov et al., 2005]. The thioredoxin reductase TrxR1 in *C. elegans* was shown to be dispensable [Stenvall et al., 2011], nonetheless this organism conserves a fully functional Sec machinery for recoding a single UGA codon in its proteome. Some plant parasitic nematodes were recently shown to have lost selenoproteins [Otero et al., 2014]. Selenoprotein extinctions in animals were first identified among insects [Chapple and Guigó, 2008, Lobanov et al., 2008]. The

species from the insect orders *Hymenoptera*, *Coleoptera* and *Hymenoptera* were described to have lost all selenoproteins in independent evolutionary events at the root of their lineages, while a more recent extinction was identified in *Diptera*, that of *Drosophila willistoni*. An additional Sec loss in insects was identified in the pea aphid [International Aphid Genomics Consortium, 2010]. Non-insect arthropods like the crustacean *Daphnia pulex*, have larger selenoproteomes (see section 3.6 in this thesis). Protists display a scattered distribution of selenoproteins [Mariotti et al., 2015].

The Sec trait in bacteria shows a highly dynamic evolution, with very much scattered distribution across lineages [Mariotti et al., 2015, Peng et al., 2016]. An estimated $\sim$25% of the sequenced genomes use Sec. Several studies have attempted to characterize bacterial selenoproteomes, analyzing both completely sequenced genomes and metagenomic sequences. More than 35 Sec-containing proteins families have been identified in bacteria, mostly through computational analysis. The most abundant selenoproteins are involved in energy metabolism, acetate formation, antioxidant defence and synthesis of selenophosphate [Stock and Rother, 2009]. Selenoprotein-rich phyla had been identified, like *Deltaproteobacteria*, *Firmicutes/Clostridia*, and *Synergistetes* [Zhang et al., 2006, Peng et al., 2016]; the largest bacterial selenoproteome to date is that of *Syntrophobacter fumaroxidans* with 39 selenoprotein genes [Peng et al., 2016]. The evolution and ecology of the Sec utilization trait in bacteria is largely unknown. Several horizontal gene transfer (HGT) events of the entire Sec utilization pathway had been described [Zhang et al., 2006, Peng et al., 2016]. It is known that HGT events can contribute to the evolution of biological processes, including the Sec trait [Romero et al., 2005].

The distribution of the Sec utilization trait in archaea is very restricted. Only 12% of the sequenced genomes encode selenoproteins [Mariotti et al., 2016], and these are limited to *Methanococcales* and a single *Methanopyrus* genome. All *Methanococcales* (20 sequenced genomes) contain selenoproteins, with eight archaeal selenoprotein families identified [Kryukov and Gladyshev, 2004, Stock and Rother, 2009]. This thesis includes the study of the selenoproteome of *Lokiarchaeota*, a recently described archaeal lineage (section 3.4, [Mariotti et al., 2016]).

Chapter 2

METHODS

2.1 Secmarker

One of the challenges I faced during my PhD was the identification of tRNA^{Sec} genes. There was no specific tool for it, and general tRNA prediction programs suffer from high false positive and false negative rates. We developed a computational tool for tRNA^{Sec} identification, and benchmarked the accuracy of tRNA^{Sec} prediction for the first time.

2.1.1 Prediction of tRNA^{Sec}

Identification of tRNA genes in genomes is carried out with tRNA detection programs. The widely used programs tRNAscan-SE [Lowe and Eddy, 1997] and aragorn [Laslett and Canback, 2004] can identify the Sec-specific tRNA (tRNA^{Sec}). Their accuracy, however, is far from optimal, mainly because the tRNA^{Sec} structure is different than that of canonical tRNAs. Since there was no specific computational tool for prediction of tRNA^{Sec}, I decided to build my own tool. Infernal [Nawrocki and Eddy, 2013] is a package for searching DNA sequence databases for RNA structure and sequence similarities based on covariance models [Eddy and Durbin, 1994]. It can be used to build a covariance model from a multiple sequence alignment with structure annotation, and then scan nucleotide sequences for RNA structure analysis. I built three different models for the three tRNA^{Sec} structures known (bacteria, archaea and eukaryotes). The models are run

with cmsearch from Infernal, and several post-processing steps were implemented in a computational pipeline. In order to compare its accuracy with other methods, we build a benchmark set based on the assumption that $tRNA^{Sec}$ is only present in genomes with selenoproteins, while it is absent in selenoproteinless organisms. Since $tRNA^{Sec}$ is a genetic marker for the use of selenocysteine, we called the program Secmarker.

2.1.2 Secmarker manuscript

Didac Santesmasses, Marco Mariotti, Roderic Guigó. Computational identification of the selenocysteine tRNA ($tRNA^{Sec}$) in genomes. *Submitted*

Santesmasses D, Mariotti M, Guigó R. Computational identification of the selenocysteine tRNA (tRNASec) in genomes. PLOS Comput Biol. 2017 Feb 13;13(2):e1005383. DOI: 10.1371/journal.pcbi.1005383

2.2 *b*Seblastian: identification of bacterial selenoproteins

2.2.1 Bacterial SECIS elements

Selenoproteins contain the non-canonical amino acid selenocysteine (Sec). Sec is present in organisms from the three domains of life, and is inserted in selenoproteins during translation. Its incorporation, however, is independent of EF-Tu, the canonical elongation factor responsible for the insertion of the universal 20 amino acids. Instead, the Sec-specific elongation factor SelB in bacteria [Forchhammer et al., 1989], and EF-Sec in eukaryotes and archaea [Fagegaltier et al., 2000], is responsible for delivering Sec-tRNA^{Sec} to the ribosome. Sequestering of the Sec specific elongation factor is mediated by an RNA secondary structure, the SECIS (Sec insertion sequence) element, present in all selenoprotein mRNAs. The SECIS elements vary in terms of localization, primary sequence and structure between bacteria, archaea and eukaryotes [Krol, 2002].

The bacterial SECIS (bSECIS) is a hairpin loop located immediately downstream the UGA codon. The best characterized bSECIS are found in genes encoding formate dehydrogenases (fdh) in *Escherichia coli* (figure 2.1) [Berg et al., 1991, Hüttenhofer and Böck, 1998]. Putative bSECIS elements in other bacterial selenoproteins, however, showed no resemblance to each other or the *E. coli* counterparts. Analysis on the regions downstream the Sec-TGA from a number of bacterial selenoprotein sequences lead to a consensus structural model for bSECIS [Zhang and Gladyshev, 2005]. The method bSECIsearch [Zhang and Gladyshev, 2005] was developed to identify bSECIS sequences based on that model. bSECISearch runs RNAfold [Lorenz et al., 2011], the segment-based sequence alignment using DIALIGN [Morgenstern et al., 1996], and position specific scoring matrices (PSSM), for prediction and statistical evaluation of bSECIS candidates. bSECISearch was used to successfully identify known and new selenoproteins in the Sec-encoding bacterial genomes available at the time.

Covariance models (CM) is a long-established approach for RNA secondary structure and primary sequence analysis [Eddy and Durbin, 1994], known for its high sensitivity. The CM approach enabled by Infernal [Nawrocki and Eddy, 2013] was successfully used for the detection of eukaryotic SECIS, implemented

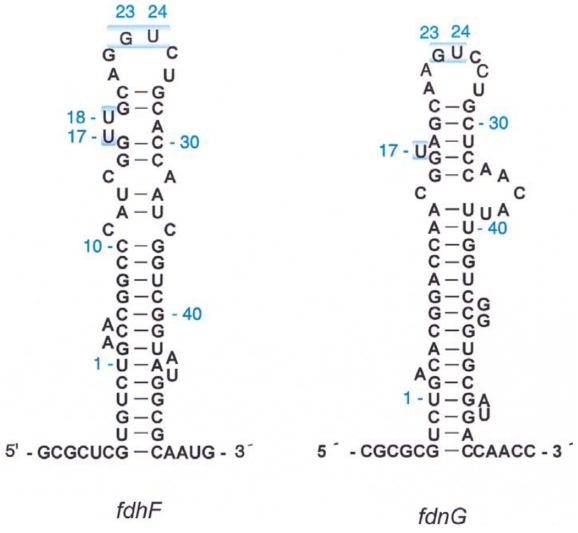


Figure 2.1: Secondary structures of bSECIS in the *E. coli* formate dehydrogenases mRNAs. From [Krol, 2002].

in SECISearch3 [Mariotti et al., 2013]. SECISearch3 was used to build Seblastian [Mariotti et al., 2013], a pipeline for the identification of eukaryotic selenoproteins based on detection of SECIS as first step. Here we developed the bacterial version of Seblastian, called bSeblastian. The new pipeline is based on the identification of bSECIS using covariance models.

2.2.2 Building a covariance model for bSECIS

We used the consensus bSECIS structure constraints [Zhang and Gladyshev, 2005] to build a bSECIS CM with Infernal 1.1 [Nawrocki and Eddy, 2013]. In order to build the CM, we used a set of bona fide bacterial selenoprotein genes, predicted with Selenoprofiles [Mariotti and Guigó, 2010] in bacterial genomes. First, their nucleotide sequences were run with RNALfold [Lorenz et al., 2011] to identify the bSECIS sequence in the region downstream the Sec-TGA. RNALfold computes locally stable RNA secondary structures with a maximal base pair span. The RNALfold output was parsed, and those structures that satisfied the bSECIS model were retrieved. The candidate bSECIS sequences were then aligned based

on their predicted secondary structure with cmalign from Infernal. The resulting alignment (figure 2.2) contained 401 sequences and was used to generate the covariance model with cmbuild.

The pipeline

bSeblastian is a new pipeline for the identification of bacterial selenoproteins based on detection of bSECIS as first step. The workflow of bSeblastian is very similar to the eukaryotic Seblastian. It inherits the main classes and functions written in the code, although many internal aspects of the pipeline were largely modified to account for the differences between the eukaryotic and bacterial SECIS. Here follows a summary of the pipeline.

2.2.3 bSECIS search phase

The search phase is implemented in the module bSECISearch2. In this module the bSECIS CM is used with cmsearch from Infernal [Nawrocki and Eddy, 2013] to scan the target nucleotide sequence. The options used for cmsearch are: a loose score threshold (-T 4) and `--max` (turn all heuristic filters off). The cmsearch output is parsed, refined and filtered. The infernal output includes the predicted secondary structure of the target sequence aligned to the model. The refinement refers to a procedure by which the non-canonical pairs are removed from the structure, or possibly additional canonical pairs are added by extending the existing stem. In the filtering step, the program excludes all those hits that do not satisfy the bSECIS constraints [Zhang and Gladyshev, 2005], i.e. a TGA triplet aligned to the TGA in the model, followed by (i) a 4-16 bp upper-stem and a 3-14 nt apical loop, (ii) at least one guanosine (G) among the first two nucleotides in the apical loop, (iii) a spacing of 16-37 nt between the TGA codon and the apical loop (figure 2.3). The hits with a TGA-bSECIS are analyzed further in the next step.

2.2.4 TGA-containing ORF

Since the bSECIS is located within the coding sequence, the remaining hits (those with a TGA-bSECIS) are analyzed for the occurrence of a TGA-containing open reading frame (ORF). The sequences flanking the TGA are translated using the

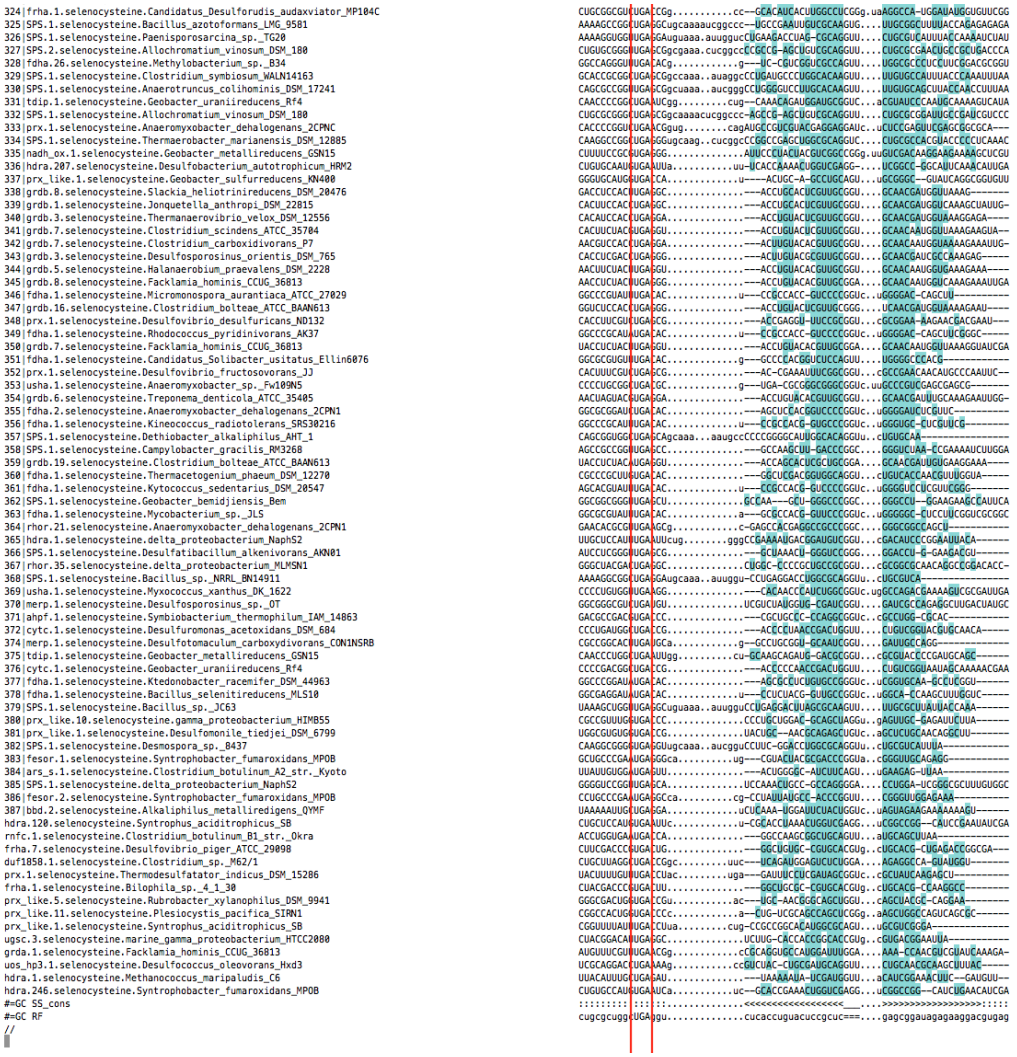


Figure 2.2: Structure-based alignment used to build the bSECIS covariance model. The Sec-UGA position is highlighted in red. The sequences were aligned using Infernal [Nawrocki and Eddy, 2013] and visualized with RALEE [Griffiths-Jones, 2005]. RALEE highlights (cyan) the nucleotides that are paired according to the consensus secondary structure (second line from the bottom, SS_cons) of the alignment, and that also respect the standard pairing rules. Only a fraction of the alignment is shown.

same frame as the TGA, until a stop codon (TGA, TAG or TAA) is found. The retrieved nucleotide sequences are analyzed for protein coding signatures with

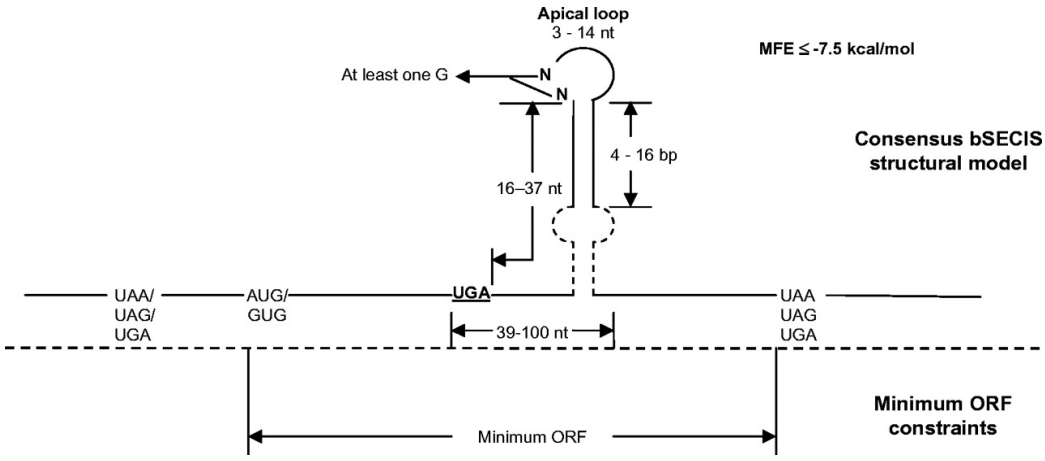


Figure 2.3: bSECIS consensus model constrains. From [Zhang and Gladyshev, 2005].

FragGeneScan [Rho et al., 2010], a de novo protein-coding gene predictor based on a Hidden Markov Model (HMM). FragGeneScan is able to identify those sequences with a high coding potential and, importantly, their optimal frame of translation. Based on the FragGeneScan predictions, the sequences with no coding potential, and those for which its optimal frame reads the TGA off-frame, are discarded. Prior to FragGeneScan, the TGA codons are changed to a TGT (Cys) codon, so they are not considered stop codons by the program. Given the low specificity of the bSECIS CM, a large fraction of candidate ORFs are filtered out in this step.

2.2.5 Sequence conservation of TGA-flanking regions

The candidate TGA-containing ORFs are then analyzed for sequence conservation using homology information. The amino acid sequences obtained from the candidate ORFs are run with blastp [Altschul et al., 1997] against a comprehensive protein database (e.g. NCBI nr). The ORFs that obtained hits spanning the putative Sec position (the TGA codon) are classified according the residue aligned to that position. Mainly two types of ORFs are considered, those that aligned the TGA codon with a U in the target sequence, and those where the TGA is aligned with a cysteine. The former correspond to known selenoproteins present

in the database, and the latter include ORFs with homology to known cysteine homologues. As the absolute majority of known selenoproteins possess cysteine homologues, bSeblastian is able to identify new selenoproteins.

2.2.6 Performance

bSeblastian includes the module bSECISearch2 for detection of bSECIS. The main difference with its predecessor, bSECISearch [Zhang and Gladyshev, 2005], is the use of covariance models, enabled by Infernal [Nawrocki and Eddy, 2013]. bSECISeacrh is based on secondary structure prediction with RNAfold v1.4. In order to test the performance of bSECISearch2, we prepared a test set consisting of 1121 bona fide selenoprotein nucleotide sequences predicted by Selenoprofiles in bacterial genomes, in which the Sec-TGA position was known. It is worth mentioning that Selenoprofiles does not use secondary structure in the identification of selenoprotein genes. We made sure that the sequences in the test set were not present in the alignment used to build the CM model. bSECISearch2 identified 89.4% of the TGA-bSECIS positions.

2.2.7 Conclusions

Here we developed bSeblastian, a pipeline for the identification of bacterial selenoproteins in nucleotide sequences. The program is based on the identification of bSECIS elements as first step. We built a covariance model for bSECIS based on the constraints of the bSECIS consensus model [Zhang and Gladyshev, 2005]. Covariance models had been used for the identification of eukaryotic SECIS, but they had been never applied before for the the bacterial SECIS. The method is able to identify both known and new selenoproteins.

2.3 SelenoDB 2.0

SelenoDB is a database that provides annotations for selenoprotein genes, proteins and SECIS elements. It contains predictions for multiple organisms. Version 1.0 was released in 2008 [Castellano2008]. It contains manually curated genes for 8 species. Although accurate, manual annotation of selenoprotein genes can not cope with the increasing number of sequenced genomes. We were contacted by Sergi Castellano, former member of our group and author of SelenoDB 1.0, because he was interested in including additional of species in SelenoDB. Since we could use Selenoprofiles [Mariotti and Guigó, 2010] for automated prediction of selenoprotein genes, we accepted the challenge. We decided to analyze the set of genomes then-available in ensembl (release 68; 59 genomes). I used Selenoprofiles to produce an annotation of known selenoprotein genes for new species included in SelenoDB 2.0. In addition, multiple isoforms for human selenoproteins, selected from Gencode (release 15), were also included into the database.

Romagné F, Santesmasses D, White L, Sarangi GK, Mariotti M, Hübler R, et al. SelenoDB 2.0: annotation of selenoprotein genes in animals and their genetic diversity in humans. Nucleic Acids Res. 2014 Jan;42(D1):D437–43. DOI: 10.1093/nar/gkt1045

Chapter 3

RESULTS

The results section is organized in two main areas. First, those works in which we applied our tools for the identification of selenoprotein genes are presented: the collaborations in two insect genome projects, the study of the Se-containing GPx in R. prolixus, and the characterization of the human selenomicrobiome. Second, I present those works in which we also applied our tools for the identification of selenoproteins, but the results are more related to the evolution of selenoproteins and the Sec machinery: my collaboration on the evolution of SPS, the selenoproteome of Lokiarchaeota, and finally, the selenoprotein extinctions in insects.

3.1 Genome projects

3.1.1 Bumble bees genome project

Bombus terrestris and *Bombus impatiens* are two bumblebees with primitive eusocial behaviour. Both are natural and agricultural pollinators, and widely utilized study species. Our group was involved in the genome project of these two hymenopteran species, coordinated by Kim Worley (Baylor College of Medicine). Francisco Camara from our group produced a genome-wide annotation of protein-coding genes. The project also included manual annotation for some specific gene families, including selenoproteins. Our contribution was the prediction of genes from selenoprotein families using Selenoprofiles [Mariotti and Guigó, 2010]. We then produced a manual annotation of those genes based on transcriptome and

comparative evidence, using the annotation editor Apollo [Lewis et al., 2002]. Like all known hymenopteran species [Chapple and Guigó, 2008, Lobanov et al., 2008], the two *Bombus* genomes did not encode any selenoprotein. All selenoprotein genes were converted to Cys homologues or lost.

Despite no UGA-Sec codons were found in the two *Bombus* genomes, the gene *SPS1* has an in-frame UGA codon. This had also been observed in *Apis mellifera* [Chapple and Guigó, 2008] and other hymenopterans [Mariotti et al., 2015]. The *Hymenoptera* SPS1 protein does not incorporate Sec: the gene lacks the SECIS element, and the factors for Sec synthesis, like $tRNA^{Sec}$, are missing from all genomes. The gene *SPS1* appeared by gene duplication of the Sec-encoding *SPS2* at the root of insects (and in other animal lineages independently, including humans), and it is never a selenoprotein. While other insects mutated the UGA-Sec to an arginine codon (e.g. CGC in *Drosophila melanogaster*), *SPS1* in *A. mellifera* maintained the UGA codon after the duplication. A stop codon readthrough event is thought to occur at this position, but the amino acid inserted remains elusive. The gene models for *SPS1* in the two *Bombus* genomes had to be modified manually in order to extend the coding sequence passed the TGA codon.

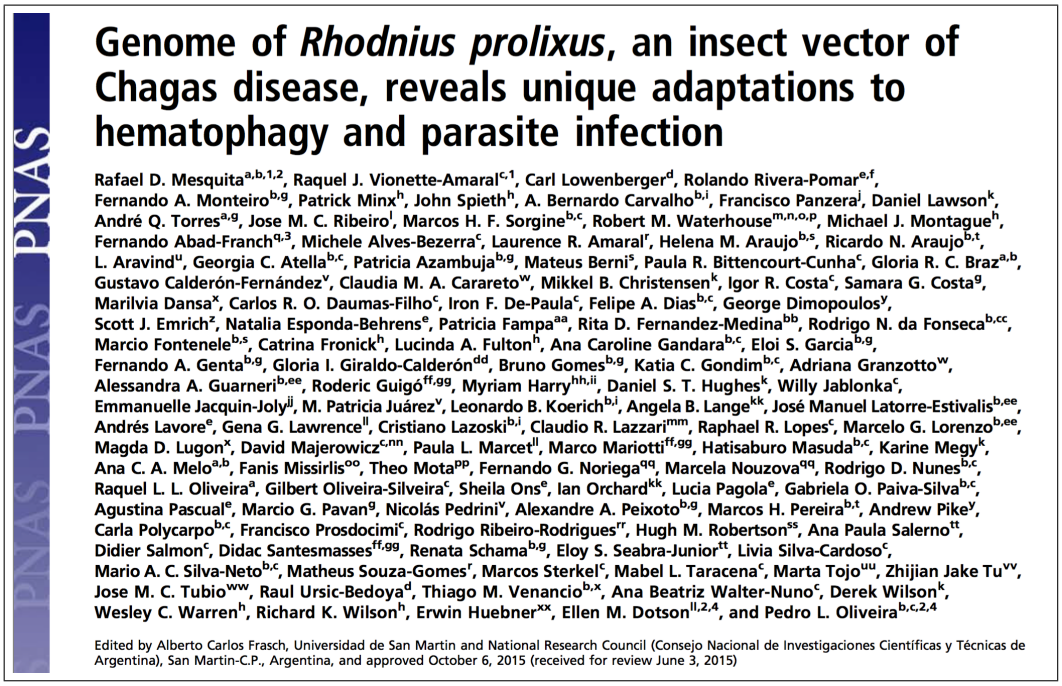
Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P, Elsik CG, et al. The genomes of two key bumblebee species with primitive eusocial organization. Genome Biol. 2015 Apr 24;16(1):76. DOI: 10.1186/s13059-015-0623-3

3.1.2 *Rhodnius prolixus* genome project

Rhodnius prolixus is a major insect vector of the Chagas disease, an illness caused by *Trypanosoma cruzi*. We were contacted by Carla Polycarpo (Universidade Federal do Rio de Janeiro, Brazil). tThey were working on the annotation of the *R. prolixus* genome, and they were interested using our tools to identify selenoproteins. Given our interest in arthropod selenoproteins we got involved in the project.

Our contribution to this work was based in the analysis of the genome of *R. prolixus* using Selenoprofiles [Mariotti and Guigó, 2010]. Two selenoproteins were identified: selenophosphate synthetase (SPS2), an enzyme involved in the synthesis of selenoproteins, providing the activated selenium donor for Sec synthesis, and a glutathione peroxidase (GPx). GPxs comprise a major antioxidant protein family. Other known insect GPx genes encoded cysteine-based enzymes (see section 3.2). Accordingly, the Sec machinery for selenoprotein synthesis was also found. No other Sec-containing genes were identified. We also analyzed the genome of other paraneopteran species, and identified several Sec-containing TR genes in this lineage. The phylogenetic analysis of the TR family in *Paraneoptera* revealed a recent Sec-to-Cys conversion in one of the two *R. prolixus* TR genes (see figure 3.2).

Mesquita RD, Vionette-Amaral RJ, Lowenberger C, Rivera-Pomar R, Monteiro FA, Minx P, et al. Genome of Rhodnius prolixus, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. Proc Natl Acad Sci. 2015 Dec 1;112(48):14936–41. DOI: 10.1073/pnas.1506226112



Genome of *Rhodnius prolixus*, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection

Rafael D. Mesquita[a,b,1,2], Raquel J. Vionette-Amaral[c,1], Carl Lowenberger[d], Rolando Rivera-Pomar[e,f], Fernando A. Monteiro[b,g], Patrick Minx[h], John Spieth[h], A. Bernardo Carvalho[b,i], Francisco Panzera[j], Daniel Lawson[k], André Q. Torres[a,g], Jose M. C. Ribeiro[l], Marcos H. F. Sorgine[b,c], Robert M. Waterhouse[m,n,o,p], Michael J. Montague[h], Fernando Abad-Franch[q,3], Michele Alves-Bezerra[c], Laurence R. Amaral[r], Helena M. Araujo[b,s], Ricardo N. Araujo[b,t], L. Aravind[u], Georgia C. Atella[b,c], Patricia Azambuja[b,g], Mateus Berni[s], Paula R. Bittencourt-Cunha[c], Gloria R. C. Braz[a,b], Gustavo Calderón-Fernández[v], Claudia M. A. Carareto[w], Mikkel B. Christensen[k], Igor R. Costa[c], Samara G. Costa[g], Marilvia Dansa[x], Carlos R. O. Daumas-Filho[c], Iron F. De-Paula[c], Felipe A. Dias[b,c], George Dimopoulos[y], Scott J. Emrich[z], Natalia Esponda-Behrens[e], Patricia Fampa[aa], Rita D. Fernandez-Medina[bb], Rodrigo N. da Fonseca[b,cc], Marcio Fontenele[b,s], Catrina Fronick[h], Lucinda A. Fulton[h], Ana Caroline Gandara[b,c], Eloi S. Garcia[b,g], Fernando A. Genta[b,g], Gloria I. Giraldo-Calderón[dd], Bruno Gomes[b,g], Katia C. Gondim[b,c], Adriana Granzotto[w], Alessandra A. Guarneri[b,ee], Roderic Guigó[ff,gg], Myriam Harry[hh,ii], Daniel S. T. Hughes[k], Willy Jablonka[c], Emmanuelle Jacquin-Joly[jj], M. Patricia Juárez[v], Leonardo B. Koerich[b,i], Angela B. Lange[kk], José Manuel Latorre-Estivalis[b,ee], Andrés Lavore[e], Gena G. Lawrence[ll], Cristiano Lazoski[b,i], Claudio R. Lazzari[mm], Raphael R. Lopes[c], Marcelo G. Lorenzo[b,ee], Magda D. Lugon[x], David Majerowicz[c,nn], Paula L. Marcet[ll], Marco Mariotti[ff,gg], Hatisaburo Masuda[b,c], Karine Megy[k], Ana C. A. Melo[a,b], Fanis Missirlis[oo], Theo Mota[pp], Fernando G. Noriega[qq], Marcela Nouzova[qq], Rodrigo D. Nunes[b,c], Raquel L. L. Oliveira[a], Gilbert Oliveira-Silveira[c], Sheila Ons[e], Ian Orchard[kk], Lucia Pagola[e], Gabriela O. Paiva-Silva[b,c], Agustina Pascual[e], Marcio G. Pavan[g], Nicolás Pedrini[v], Alexandre A. Peixoto[b,g], Marcos H. Pereira[b,t], Andrew Pike[y], Carla Polycarpo[b,c], Francisco Prosdocimi[c], Rodrigo Ribeiro-Rodrigues[rr], Hugh M. Robertson[ss], Ana Paula Salerno[tt], Didier Salmon[c], Didac Santesmasses[ff,gg], Renata Schama[b,g], Eloy S. Seabra-Junior[tt], Livia Silva-Cardoso[c], Mario A. C. Silva-Neto[b,c], Matheus Souza-Gomes[r], Marcos Sterkel[c], Mabel L. Taracena[c], Marta Tojo[uu], Zhijian Jake Tu[vv], Jose M. C. Tubio[ww], Raul Ursic-Bedoya[d], Thiago M. Venancio[b,x], Ana Beatriz Walter-Nuno[c], Derek Wilson[k], Wesley C. Warren[h], Richard K. Wilson[h], Erwin Huebner[xx], Ellen M. Dotson[ll,2,4], and Pedro L. Oliveira[b,c,2,4]

Edited by Alberto Carlos Frasch, Universidad de San Martin and National Research Council (Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina), San Martin-C.P., Argentina, and approved October 6, 2015 (received for review June 3, 2015)

Figure 3.1: *R. prolixus* genome project publication.

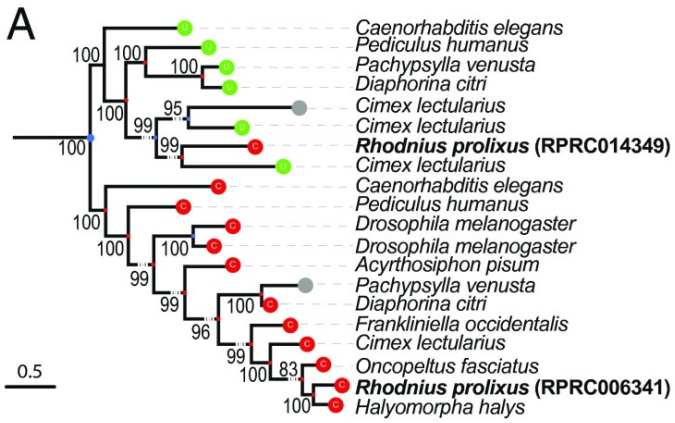


Figure 3.2: Phylogenetic relationship of thioredoxin reductases (TR) in nine paraneopterans. Sec position: green (U) selenocysteine, red (C) cysteine, and gray unknown/unaligned. Adapted from [Mesquita et al., 2015].

3.2 *Rhodnius prolixus* Se-dependent GPx

In this paper, the Sec-containing GPx was characterized. In most insect genomes sequenced to date, the known GPx genes encoded non-selenium, cysteine-based enzymes. It was previously hypothesized that Sec was replaced by cysteine in an ancient common ancestor of all insects. The discovery of a Sec-based GPx in an insect brought new insight into the evolution of this gene family. Our collaboration included the accurate prediction of the GPx selenoprotein gene with Selenoprofiles [Mariotti and Guigó, 2010] and its SECIS element with Seblastian [Mariotti et al., 2013].

Dias FA, Gandara ACP, Perdomo HD, Gonçalves RS, Oliveira CR, Oliveira RLL, et al. Identification of a selenium-dependent glutathione peroxidase in the blood-sucking insect Rhodnius prolixus. Insect Biochem Mol Biol. 2016 Feb;69:105–14. DOI: 10.1016/j.ibmb.2015.08.007

3.3 The human selenomicrobiome

3.3.1 Introduction

The human microbiota is composed by bacteria, archaea, viruses, and microbial eukaryotes that inhabit the human body. These microbial communities interact with the host and play fundamental roles in human health and disease. Selenium (Se) is an essential trace element in humans and in organisms throughout the tree of life, including the human microbiota. The main biological form of Se is the non-canonical amino acid selenocysteine (Sec), incorporated in selenoproteins. Sec is an analogous to cysteine (Cys) with Se replacing sulphur, and is generally found in the active site of oxidoreductase enzymes. Sec is inserted co-translationally through a peculiar mechanism in which a UGA codon (a stop in most organisms) is recoded to Sec [Labunskyy et al., 2014]. Specific RNA structures (SECISes, SEC Insertion Sequence) are present on selenoprotein transcripts and act as Sec recoding signals.

The human selenoproteome, the set of selenoproteins encoded in its genome, consists of 25 genes [Kryukov et al., 2003]. However, the number of selenoproteins present in our body is much larger than that. The microbes within a human body are estimated to outnumber human cells by an order of magnitude. Yet, the selenoproteins in the human microbiota have never been analyzed thus far. In addition to Sec, two other biological forms of Se exist in prokaryotes. Se is found in selenouridine (SeU), a modified nucleoside found in certain tRNAs, and it is used as cofactor to certain molybdenum-containing hydroxylases (Se-cofactor). The three Se utilization pathways use Se in the form of selenophosphate and are selenophosphate synthetase (SelD) dependent. Each of the three pathways are identified by the corresponding genetic markers [Lin et al., 2015]: *SelA*, *SelB*, *SelC* ($tRNA^{Sec}$) for selenocysteine; *ybbB* (2-selenouridine synthetase) for SeU; and *yqeB* and *yqeC*, of unknown function, for Se-cofactor. The distribution of the Se utilization traits in prokaryotes has been studied recently through analyses of large sets of completely sequenced genomes. The fraction of prokaryotes that use Se (those with *SelD* genes) was estimated to range between 26-38%, depending on the set of genomes analyzed. Among the Se utilization traits, Sec was found to be the most abundant (18-25%), followed by SeU (16-22%) and Se-cofactor

(6-8%) [Lin et al., 2015, Mariotti et al., 2015, Peng et al., 2016]. To date, the distribution of the three Se utilization traits in the human microbiome has never been studied.

Here, we characterized the composition and distribution of the selenoproteome of the human microbiota. We analyzed the metagenomic assemblies provided by the Human Microbiome Project (HMP), a NIH funded project implemented to provide a catalog of the microbial communities found in multiple body sites [The Human Microbiome Project Consortium, 2012]. We analyzed both the whole metagenomic assemblies (748) and the body-site specific assemblies (15) for the occurrence of selenoprotein genes, and of the genetic markers for the other Se utilization traits. Complete reference genome assemblies provided by HMP (1096) were also used. On the whole, more than 60 billion nucleotides (Gb) were analyzed.

3.3.2 Results

The selenoproteome of the human microbiota

We searched selenoprotein genes in 748 whole metagenomic assemblies (HMASM, Methods) from the HMP project. The assemblies correspond to samples obtained from 106 healthy adult individuals (46 females and 60 males) targeted in five major body sites: oral cavity, gastrointestinal tract, airways, skin and vagina. The 748 assemblies comprised a total of 46.5 Gb. We identified at least 10,726 selenoprotein genes distributed across all five body habitats. Surprisingly, more than 90% (9,665) of the selenoproteins were predicted in oral samples. The 415 oral samples corresponded to only 55% of all samples, and totalled 60% in sequence size (table 3.1). On the other hand, no selenoproteins were found in 57 of the 65 samples from vagina (urogenital tract). Those 65 samples (9%) accounted for less than 1% of the total sequence size. The vaginal habitat had been previously reported to harbor particularly simple communities, with low diversity both within samples and between subjects [The Human Microbiome Project Consortium, 2012]. In these bacterial communities, the presence of selenoproteins might be very restricted. In the lower gastrointestinal tract, represented by 147 stool samples (20% of all samples, 38% in sequence size), the number of selenoproteins was surprisingly low, 834 (less than 8%).

Table 3.1: Number of selenoproteins, samples and total nucleotides in each body site, and their corresponding percentages over the total.

	Selenoproteins (%)		Samples (%)		Nucleotides (Gb) (%)	
Urogenital tract	43	(0.4)	65	(8.7)	0.31	(0.7)
Skin	54	(0.5)	27	(3.6)	0.44	(0.9)
Airways	102	(1.0)	94	(12.6)	0.26	(0.6)
Stool	834	(7.8)	147	(19.7)	17.60	(37.8)
Oral	9665	(90.3)	415	(55.5)	27.90	(60.0)
TOTAL	10698	(100.0)	748	(100.0)	46.52	(100.0)

These observed differences may be due to actual variations in Sec usage across body sites, or due to disparities in genome size, or artifacts introduced in the generation of the meta-assemblies. To cancel any potential bias, we normalized the number of selenoproteins across samples using the RNA subunit of Ribonuclease P (RNase P) as a proxy for the number of genomes present in each assembly, since this gene appears precisely in a single copy in complete genome assemblies (Methods). The normalized occurrence of selenoproteins for each individual and body site is shown in figure 3.3. After normalization, we applied the Kruskal-Wallis rank sum test (kruskal.test from the R package "stats") to see whether there were differences in the distribution of selenoproteins across body sites. The test returned a very small p-value (2.2e-16). To identify which pairwise combination of samples were significantly different, we used the post-hoc Nemenyi test (posthoc.kruskal.nemenyi.test from R package "PMCMR"). The comparisons between oral and stool samples and between oral and urogenital samples obtained highly significant p-values (table S1). The distribution of the normalized counts of selenoproteins showed that the lower gastrointestinal tract (stool samples) and vagina appeared to be depleted of selenoproteins compared to the oral cavity (figure 3.4).

Known selenoprotein families in the human microbiome

The selenoproteins identified in the HMP assemblies belonged to 31 distinct selenoprotein families. The most abundant families were FDH, SelD, GrdB, PrdB and GdrA, described hereafter. All of them were found as selenoproteins in all the

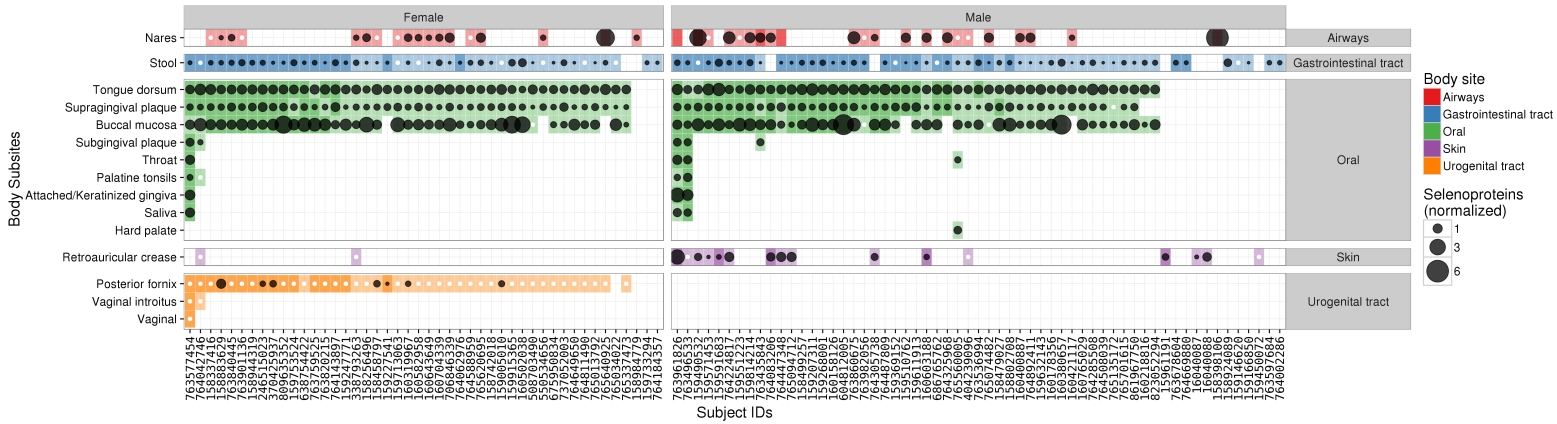


Figure 3.3: Selenoprotein abundance across individuals and body subsites. The subject individuals (columns) were sampled in different body subsites (rows). The size of the black dots is proportional to number of selenoproteins identified after normalization the number of genomes per sample (Methods). White dots correspond to samples with no selenoproteins. Some individuals were sampled multiple times on the same site, those cells are marked with a darker color, the average selenoprotein count is shown.

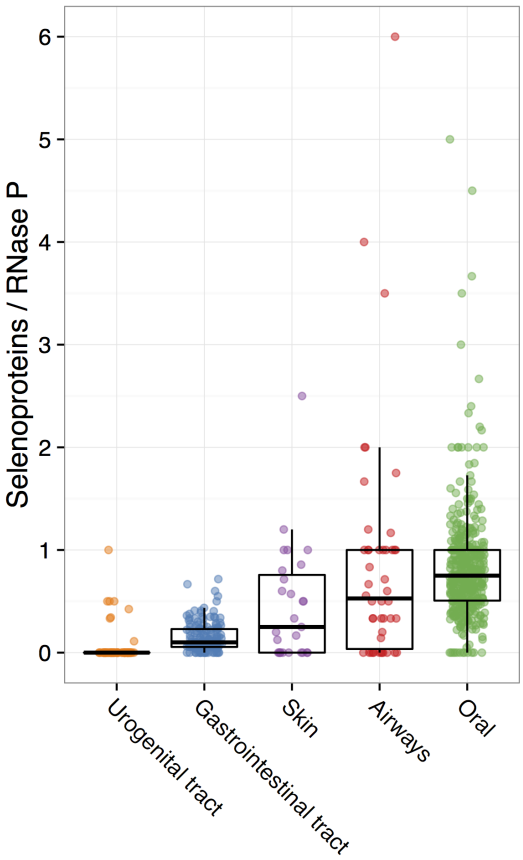


Figure 3.4: Abundance of selenoprotein genes across body sites. The number of selenoproteins identified in each genomic sample (dots) was normalized to the number of RNase P genes found in the corresponding sample (Methods).

body sites analyzed (table 3.2).

The most abundant selenoprotein in the human microbiota was formate dehydrogenase (FDH), which was detected in all body sites. FDH catalyzes the reversible oxidation of CO2 to formate, and is involved in energy metabolism, carbon fixation and pH homeostasis. FDH is the most abundant selenoprotein found in prokaryotes [Stock and Rother, 2009]. It was the most abundant selenoprotein gene in all body habitats except for the lower gut (stool samples).

The second most abundant selenoprotein family was Glycine reductase subunit B (GrdB). The glycine reductase system is involved in the acetate formation

Table 3.2: Abundance of selenoproteins families in each body site. Only families with more than ten occurrences are shown.

	Oral	Stool	Airways	Skin	Vagina
Formate dehydrogenase (FDH)	2424	126	30	17	14
Glycine reductase B (GrdB)	1552	150	12	5	9
Selenophosphate synthetase (SPS)	1417	202	6	3	5
Proline reductase (PrdB)	1066	85	9	9	2
Glycine reductase A (GrdA)	682	101	13	0	3
Arsenate reductase	478	6	0	0	1
NADH:ubiquinone ox-red (RnfC)	407	0	0	0	1
Radical SAM	282	9	1	2	5
BFD/(2Fe-2S)-binding	149	1	1	1	0
Iodothyronine deodinase (DI)	33	4	5	3	1
Ferredoxin-Thioredoxin reductase	34	7	0	1	2
Mercuric transport protein	1	25	0	4	0
Rhodanese-related	3	12	10	3	0
UGSC-containing	3	20	0	0	0
Glutathione peroxidase (GPx)	13	8	0	1	0
HesB-like	1	15	0	0	0
Prx-like	13	2	0	1	0

via glycine, and consists of subunits A, B and C. GrdB catalyzes the specific activation of glycine, with Sec presumably directly involved in the catalysis [Stock and Rother, 2009]. Subunit A (GrdA) was also abundantly observed in HMP samples. This is a small redox-active protein which accepts the carboxymethyl group from GrdB [Stock and Rother, 2009].

The gene *SelD* encodes for selenophosphate synthetase. This was found as a selenoprotein in almost all body sites, ranking as the 2nd and 4th most abundant selenoprotein in genomes and metagenomes, respectively. SelD catalyzes the phosphorylation of selenide, a necessary activation step in the utilization of Se. The presence of *SelD* genes has been used as a genetic marker for Se utilization [Mariotti et al., 2015, Lin et al., 2015].

Proline reductase (PrdB) was also among the most abundant selenoproteins found. This enzyme reduces D-proline to 5-aminovalerate. PrdB proteins share similarity with the GrdB family, and contain Sec in a similar motif. Yet, the two families use a different catalytic mechanism [Stock and Rother, 2009].

Novel selenoproteins in the human microbiome

Given the rich diversity represented in the human microbiome datasets, we expected that novel selenoprotein families could be discovered. By applying our new method bSeblastian (Methods), we could identify four candidate new selenoproteins in the whole metagenomic assemblies from the human microbiota. All candidate genes feature a bacterial SECIS downstream of the putative coding sequence and at least one Sec-TGA aligned to conserved Cys residues in homologous protein sequences. We further validated the candidate selenoprotein families investigating whether they co-occurred in fully sequenced genomes with genetic markers for the Sec encoding capacity, such as $tRNA^{Sec}$ and *SelD* (Methods).

MetE-like We identified a previously unreported selenoprotein family homologous to the C-terminal domain of Cobalamin-independent methionine synthase (MetE; Meth_synt_2, PF01717 Pfam domain). To our best knowledge, no one has ever reported Meth_synt_2 as a Sec-containing domain. Yet, surprisingly, we found a MetE protein sequence from *Desulfonatronospira thiodismutans* already annotated with a Sec residue ("U" symbol) in Uniprot (D6SNM6, status "unreviewed") (see figure 3.5). We identified at least 1,271 gene sequences belonging to this family in the HMP metagenomic assemblies. 425 of them (33%) contained a TGA codon all at the same homologous position, while the remaining 846 (67%) had a Cys codon instead (figure S1). Interestingly, this protein family was almost exclusively found in oral samples (figure S1). A conserved bacterial SECIS (bSECIS) was identified downstream the Sec-TGA codon (figure S2A). In the HMP reference genomes (HMRGD), we identified 21 gene sequences (six with Sec and 15 with Cys) in 19 genomes. All 19 genomes also encoded genes of the factors for Sec synthesis, thus having the ability to incorporate Sec in proteins (figure S3).

Cobalamin-independent methionine synthase (MetE) catalyzes the final step in the biosynthesis of methionine by transferring a methyl group from methylte-

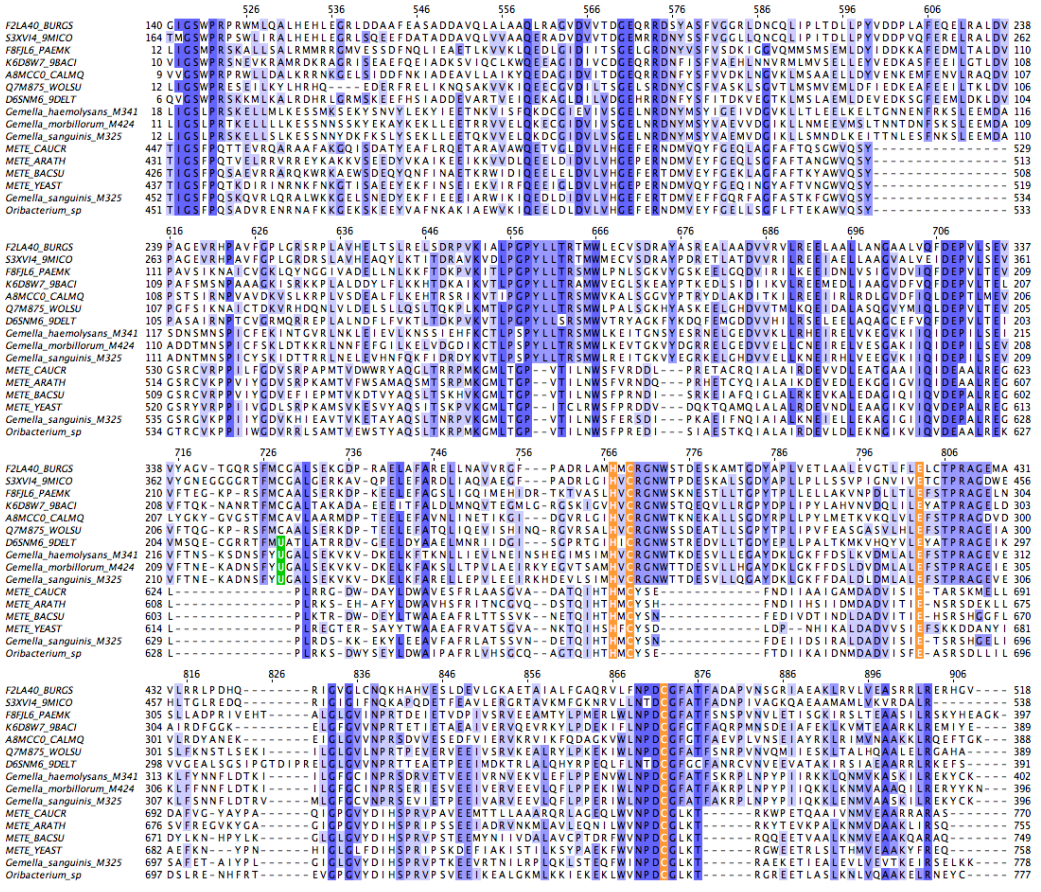


Figure 3.5: Multiple sequence alignment of the C-terminal half of MetE and MetE-like proteins. The first 10 sequences correspond to MetE-like. The Sec position (green) is only present in MetE-like. Known zinc ligands in MetE [Pejchal and Ludwig, 2005] are highlighted in orange. The sequences were obtained from Uniprot (uniprot identifiers) and bacterial genome (species names).

trahydrofolate to L-homocysteine (Hcy), and requires zinc for activation. MetE has a dual barrel structure, where the N-terminal barrel (Meth_synt_1, PF08267) is thought to have evolved from the C-terminal polypeptide by gene duplication. The catalytic C-terminal half is more highly conserved than the N-terminal half and has homologous thiol methyltransferases that are approximately half the size of MetE [Pejchal and Ludwig, 2005]. The sequence alignment of the Meth_synt_2 domain and the Sec-containing MetE-like shows several indel blocks (deletions or

insertions) (figure 3.5). The Sec position maps to one of those regions. Nonetheless, the residues known to interact with zinc in MetE [Pejchal and Ludwig, 2005], are conserved in the MetE-like (figure 3.5). We speculate that MetE-like could be a distant homologue of MetE with thiol methyltransferase activity.

Two putative novel selenoproteins were identified in the HMP samples and fully sequenced genomes. They do not present any domain of known function, and their homologues are annotated in databases as "uncharacterized". Here they are referred as bseb1 and bseb2.

bseb1 is a short ∼80-120 amino acids protein with relatively low sequence conservation, whose alignment required careful manual curation. In this protein family, Sec residues were observed in two possible positions close to each other, and aligned to conserved Cys residues in non-selenoprotein homologues (figure 3.6A). At least 464 Sec-containing sequences were found in the HMP metagenomic assemblies, with a Sec-TGA in either of the two positions. Interestingly, in most of the bseb1 proteins the two Sec/Cys positions were located adjacent, but in many of the Sec-containing sequences they were separated by one (CxU) or two (CxxU) residues, a common motif in redox enzymes often referred to as "redox box" [Chivers et al., 1997]. A search of this protein in fully sequenced genomes showed that Sec containing bseb1 were found only in genomes with a $tRNA^{Sec}$ gene, while this was not necessarily the case for bseb1 Cys-homologues (figure S4A).

Our second novel selenoprotein gene candidate, bseb2, encodes a ∼300 amino acids protein with a conserved DCC (Asp-Cys-Cys) motif in which the second Cys is replaced by Sec (figure 3.6B). We identified at least 21 genes in HMP samples, 15 of them with a Sec-TGA codon. The gene was also found in 53 fully sequenced bacterial genomes; 39 of them had a Sec-TGA codon. All 53 genomes also encoded a $tRNA^{Sec}$ (figure S4B).

Finally, we identified a putative Sec-containing Transposase DDE domain (Tnp_DDE_dom, IPR025668), here referred as TnpSec. The domain contained a "redox box" motif (CXXC), in which the second Cys aligned with putative Sec-TGA in two almost identical sequences (figure 3.6C). One of the sequences was observed identical in eight distinct samples. The Sec-containing domain was not found in fully sequenced genomes. Although we cannot discard a sequencing error or a common pseudogene, given that the Sec-TGA codon was observed in nine

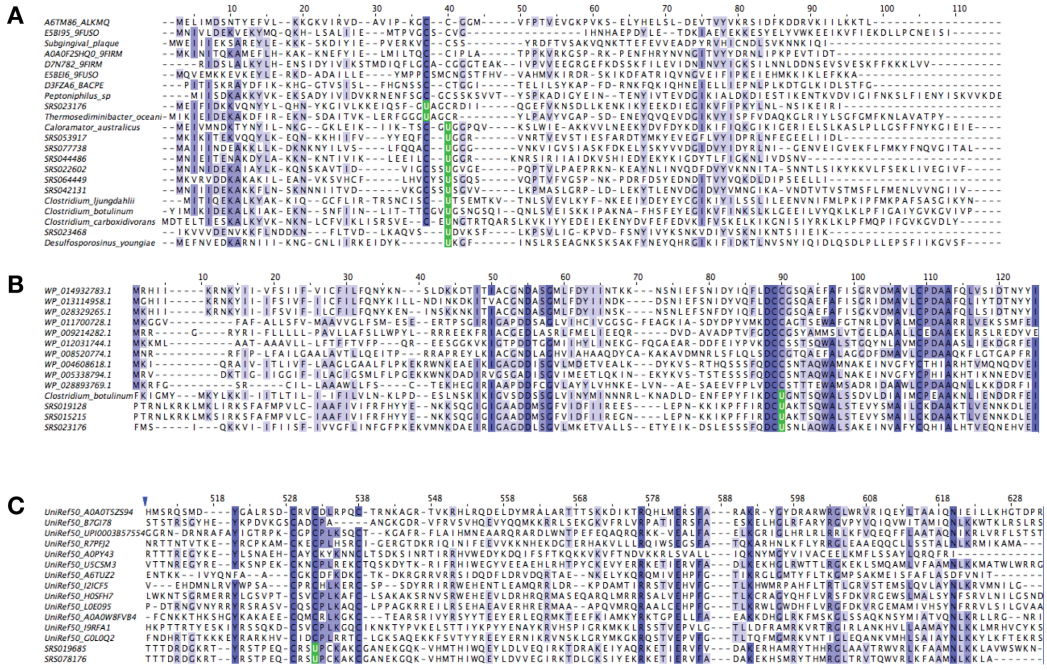


Figure 3.6: Novel candidate selenoproteins: bseb1 (A), bseb2 (B), TnpSec (C). The putative Sec-TGA residues are highlighted in green. The sequences were obtained from different sources, reflected in their identifiers: the SRS prefix and "Subgingival_plaque" denote sequences from HMP metagenome assemblies; WP from RefSeq NCBI; species names denote sequences that were found in complete bacterial genomes; the rest of identifiers are from UniProtKB.

different samples we believe that TnpSec is likely a true selenoprotein gene.

tRNA^{Sec} genes in the human microbiota

tRNA^{Sec}, the Sec specific tRNA, is a key molecule in the synthesis of selenoproteins. We recently developed Secmarker [Santesmasses et al., section 2.1 in this thesis], a program to accurately predict tRNA^{Sec} genes in genomes, which we applied here to search all the 770 HMP assemblies (including all HMASM, Methods). We identified a total of 2,169 candidate tRNA^{Sec} sequences in samples from all body sites. The vast majority of tRNA^{Sec} belonged to bacteria, although 7 of them were of eukaryotic origin. The eukaryotic sequences, all distinct from

human tRNA^{Sec}, were detected in samples from the retroauricular crease (skin) and stool (lower gut). No archaeal tRNA^{Sec} was detected, consistently with selenoproteins being a rare trait among archaea [Stock and Rother, 2009]. Similar to the distribution of selenoproteins observed in this study, most tRNA^{Sec} genes (85%) were detected in oral samples, while that fraction was much lower (12%) in stool samples.

We found 86 bacterial tRNA^{Sec} sequences with an unusual structure, different from the canonical 13 base pairs (8+5) AT-stem. These candidates had a 7 base pairs acceptor stem (with 7 nucleotides between the T-stem and the discriminator base) and 5 a bp T-stem. We recently identified similar sequences in *Gammaproteobacteria*, *Clostridiales* and *Spirochetes* [Santesmasses et al., section 2.1 in this thesis].

Selenium utilization traits

We investigated the presence of the three Se utilization traits in the human microbiome across body sites. We identified the different genetic markers for the three pathways in the metagenomic samples and quantified their abundances (Methods). Our results suggest that, consistently with previous studies in fully sequenced genomes, Sec is more abundant than SeU and Se-cofactor, in most samples (figure S5). Yet, the gastrointestinal tract constitutes an exception, as Se-cofactor appears more abundant than Sec (figure 3.7). On average, stool samples had twice as many Se-cofactor markers as Sec markers. No markers for SeU and Se-cofactor were detected in Airways (figure 3.7).

3.3.3 Discussion

Here we looked at selenoproteins in the human microbiota for the first time. This novel study is important because these selenoproteins are present in our body, and they need to be taken into account, along with the other Se utilization traits, to better understand the effects of selenium in our organism.

Our results show that selenoproteins are commonly present in the microbiota of the different body habitats of healthy individuals. The most abundant selenoprotein families are, similar to other studies in completely sequenced genomes and environmental metagenomes, formate dehydrogenases, selenophosphate syn-

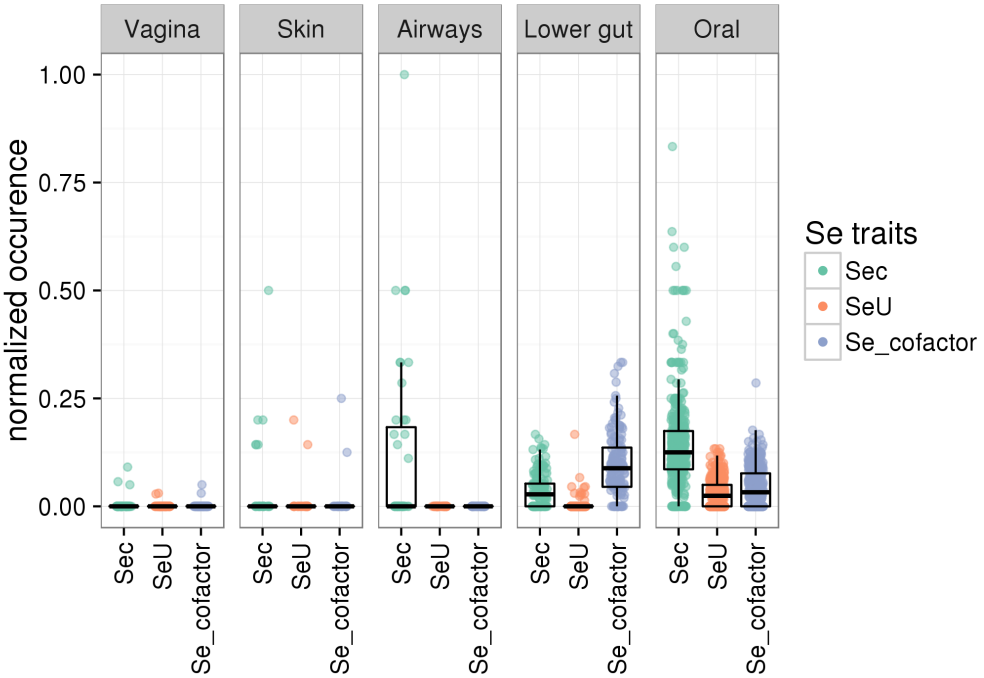


Figure 3.7: Selenium utilization traits across body sites. The occurrence of the genetic markers for the three Se traits in each samples was normalized to RNase P (Methods).

thetases and glycine and proline reductase systems. We observed that selenoproteins are remarkably abundant in the oral cavity, and they appeared to be scarce in stool samples. Surprisingly, in stool samples the Se-cofactor was found to be more abundant than the Sec trait.

To do this analysis we applied computational tools developed in the last years and successfully used in numerous studies, as well as newly developed ones. bSeblastian is a new pipeline for the identification of bacterial selenoproteins based on the identification of bSECIS as first step. The detection of bSECIS is carried out using covariance models enabled by Infernal [Nawrocki and Eddy, 2013].

The program bSeblastian allowed us to identify four novel candidate selenoproteins, three of which have support from different sources, including completely sequenced genomes that also encode the factors for the synthesis of selenopro-

teins.

This study advances our understanding of the human microbiota. The results shed light into the use and the distribution of selenocysteine, as well as the other selenium utilization traits, by the bacterial cells we host in and on our body. The selenium availability in the human body might be important for the regulation, not only selenoproteins encoded in our genome, but also those encoded in our microbiota, whose impact in human health and disease is not well understood, but increasingly studied.

3.3.4 Methods

Human Microbiome Project data

The genomic data used in this work is part of the Human Microbiome Project (HMP) [The Human Microbiome Project Consortium, 2012]. In the context of that project, 242 healthy adult individuals were sampled targeting five clinically relevant major body sites, as a catalog of the human microbiota. Nine specimens from the oral cavity and oropharynx: saliva; buccal mucosa (cheek), keratinized gingiva (gums), palate, tonsils, throat, and tongue soft tissues; and supra- and subgingival dental plaque (tooth biofilm above and below the gum). Four skin specimens were collected from the two retroauricular creases (behind each ear) and the two antecubital fossae (inner elbows), and one specimen for the anterior nares (nostrils). One stool specimen represented the microbiota of the lower gastrointestinal tract, and three vaginal specimens were collected from the vaginal introitus, midpoint, and posterior fornix (figure 3.8). We downloaded whole metagenomic assemblies from the Data Analysis and Coordination Center (DACC) (http://www.hmpdacc.org): a total of 755 sample assemblies were downloaded from (http://hmpdacc.org/HMASM/), and 15 body-site specific assemblies were downloaded from (http://hmpdacc.org/HMBSA/). The metadata used to map the sample identifier with the subject identifier and its gender was downloaded from IMG/HMP (https://img.jgi.doe.gov/cgi-bin/imgm_hmp/main.cgi). Seven of the 755 HMASM assemblies could not be associated with a subject identifier; those samples were excluded from the analyses, except when noted. The 748 samples were obtained from 106 individuals. In addition, reference genomes from bacteria cells isolated from human body sites were downloaded from

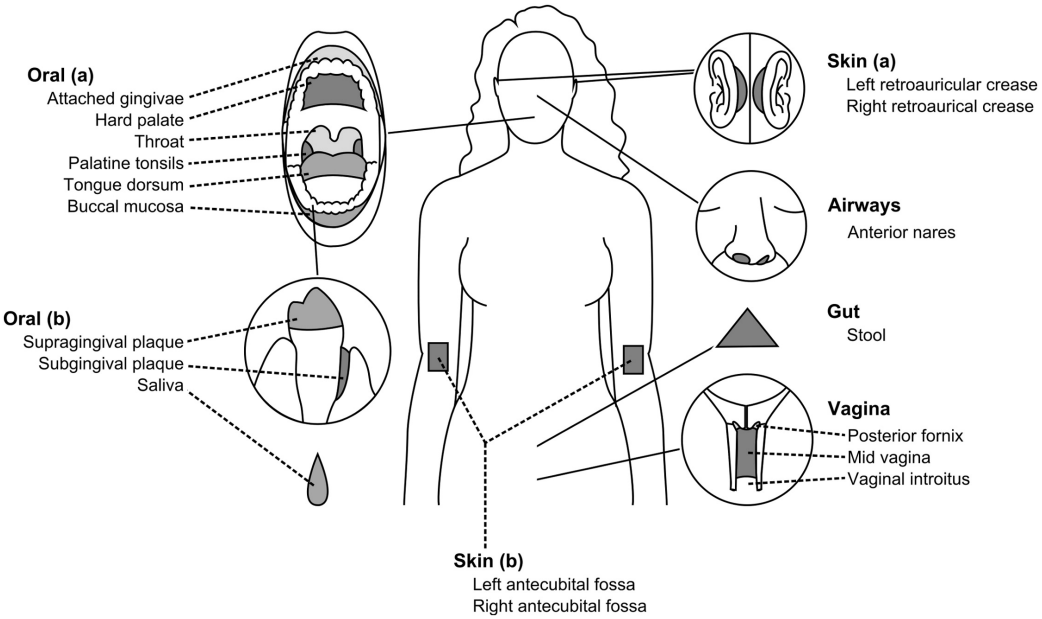


Figure 3.8: Schematic of the Body Sites sampled for the HMP Healthy Adult Cohort Study. 18 body subsites from 5 main body areas were sampled. Note that some subsites were poorly represented, or not at all (antecubital fossa), among the whole metagenome assemblies used in this work (HMASM). From [Proctor, 2011].

http://hmpdacc.org/HMRGD/ (1096, accessed on Nov 10, 2014). Further information on data collection and protocols is available in http://hmpdacc.org.

Identification of known selenoproteins and genetic markers for the Se utilization traits

We used Selenoprofiles [Mariotti and Guigó, 2010] and bSeblastian, a newly developed method for identification of selenoproteins based on the identification of bacterial SECIS as first step (section 2.2 in this thesis), to identify known selenoprotein families in the HMP assemblies. The union of the predictions coming from the two programs were used as the final set of genes. Selenoprofiles was also used to identify the protein coding genes used as genetic markers of the Se utilization traits: *SelA* for the Sec trait; *yqeB* and *yqeC* for the Se-cofactor; and *ybbB* for the SeU trait. $tRNA^{Sec}$, also for the Sec trait, was predicted with Secmarker

[Santesmasses et al., section 2.1 in this thesis]. We found a good correlation in the number of predictions for *SelA* and $tRNA^{Sec}$ in each sample (0.91 spearman's rho, $p<0.05$); for the Sec trait we used the lowest value. A good correlation was also found for *yqeB* and *yqeC* (0.88 spearman's rho, $p<0.05$), the lowest value in each sample was used.

Estimation of number of genomes in a metagenomic assembly

Each metagenomic assembly contains DNA from an unknown number of organisms. We estimated the number of genomes present in each assembly using the abundance of the Ribonuclease P (RNase P). RNase P was selected after a search of the Rfam database (RNA families) with Infernal [Nawrocki and Eddy, 2013] in a the set of 223 prokaryotic fully sequenced genomes revealed that the RNA subunit of RNase P was present as a single copy gene in each of the genomes. RNase P is a ribonucleoprotein that cleaves RNA, known for its role in 5'-processing of tRNA precursors [Ellis and Brown, 2009]. Homologues of the catalytic RNA subunit are conserved across archaea, bacteria and eukaryotes. We identified the RNA subunits of RNase P in the metagenomic assemblies using three Rfam models (bacteria type A (RF00010), bacteria type B (RF00011) and archaea (RF00373)) with infernal (`--rfam` option active and e-value <= 1e-5). We found a good correlation between the number of RNase P genes and the assembly length in the metagenomic samples (0.93 spearman's rho, $p<0.05$). However, regression analysis by body site showed a noticeable lower slope in the gut samples, compared to the oral ones (figure S6B): for a given assembly length, the gut samples harbored less RNase P genes than oral samples. Since we expected every genome present in the metagenomic assemblies to contain one copy of RNase P, the observed shift towards longer lengths could be explained if the genomes from the gut samples were on average larger in total sequence length than the ones from the oral samples. To investigate the length of microbial genomes from different body sites we used the HMP reference genome assemblies (HMRGD). We identified a single RNase P gene in the vast majority of them, with the exception of 3 genomes with two copies (figure S7). We found indeed that the microbial species from the gastrointestinal tract had longer genomes (figure S6D). Since no RNase P could be identified in some assemblies, normally the shortest ones (Fig S8), assemblies

shorter than 1Mb were removed from the analysis.

Search for new selenoproteins in HMP assemblies

The last step of bSeblastian (see section 2.1 in this thesis) consists of a blast search of the candidate TGA-containing ORFs against a protein database to identify those with selenoprotein coding potential. In order to minimize the computational time, we used a small database consisting only of known selenoproteins, which allowed us to identify most of the known selenoproteins among the candidate ORFs. Additional steps were performed to analyze all the remaining TGA-containing ORFs. bSeblastian predicted more than 2 million candidate ORFs in the HMP metagenomic assemblies. We pulled together all ORFs to generate a single database (here referred to as "ORF full db"). We then removed redundancy using CD-HIT [Li and Godzik, 2006], reducing approximately fourfold the size of the database–from $\sim$2.6 million ORFs to $\sim$0.59 millions. The non-redundant database (here referred as "ORF nr db") was run with blast [Altschul et al., 1997] against two different sequence databases: UniRef50 and the full ORF db. An e-value threshold of 1e-4 was used. The blast outputs were then parsed and filtered. We first excluded those hits spanning less than 30% of the candidate ORF length. For each candidate, the total number of hits, the number of hits spanning the TGA, and the residues aligned to the TGA were computed. We then selected those candidates that satisfied all following requirements: i) a minimum of 50% of the hits spanned the TGA, ii) 50% of the hits that spanned the TGA had a C or U aligned to the TGA, and iii) no hits that aligned a tryptophan (Trp, W) with the TGA (Trp is encoded by UGA in the mitochondrial and *Mycoplasma*/*Spiroplasma* codes). The requirements had to be satisfied in both databases (UniRef50 and ORF full db). This procedure resulted in a filtered list of 529 ORFs. These candidates were further analyzed. For each ORF, a multiple sequence alignment was built using mafft [Katoh et al., 2002] with the top 100 non-identical target sequences from the blast hits (against the full ORF db). The alignments were used to build profiles that were then searched using Selenoprofiles [Mariotti and Guigó, 2010] in a large collection of completely sequenced bacterial genomes obtained from NCBI. The results allowed to identify the four candidate novel selenoprotein families presented in Results.

3.3.5 Supplementary materials

The following pages correspond to the supplementary materials in this study.

The human selenomicrobiome.
Supplementary materials

Table S1. Nemenyi-test p-values for the pairwise comparisons of the normalized number of selenoproteins in the different body sites. The Chi-squared approximation for independent values was used.

	Airways	Stool	Oral	Skin
Stool	5.5e-07	-	-	-
Oral	0.1968	<2e-16	-	-
Skin	0.4176	0.1638	0.0006	-
Urogenital tract	8.4e-11	0.1033	<2e-16	0.0018

Figure S1. Occurrence of Sec- and Cys-containing MetE-like genes in HMP metagenome assemblies across body subsites.

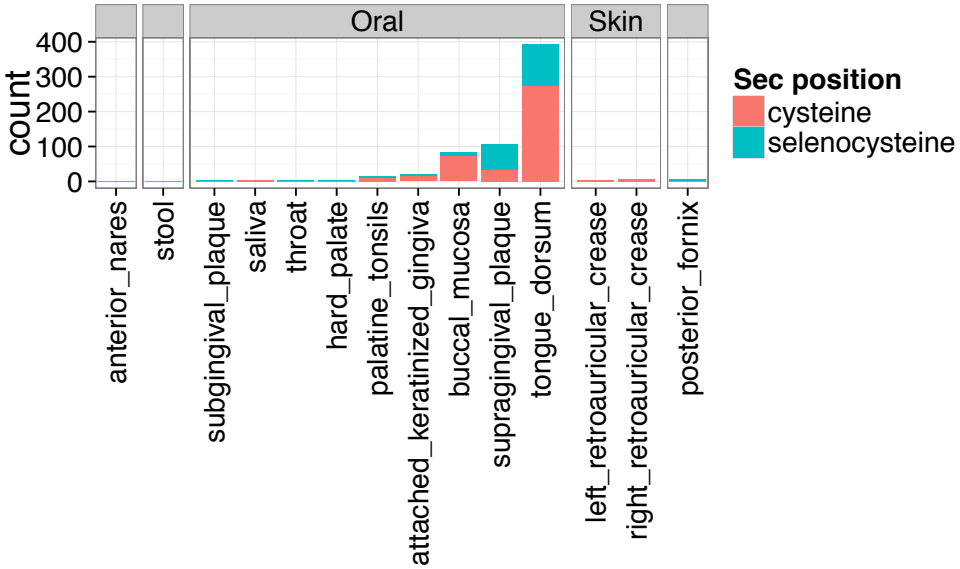


Figure S2. Conserved bSECIS structure in Sec-containing MetE-like (A) and the candidate novel selenoprotein bseb1 (B). The structure-based multiple sequence alignment includes the region downstream the Sec-UGA codon (red underline). The alignments were produced with cmalign from Infernal [Nawrocki and Eddy, 2013] and visualized with RALEE [Griffiths-Jones et al., 2005]. RALEE highlights nucleotides that form canonical pairs based on the consensus secondary structure (second line from bottom).

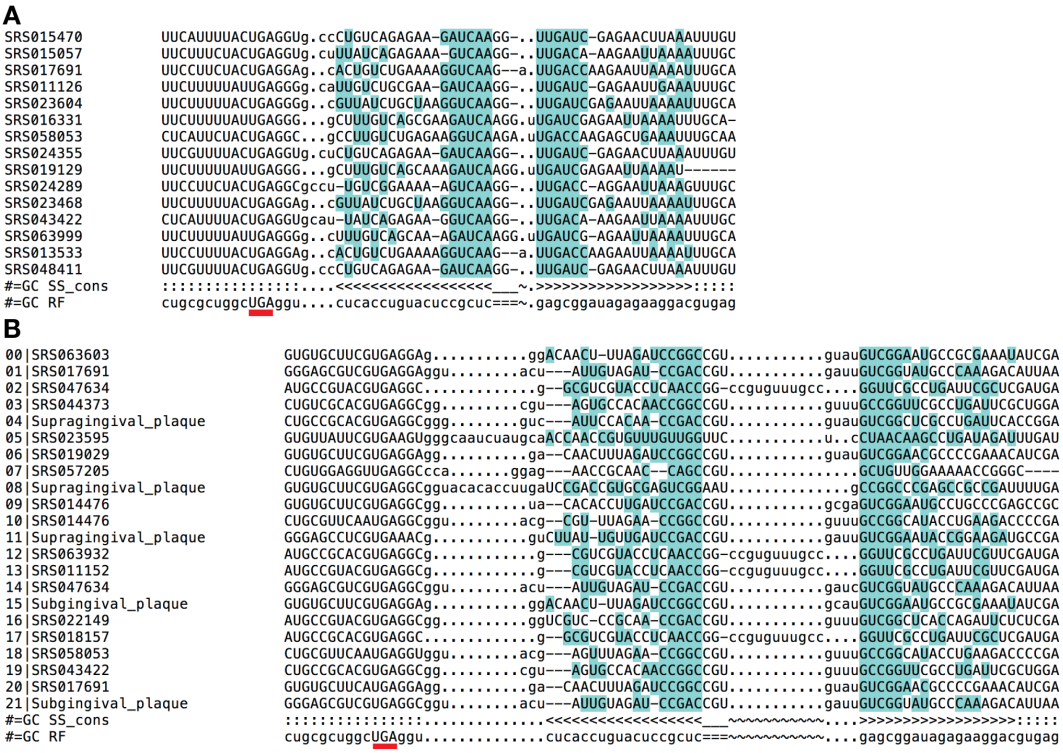


Figure S3. Species tree of the 19 bacterial genomes from HMP (HMRGD) with MetE-like genes. All species contain the Sec synthesis factors (*SelA*, *SelB*, *SelC*/tRNA-Sec, and *SelD*) and MetE-like with either Sec (green), Cys (red), or both. Phylogeny obtained from NCBI taxonomy.

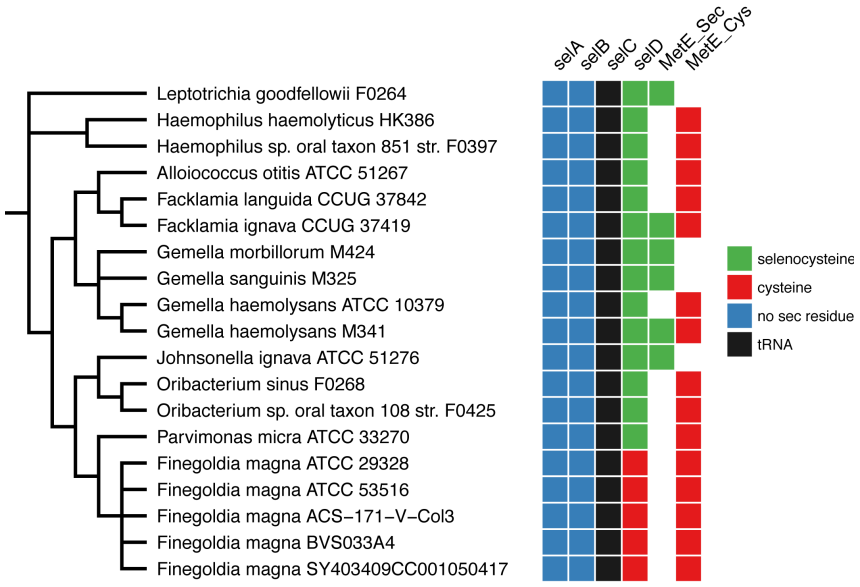


Figure S4. Genomic co-occurrence of the novel selenoprotein candidate and Sec-tRNA, the Sec encoding marker. The phylogenetic tree of the bacterial species was obtained from NCBI taxonomy and annotated with the presence of the two putative selenoproteins. The colored cells correspond to the presence of bseb1 (A) and bseb2 (B), either with Sec (green) or Cys (red). The black dots at the tip of the branches of the tree indicate the presence of tRNA-Sec.

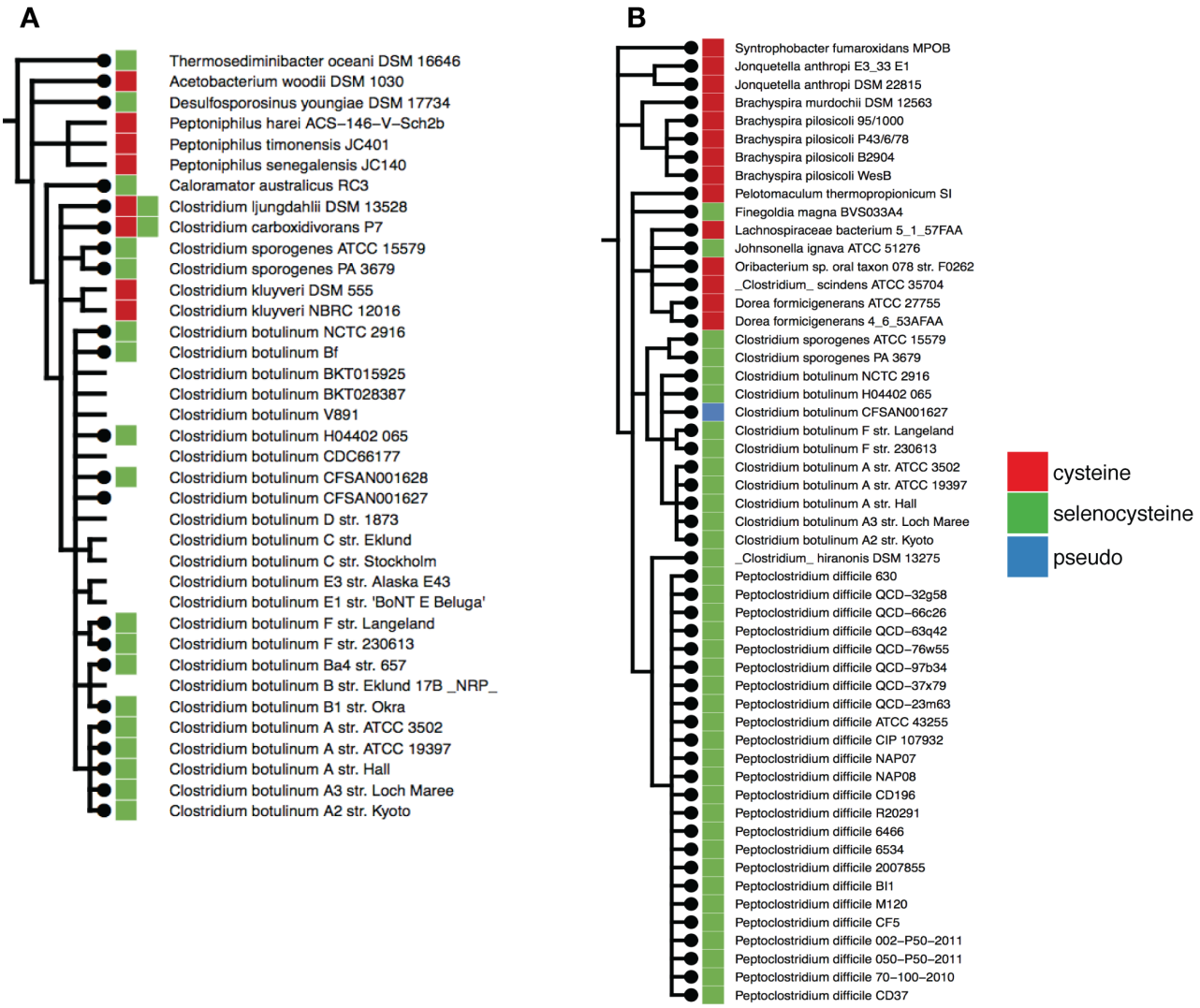


Figure S5. Selenium utilization traits across body sites. The abundances of Se-cofactor and SeU were compared to the abundance of Sec in each sample (dots), across the five body areas (rows). The x axis corresponds always to the quantification of the Sec trait, and the y axis corresponds to Se-cofactor (left column) and SeU (right column). In those samples found below the identity line, the Sec trait was more abundant than the corresponding trait in the y axis. Darker dots indicate the overlap of multiple samples with the same quantifications.

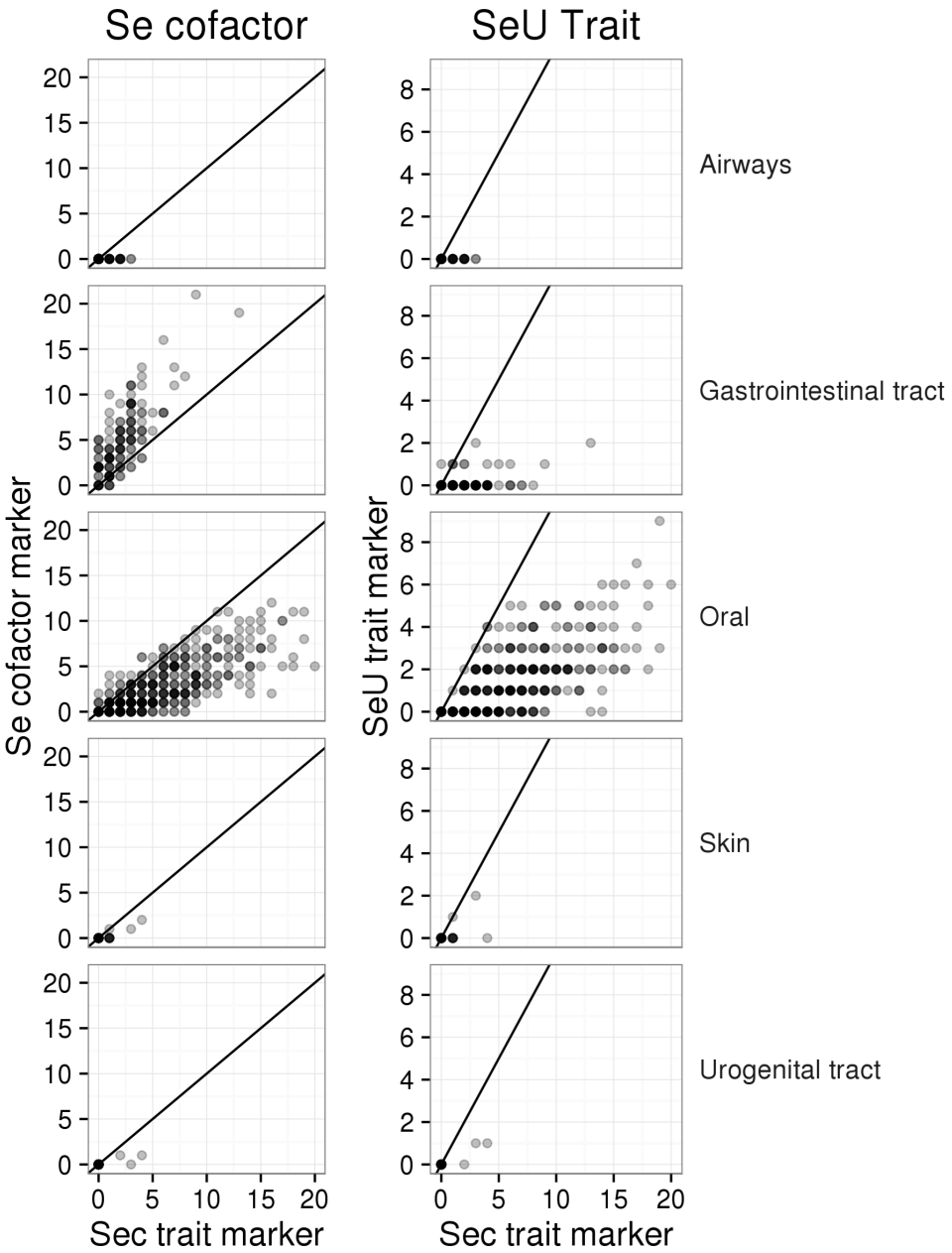


Figure S6. RNA subunit of RNase P in the HMP samples. A) the predicted secondary structure of the RNA subunit of RNase P in *Bacillus subtilis* (source: wikipedia). B) Relationship between the number of RNase P predictions in HMP metagenome assemblies (dots) and their size in nucleotides. C) Distribution of RNase P predictions per nucleotide in HMP samples, across body sites. D) Distribution of the genome size in completely sequenced genomes, obtained from specific body sites (HMRGD). The colors of the legend apply to B,C and D.

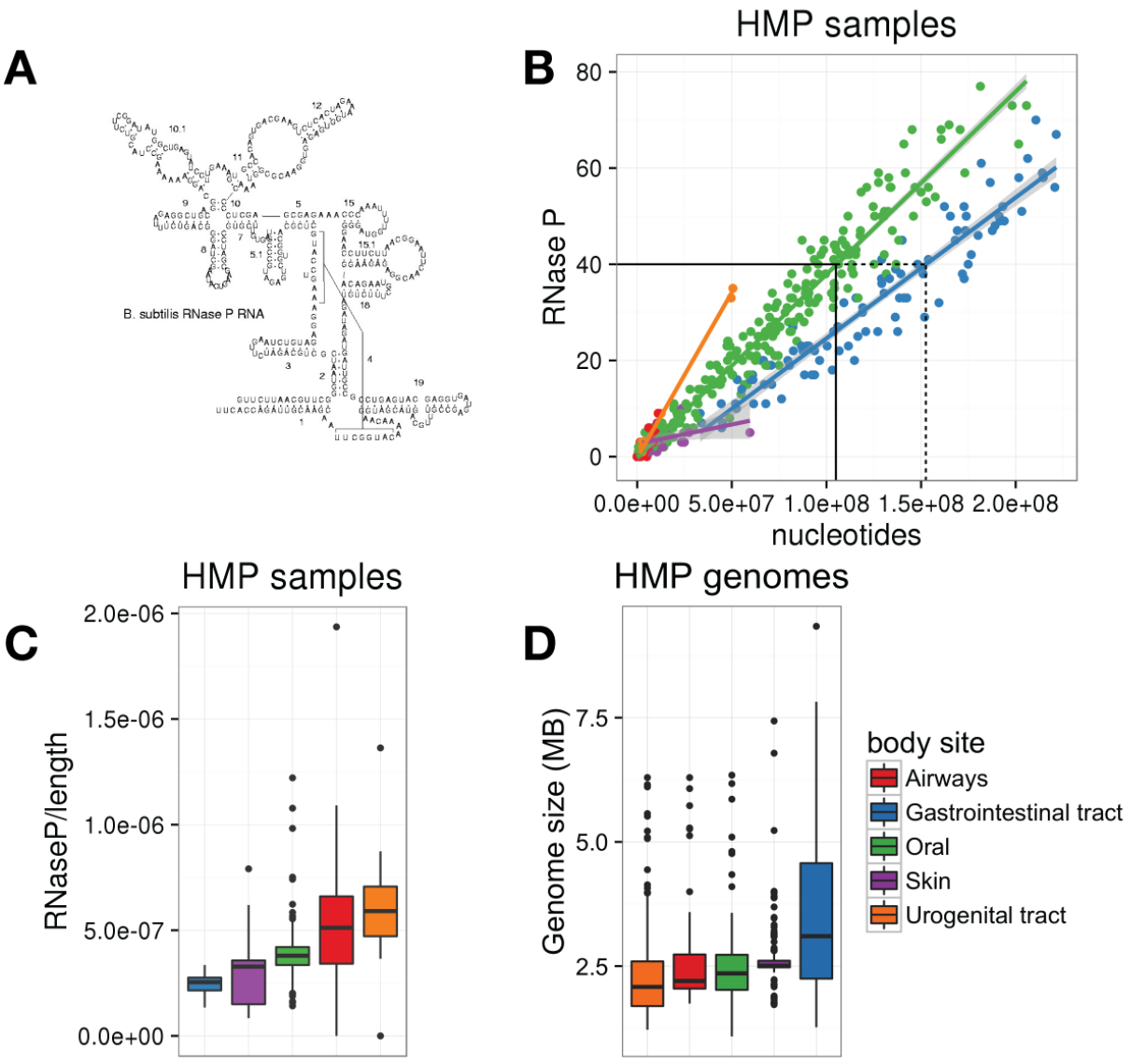


Figure S7. Sunburst diagram showing the phylogeny of the 1096 completely sequenced genomes (HMRGD), annotated with the number of copies of the RNA subunit of RNase P, identified in their genomes. The three external rings correspond to the results obtained with the three Rfam models: RNaseP_arch (archaea), RNaseP_bact_a (bacteria type A), RNaseP_bact_b (bacteria type B).

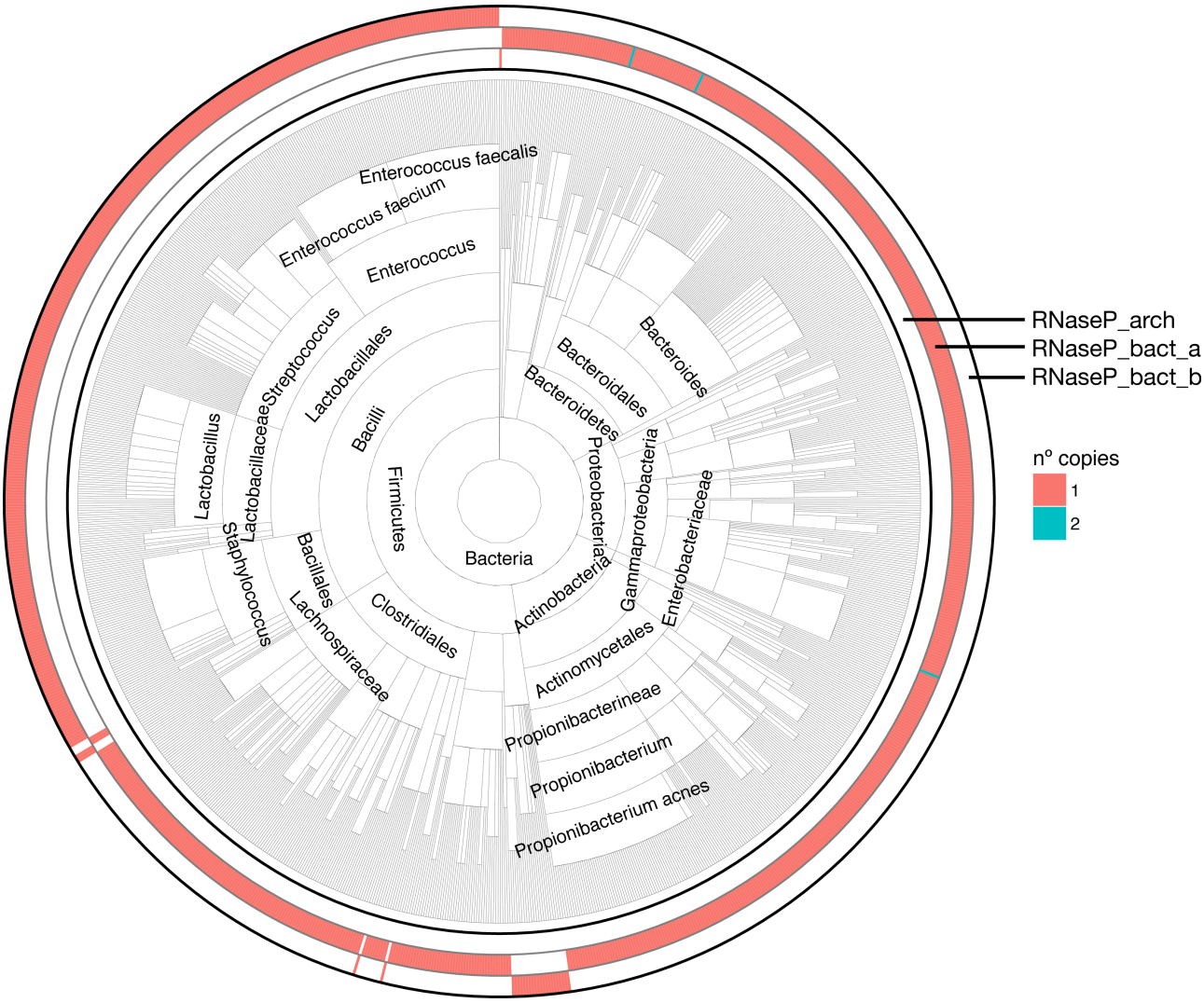
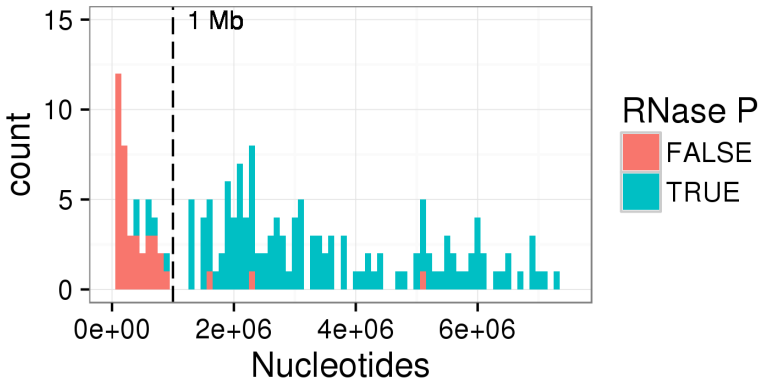
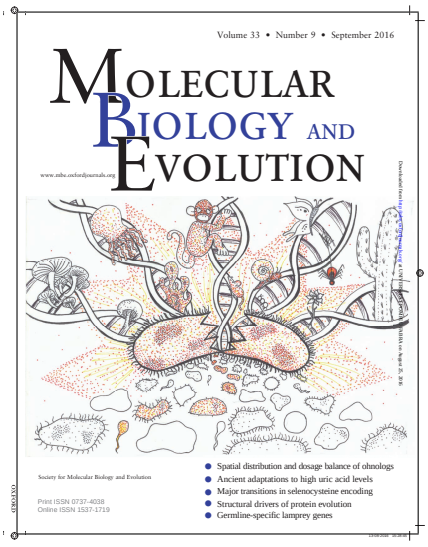


Figure S8. Metagenome assemblies shorter than 1Mb were removed from the analysis. The histogram shows the distribution of the size of the assemblies (shorter than 7MB). Red indicates assemblies without RNase P predictions.



3.4 *Lokiarchaeota* selenoproteome

Lokiarchaeota is an archaeal phylum that forms a monophyletic group with eukaryotes. Until recently, archaeal selenoproteins were only know from *Methanococcales* and *Methanopyrus*. This study describes the selenoprotein genes present in *Lokiarchaeota*. The detailed analysis of the SECIS structures from *Lokiarchaeota* selenoproteins showed that they have eukaryotic-like features, suggesting that these features were already established in archaea and propagated to eukaryotes. Secmarker was used to identify $tRNA^{Sec}$ genes, and we reported here the first known intron-containing $tRNA^{Sec}$.



Mariotti M, Lobanov A V., Manta B, Santesmasses D, Bofill A, Guigó R, et al. Lokiarchaeota Marks the Transition between the Archaeal and Eukaryotic Selenocysteine Encoding Systems. Mol Biol Evol. 2016 Sep;33(9):2441–53. DOI: 10.1093/molbev/msw122

3.5 Evolution of selenophosphate synthetases

In this work we described the evolution of selenophosphate synthetases, the enzyme responsible for the production of selenophosphate, the Se activated donor required for Sec synthesis (and other Se-dependent pathways in prokaryotes). In many organisms, the protein carries a Sec rescue itself. The gene underwent several independent duplication events in different metazoan lineages, in a nice example of convergent evolution. The new protein (SPS1), despite having independent origins, is characterised for having lost the Sec residue—it is not a selenoprotein. SPS1 performs a function different than that of the ancestral enzyme (SPS2). The work describes with great detail the phylogenetic distribution of the two genes across the Tree of Life, that serves as a map for the utilisation of Se.

This paper is not included in this thesis because the results were already presented in the thesis [Mariotti, 2013]. The publication of this paper was highlighted in the cover of the journal, and the figures 1 and 2, produced with ggsunburst (an R package I developed, see appendix A) were published as a poster (figure 3.10).

Mariotti M, Santesmasses D, Capella-Gutierrez S, Mateo A, Arnan C, Johnson R, et al. Evolution of selenophosphate synthetases: emergence and relocation of function through independent duplications and recurrent subfunctionalization. Genome Res. 2015 Sep;25(9):1256–67. DOI: 10.1101/gr.190538.115

Research

Evolution of selenophosphate synthetases: emergence and relocation of function through independent duplications and recurrent subfunctionalization

Marco Mariotti,[1,2,3,4] Didac Santesmasses,[1,2,3] Salvador Capella-Gutierrez,[1,2] Andrea Mateo,[5] Carme Arnan,[1,2,3] Rory Johnson,[1,2,3] Salvatore D'Aniello,[6] Sun Hee Yim,[4] Vadim N. Gladyshev,[4] Florenci Serras,[5] Montserrat Corominas,[5] Toni Gabaldón,[1,2,7] and Roderic Guigó[1,2,3]

[1]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), 08003 Barcelona, Catalonia, Spain; [2]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Catalonia, Spain; [3]Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), 08003 Barcelona, Catalonia, Spain; [4]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA; [5]Departament de Genètica, Facultat de Biologia and Institut de Biomedicina (IBUB) de la Universitat de Barcelona (UB), 08028 Barcelona, Catalonia, Spain; [6]Department of Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dorhn, Villa Comunale, 80121, Napoli, Italy; [7]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Catalonia, Spain

Figure 3.9: Evolution of selenophosphate synthetases [Mariotti et al., 2015].

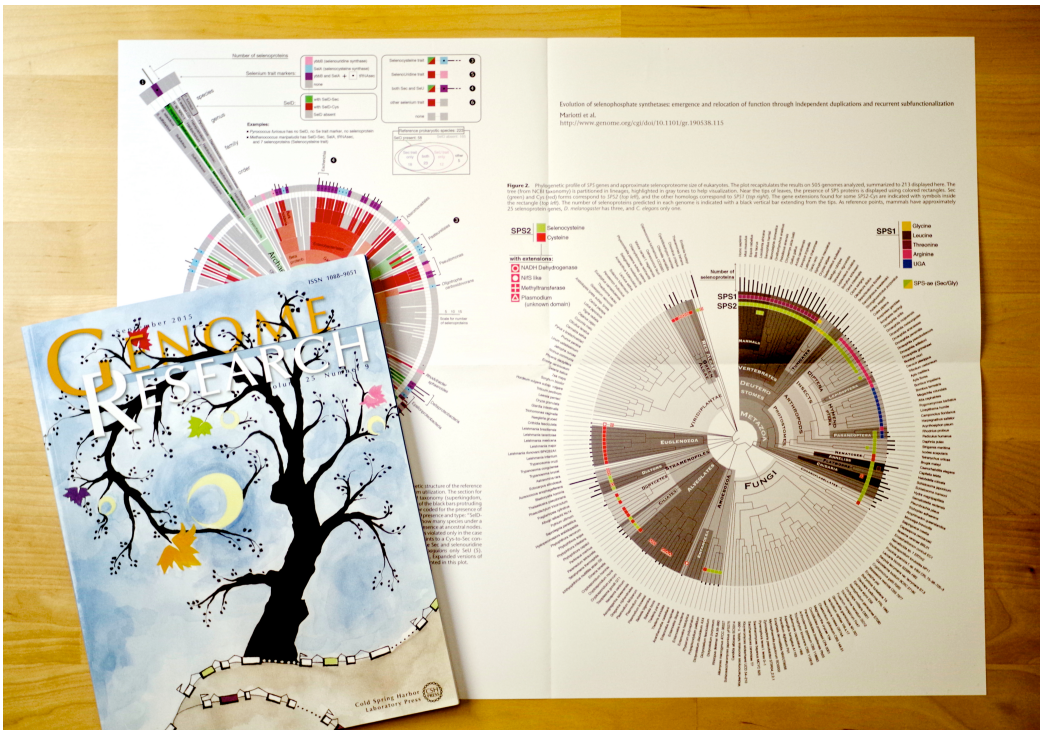


Figure 3.10: Cover and poster [Mariotti et al., 2015] from the *Genome Research* September 2015 issue.

3.6 Selenoprotein extinctions (cont.)

This section includes a summary of the results presented in [Mariotti, 2013], and includes new results in the study of selenoprotein extinctions in insects. The new results are the analysis of expression data (RNAseq) from multiple Drosophila species, and the analysis of selenoprotein genes in additional species.

3.6.1 Known Sec extinction in *Drosophila*

In the work of [Chapple and Guigó, 2008] and [Lobanov et al., 2007], it was described how several insects lost selenoproteins, along with the ability to make selenocysteine. Sec extinctions happened in parallel insect lineages. In these lineages, selenoprotein genes are either converted to cysteine homologs or lost, and the synthesis machinery degenerated concomitantly. However, little is known about the process of selenoprotein extinction, its evolutionary pathway, and the causes or consequences. We investigated the most recent Sec extinction known, that of *Drosophila willistoni*. This species is estimated to have diverged from the rest of sequenced drosophila about 35 million years ago. A few features set it apart from the other drosophila: a lower genomic GC content, a lower codon bias in coding sequences (favouring AT nucleotides) [Powell et al., 2003, Clark et al., 2007]. We attempted to widen the spotlight around *D. willistoni*, trying to map more precisely its Sec loss. A survey using degenerate PCR primers in 23 species from three lineages: *willistoni*, *obscura* and *saltans*. From PCR results the *saltans* group was the most interesting, we thought it may contain both species with and without selenoproteins. In view of these results, we sequenced the full genome of 8 species in the saltans group: *D. austrosaltans*, *D. emarginata*, *D. lusaltans*, *D. milleri*, *D. neocordata*, *D. prosaltans*, *D. saltans* and *D. sturtevanti*. A whole-genome annotation was produced for 29 Drosophila genomes, with orthology and paralogy predictions. The phylogeny of 30 dipteran species, including the eight species from the *saltans* group, was inferred using the protein sequences of 566 one-to-one orthologous genes, in collaboration with Jaime Huerta-Cepas and Salvador Capella-Gutierrez from the group of Toni Gabaldón, in our department.

We searched selenoproteins and Sec machinery in the *saltans* group genomes, as well as in the rest of available *Drosophila* genomes, using Selenoprofiles [Mar-

iotti and Guigó, 2010]. Our predictions replicate well the results in [Chapple and Guigó, 2008]. The same selenoproteome of *D. melanogaster* (SPS2, SelG and SelH) [Castellano et al., 2001] and Sec machinery were found in all 12 reference *Drosophila* genomes as well as the public genomes previously not analyzed. A few genes were predicted with pseudogene features (in-frame stop codons or frameshifts), but considering the imperfect quality of genome assemblies, we must assume that these are actually intact in the real genome.

3.6.2 Novel Sec extinctions in *Drosophila*

The eight species in the *saltans* group revealed to be very interesting for selenoproteins, as expected from the PCR results. Four of the species, *D. saltans*, *D. austrosaltans*, *D. prosaltans* and *D. lusaltans* had the same selenoproteome and Sec machinery as *melanogaster*, while the rest, *D. sturtevanti*, *D. milleri*, *D. neocordata* and *D. emarginata*, had lost the Sec genes, or converted them to Cys homologues, and the Sec machinery was incomplete. After analyzing all selenoprotein and Sec machinery genes in our species set, we inferred their phylogenetic history, in terms of gene losses or conversions. Figure 3.11 displays a summary of the extant genes and events in the *willistoni/saltans* lineage.

We consider *D. neocordata* the most interesting species in our set. Here, selenoproteins SelG and SelH have been converted to cysteine. SPS2 and other Sec machinery genes could be detected, but with pseudogene features, which could be confirmed in RNAseq samples. A $tRNA^{Sec}$ was also detected, but some point mutations appeared in otherwise conserved positions in *Drosophila* $tRNA^{Sec}$. Taken altogether, these observations indicate that *D. neocordata* underwent a selenoprotein extinction very recently.

Summarizing, we found 3 more Sec extinction events in the *saltans* group, one of which is so recent that all Sec machinery genes are still recognizable (*D. neocordata*). Including *D. willistoni*, we have now 4 events of Sec extinctions that happened in parallel drosophila lineages. Considering that the *saltans* and *willistoni* groups are phylogenetically sisters, we can say that all such events (although independent) happened in a single lineage of drosophila. This prompted us to think that a physiological change occurred at the root of this lineage, favoring later Sec extinctions. This hypothesis is analogous to the one proposed in [Chap-

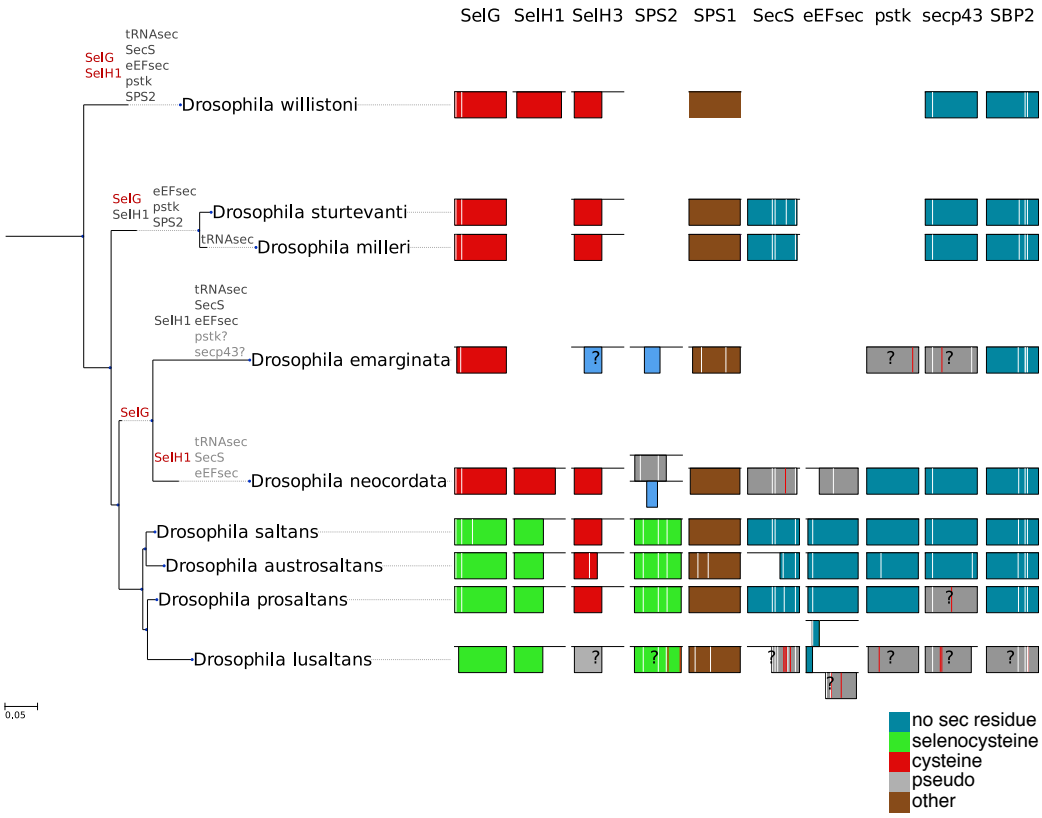


Figure 3.11: Phylogenetic tree of the *willistoni/saltans* lineage. The gene predictions for selenoproteins and Sec machinery are included. The evolutionary events involving the three selenoproteins are indicated in the corresponding node of the tree: in red, Sec to Cys conversion; in grey gene losses. (Plot provided by Marco Mariotti).

ple and Guigó, 2008] for the root of insects, and must be seen complementary to it.

3.6.3 *willistoni/saltans*: GC content and codon bias

Having whole-genome annotations for all *Drosophila* species used in this work, we analyzed their GC content and codon usage. From literature [Powell et al., 2003], we expected *saltans* and *willistoni* to be homogenous for GC content and

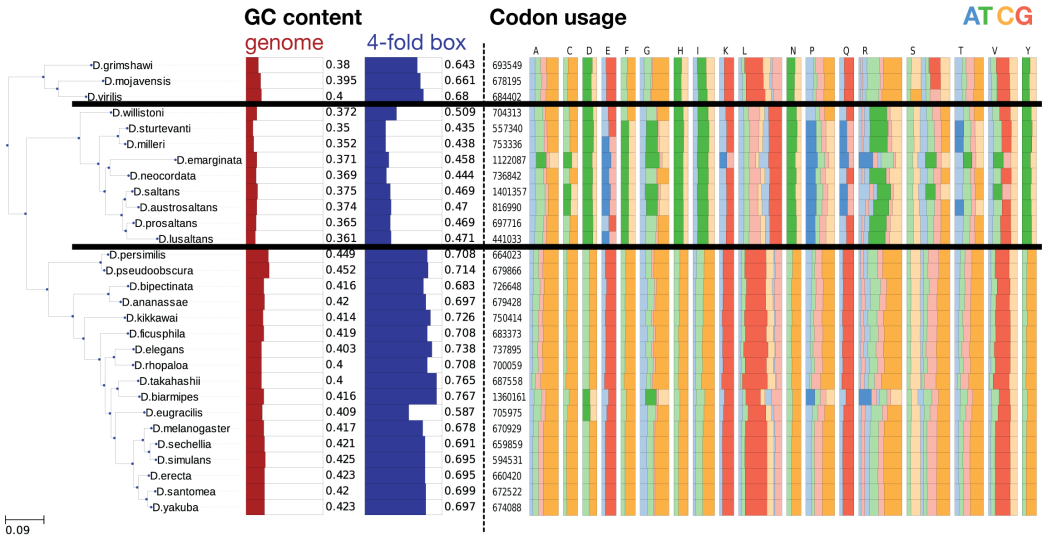


Figure 3.12: Phylogenetic tree of the 29 *Drosophila* species with the GC content at genomic level and in coding sequences (3rd position of 4-fold box codons). The codon usage table shows the proportion of codons observed for each amino acid (columns). Codons are colored according to its ending nucleotide. (Plot provided by Marco Mariotti).

codon bias. Indeed, the genomic GC content of all species belonging to the *willistoni/saltans* lineage is lower than any other *Drosophila*. The GC content in coding sequences is also lower and exhibits a much bigger difference, almost 2-fold (figure 3.12).

When codon bias is considered, the *willistoni/saltans* group again appears homogeneous, and different from the rest of drosophila. The relative synonymous codon usage (RSCU) is a measure for each codon, and it quantifies how much this codon is overrepresented comparing to neutral expectations (all synonymous codons with equal frequency). RSCU can pinpoint the differences in usage for each codon. The preferred codons for many amino acids changed in this lineage favouring A or T ending codons (figures 3.12 and 3.13).

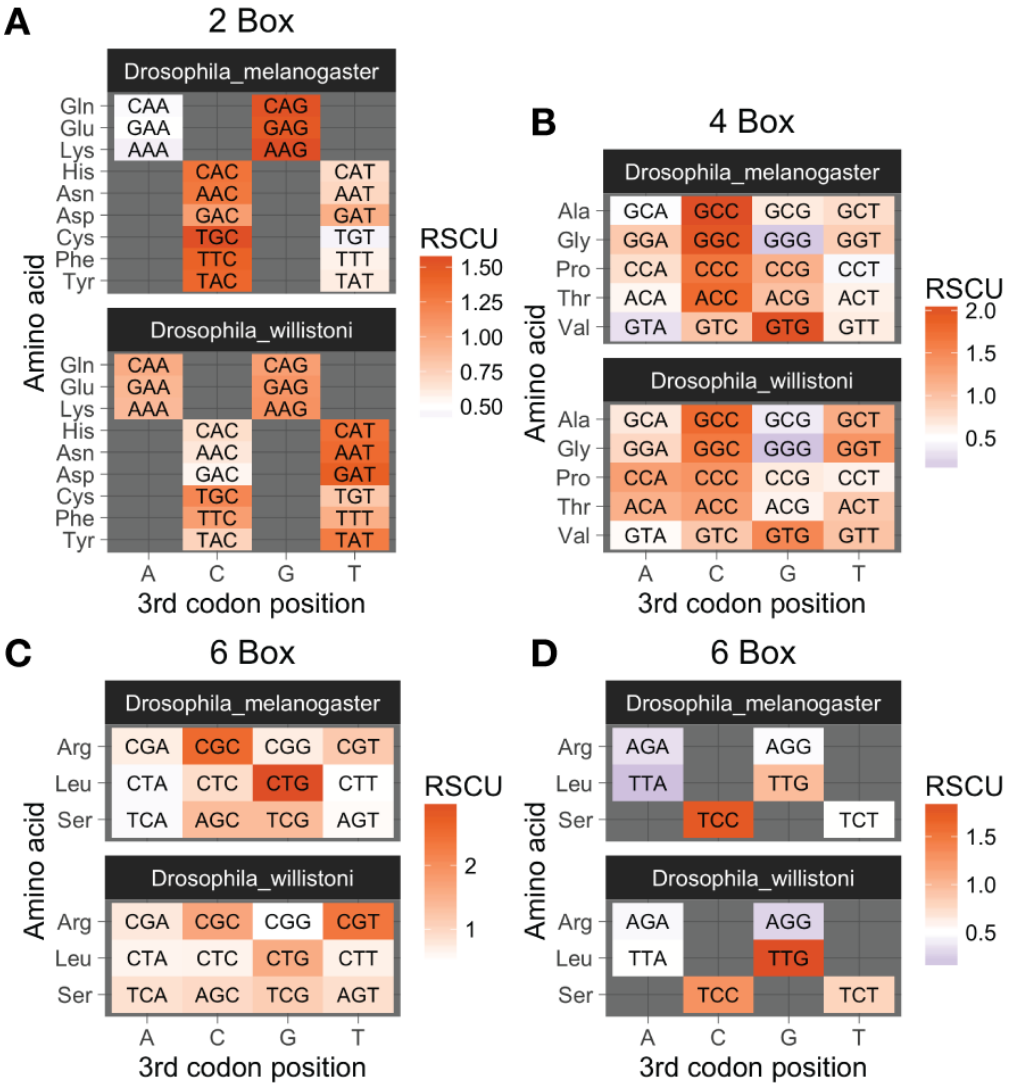


Figure 3.13: Relative synonymous codon usage (RSCU) in *D. melanogaster* vs *D. willistoni*, in 1160 one-to-one orthologues. Each coloured cell corresponds to one codon. Codons with higher RSCU (red) are more frequently used. In each of the panels, the rows corresponds to the different amino acids, and the columns correspond to the 3rd position of the codon. A) twofold degenerate codons; B) fourfold degenerate codons; C) and D) sixfold degenerate codons. Note that C and D correspond to the same three amino acids.

3.6.4 Widening the picture: other arthropods

In order to put the *Drosophila* Sec extinctions in context, we investigated other insect and non-insect arthropod species. All analyzed genomes were downloaded from NCBI, and scanned with Selenoprofiles [Mariotti and Guigó, 2010]. $tRNA^{Sec}$ predictions were obtained using Secmarker (section 2.1 in this thesis). Figure 3.14 shows a summary of the results. In accordance with our previous results on fewer species, all organisms belonging to *Hymenoptera* (sawflies, wasps, bees, and ants) and *Lepidoptera* (butterflies and moths) showed no intact selenoprotein genes, and also lacked a complete machinery. Among *Coleoptera* (beetles) instead, not all genomes lacked selenoproteins, as previously thought. Two selenoproteins, and a complete Sec machinery were found in *Onthophagus taurus* (taurus scarab). Phylogenetically, *O. taurus* is placed basal to the other coleopterans analyzed so far, which suggests that the Sec loss described in *Coleoptera* is more recent than previously thought (figure 3.14). No selenoproteins other than those observed in *D. melanogaster* could be found among *Diptera* (flies and mosquitoes).

Other eukaryotic families were found as selenoproteins in *Paraneoptera*, with genomes of *Pediculus humanus* (human louse), *Rhodnius prolixus* (kissing bug), *Acyrthosiphum pisum* (pea aphid) and *Diaphorina citri* (the citrus psyllid). Among them only pea aphid lacks selenoproteins [International Aphid Genomics Consortium, 2010]. *P. humanus* possesses a rich selenoproteome, including three important antioxidant selenoprotein families: glutathione peroxidase (GPx), thioredoxin reductase (TR) and methionine-S-sulfoxide reductase (SelR). Notably, GPx in insects has only been observed as a selenoprotein in Paraneoptera, other insects only have cysteine based GPx enzymes. *R. prolixus* Sec-containing GPx was analyzed in [Dias et al., 2016] (publication included in section 3.2 of this thesis). Walking away from *Drosophila*, *Ladona fulva* (dragonfly, order *Odonata*) showed the richest selenoproteome among insects. Sec forms of protein families SelH, GPx, SelT, SelR, SelW, SelU and TR were found.

Regarding non-insect arthropods, the same selenoprotein families, plus others, were found. For example, *Crustacea* (*Daphnia pulex*, *Eurytemora affinis*, *Lepeophtheirus salmonis*) and *Myriapoda* (*Strigamia maritima*) possess a very rich selenoproteome, quite similar to the vertebrate one. In one or both these subphylums, we found Sec forms for 16 selenoprotein families. Our analyses on

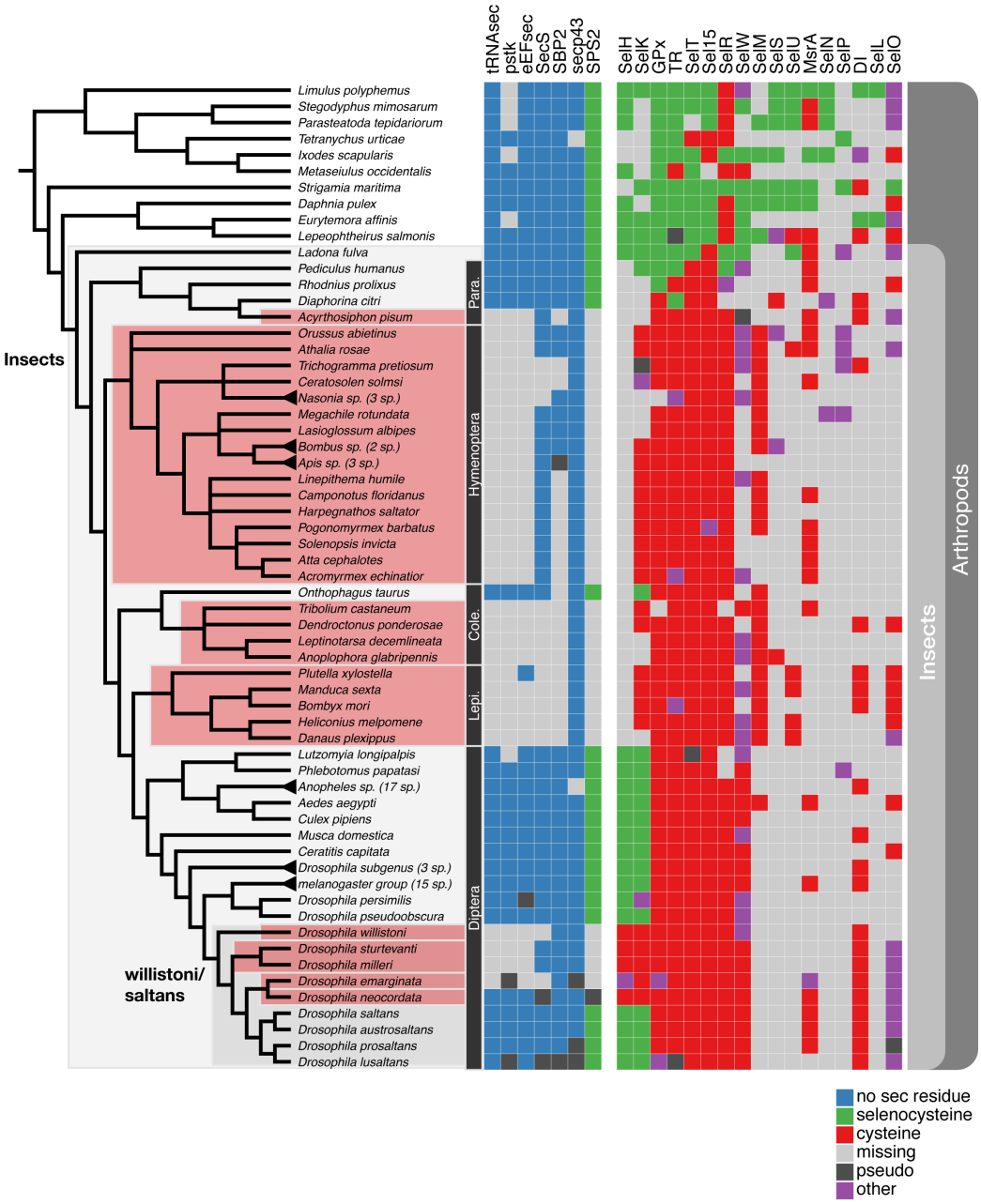


Figure 3.14: Selenoprotein extinctions in insects. Phylogenetic tree of arthropods annotated with prediction selenoprotein genes and Sec machinery factors. The red shaded boxed correspond to selenoprotein extinctions.

Arachnida, however, deserve a special mention, for it revealed a novel Sec loss event. Among the 10 *Acari* genome (mites and ticks), those of *Sarcoptes scabiei* and *Dermatophagoides farinae* lacked selenoproteins, and the Sec machinery genes. The two species cluster together phylogenetically, and are the only two representatives of the order *Astigmata* (mites). Incidentally, the genome of *Rhizoglyphus robini* (bulb mite) is currently being sequenced by the group of Fyodor Kondrashov, in our department. We analyzed the assembly version 2 (Mateusz Konczal, personal communication), and consistent with the other two genomes, no selenoproteins nor Sec machinery could be detected in *R. robini* (figure 3.15). We speculate that selenoproteins were lost in *Astigmata*, or possibly in the lineage leading to the common ancestor of these three genomes. Unlike selenoprotein-less insects (*Astigmata* are not insects), these three genomes don't have the *SPS1* gene, a non-selenoprotein paralogue of *SPS2*. That would be similar to what was observed in nematodes (nematodes don't have SPS1), but in that case, SPS2 was converted to a Cys homologue first (the only known case among animals) and subsequently lost in some plant parasitic lineages [Otero et al., 2014]. *SPS1* was predicted to appear by gene duplication at the root of insects (and in other lineages independently), presumably in a subfunctionalization event that relocated two different functions carried by the ancestral *SPS2* [Mariotti et al., 2015].

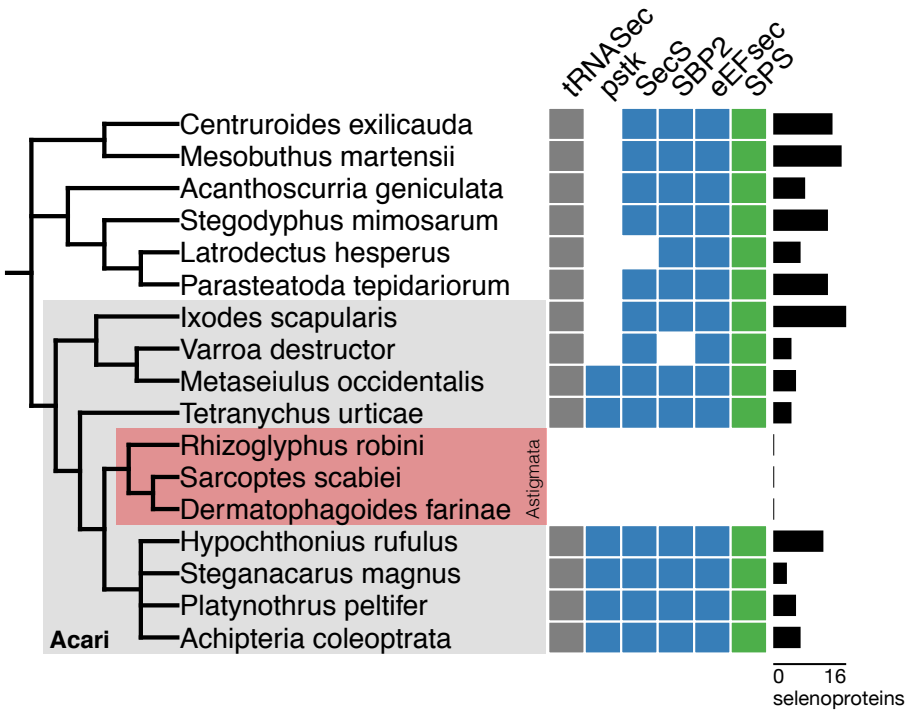


Figure 3.15: Phylogenetic tree of *Arachinda* genomes used in this work. The colored cells correspond to the Sec machinery factors (green: selenoprotein; blue: no sec residue; grey: $tRNA^{Sec}$). The number of selenoproteins in indicated by the horizontal black bars. In the three genomes shaded in red (*Astigmata*) no selenoproteins and no traces of Sec machinery were found.

Chapter 4

DISCUSSION

4.1 Prediction of tRNA^{Sec}

Researchers interested in tRNA^{Sec} traditionally had suffered the lack of a specific method for tRNA^{Sec} prediction. Due to its unusual structure, general tRNA detection programs fail to accurately predict tRNA^{Sec} in genomes, with high false positive and false negative rates. For that reason, identification of tRNA^{Sec} has barely been addressed before, and tRNA databases contain very few tRNA^{Sec} entries (only 46 among the more than 12,000 canonical tRNAs in tRNAdb [**?**]). Very few studies had addressed the evolution of tRNA^{Sec}, and possibly one of the reasons is the lack of sequences to analyze.

Another issue is the characterization of the Sec utilization trait, which has been mainly based on the prediction of the protein factors of the Sec machinery. For different reasons, it is not trivial to find a good protein coding genetic marker among the Sec machinery factors: SPS (or SelD) in bacteria is used for other pathways that require selenium [Lin et al., 2015, Mariotti et al., 2015, Peng et al., 2016], independently of selenocysteine; the predictions of SPS (SPS2) in eukaryotes can be confused by the presence of the non-selenoprotein paralogue SPS1 in several animal lineages (and to make it even worse, the puzzling *Hymenoptera* SPS1 gene has a non-Sec UGA in-frame codon [Mariotti et al., 2015]); SBP2 and SecS appeared to be present in many selenoproteinless insects (see figure 3.14). Apart from that, proteins a made up of domains, and particular domains can be conserved in distant homologues (eEFSec and SelB have sequence homology with

the canonical elongation factor) which can also complicate its prediction. $tRNA^{Sec}$ instead, as we showed in the benchmark included in the Secmarker manuscript (section 2.1 in this thesis), is the perfect marker for the use of selenocysteine in an organism. An additional benefit of the identification of $tRNA^{Sec}$ is the possibility to distinguish the domain (archaea, bacteria or eukaryote) by its secondary structure, which is particularly useful in the analysis of metagenomes, where the source organisms is not known. Secmarker can be used to quickly scan newly sequenced genomes, and can help to allocate the resources to identify selenoproteins genes only when needed. For example, the presence of an archaeal $tRNA^{Sec}$ in the *Lokiarchaeota* assembly [Mariotti et al., 2016], was a strong indication of the presence of selenoproteins.

With Secmarker, we contributed with a computational tool for accurate $tRNA^{Sec}$ identification, filling a gap in the field of the study of selenocysteine and selenoproteins.

4.2 Evolution of SECIS elements

A long-standing questions in the the evolution of selenoproteins is the link between the selenoprotein synthesis machinery in the three domains of life. Sec is used by organisms from the three lines of descent through recoding of a UGA codon. The process of Sec synthesis and insertion is conserved across living organisms. Sec was probably present in the last universal common ancestor, although many organisms lost the ability to use it.

Some important differences exist in the selenoprotein system of the different domains, but the evolutionary relationship of the three systems is not well understood. The analysis of selenoprotein genes in *Lokiarchaeota*, the closest archaeal relative to eukaryotes known to date, shed some light to the origin of the eukaryotic SECIS. The main characteristic features of the eukaryotic SECIS are the presence of a kink-turn motif at the base of the upper stem (the core), and a stretch of adenines in the apical loop [Krol, 2002]. The SECIS elements in selenoprotein genes from *Lokiarchaeota* are eukaryotic-like, in the sense that they present the same conserved residues that form the core in the eukaryotic SECIS, and a strong preference for adenosine in the apical loop. These feature were found downstream every selenoprotein gene. In addition, further analysis revealed that

these eukaryotic-like features were also present in the already known archaeal selenoprotein gene *VhuD*. This work established that the characteristic fold of the eukaryotic SECIS was already present in archaea, and propagated through the eukaryotic lineage.

Some unsolved questions regarding the evolution of SECIS elements remain. Why the SECIS element migrated from the immediate proximity of the UGA codon in bacteria to the UTR region? In the work of [Krol, 2002] it was hypothesized that the coupled transcription/translation in bacteria makes it mandatory for the bSECIS to reside next to the UGA, with the burden of having to maintain both coding capacity and pairing ability, while the uncoupled transcription/translation in eukaryotes would enable the migration towards the 3'UTR releasing the constraints of the coding capacity. That is an interesting hypothesis, but the data at that time was already against it, because the prokaryotic archaeal SECIS was known to reside in the 3'UTR. Today we know that the eukaryotic SECIS is the direct descendant of the archaeal SECIS, so the system was already established before the acquisition of the nucleus by the cell. Krol continued by saying that the degree of freedom acquired by increasing the distance UGA-SECIS, the possibility of inserting multiple Sec residues appeared. It is true that the single SECIS element in the HdrA-VhuD gene tandem in *Lokiarchaeota* must act in multiple distant UGA codons, and that would be not possible with a single bacerial SECIS. However, in the majority of eukaryotic selenoproteins, the distant SECIS element acts in a single Sec residue. If having multiple Sec residues was beneficial for selenoproteins, and the insertion of multiple Sec residues was possible with a single distant SECIS, one would expect to see that trait more often in eukaryotes. The notable exception is selenoprotein P, which is well known for carrying multiple Sec residues (ten in human). But SelP requires the action of two SECIS elements for insertion of multiple selenocysteines [Tujebajeva et al., 2000, Shetty et al., 2014].

4.3 *willistoni/saltans* lineage

The *Drosophila* genomes from the *saltans* group sequenced in our group are a great resource for studies on molecular evolution. The peculiar genomic features of the members of this lineage can be used to address many questions. The causes

of the change in GC content and codon usage bias remain obscure to us. The analysis of tRNAs and tRNA modification enzymes could provide interesting insights in the evolution of the codon usage in *Drosophila*. Having whole-genome annotations and orthology assignment for several drosophilas, we devised possible methods for searching expansions or depletions of specific gene families, or with particular evolutionary patterns, that hopefully will help in understanding the nature of the genome catastrophes[1] that occurred at the root of this lineage.

4.4 Visualization of large phylogenies

We have now access to an unprecedented number of genome sequences, from highly diverse lineages. We have the proper tools and enough computational resources to analyze a large number of genomes. But one of the challenges is just looking at the results. The phylogenetic context of the results is critical to interpret them correctly and obtain meaningful conclusions. That can be achieved by visualization of phylogenetic trees. Many softwares exist for the construction and visualization of phylogenetic trees. But the ones we tried were not able to properly display and annotate large phylogenies. With ggsunburst (appendix **??**), the package I developed, we were able to visualize and annotate large trees (for example, more than 8,000 genomes in supplementary SM1.1 in [Mariotti et al., 2015]). One of the features of the package is the possibility to use the sunburst layout (http://www.cc.gatech.edu/gvu/ii/sunburst/), particularly useful for large phylogenies.

[1] Catastrophe, as used by Thom (1975), describes the sudden effects of gradual, continuously changing forces, often in ways which are not intuitively expected and which may seem quite radical.

Chapter 5

CONCLUSIONS

During my PhD I developed computational methods for the identification of selenoproteins and $tRNA^{Sec}$

- I improved the detection of $tRNA^{Sec}$ with Secmarker, and created a web server that can be used for online analysis. By applying Secmarker in more than 10,000 genomes, we obtained a precise map of the use of selenocysteine across the Tree of Life;

- I developed bSeblastian, a pipeline for the identification of bacterial selenoproteins based on the detection of bSECIS as first step. bSeblastian is able to identify known and novel selenoproteins.

By using these tools, and already existing ones, I contributed to selenoprotein research in several projects.

- I produced whole-genome selenoprotein gene annotations for 59 species, included in SelenoDB 2.0;

- I contributed with selenoprotein annotations in the genome projects of two bumble bees (*Bombus terrestris* and *Bombus impatiens*), and the insect vector of Chagas disease *Rhodnius prolixus*;

- I contributed in the study of the evolution of selenophosphate synthetases, which delineates an insightful story of function evolution;

- I contributed to the characterization of the selenoproteome of *Lokiarchaeota*. With Secmarker, we identified the first reported intron-containing $tRNA^{Sec}$;
- I characterized the abundance and distribution of selenoproteins in the human microbiota, as well as the other selenium utilization traits. bSeblastian allowed us to identify 4 novel candidate selenoprotein families.

Bibliography

[Altschul et al., 1997] Altschul, S. F., Madden, T. L., Schäffer, 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 research*, 25(17):3389–402.

[Ambrogelly et al., 2007] Ambrogelly, A., Palioura, S., and Söll, D. (2007). Natural expansion of the genetic code. *Nature chemical biology*, 3(1):29–35.

[Andreesen and Ljungdahl, 1973] Andreesen, J. R. and Ljungdahl, L. G. (1973). Formate dehydrogenase of Clostridium thermoaceticum: incorporation of selenium 75, and the effects of selenite, molybdate, and tungstate on the enzyme. *Journal of Bacteriology*, 116(2):867–873.

[Arbogast and Ferreiro, 2010] Arbogast, S. and Ferreiro, A. (2010). Selenoproteins and protection against oxidative stress: selenoprotein N as a novel player at the crossroads of redox signaling and calcium homeostasis. *Antioxidants & redox signaling*, 12(7):893–904.

[Arnér and Holmgren, 2000] Arnér, E. S. J. and Holmgren, A. (2000). Physiological functions of thioredoxin and thioredoxin reductase.

[Ballihaut et al., 2007] Ballihaut, G., Mounicou, S., and Lobinski, R. (2007). Multitechnique mass-spectrometric approach for the detection of bovine glutathione peroxidase selenoprotein: focus on the selenopeptide. *Analytical and bioanalytical chemistry*, 388(3):585–91.

[Baron et al., 1990] Baron, C., Heider, J., and Böck, A. (1990). Mutagenesis of selC, the gene for the selenocysteine-inserting tRNA-species in E. coli: effects on in vivo function. *Nucleic acids research*, 18(23):6761–6.

[Baron et al., 1993] Baron, C., Heider, J., and Böck, A. (1993). Interaction of translation factor SELB with the formate dehydrogenase H selenopolypeptide mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, 90(9):4181–4185.

[Behne et al., 1990] Behne, D., Kyriakopoulos, A., Meinhold, H., and K??hrle, J. (1990). Identification of type I iodothyronine 5deiodinase as a selenoenzyme. *Biochemical and Biophysical Research Communications*, 173(3):1143–1149.

[Berg et al., 1991] Berg, B. L., Baron, C., and Stewart, V. (1991). Nitrate-inducible formate dehydrogenase in Escherichia coli K-12. II. Evidence that a mRNA stem-loop structure is essential for decoding opal (UGA) as selenocysteine. *J. Biol. Chem.*, 266(33):22386–22391.

[Berry et al., 1991] Berry, M. J., Banu, L., Chen, Y. Y., Mandel, S. J., Kieffer, J. D., Harney, J. W., and Larsen, P. R. (1991). Recognition of UGA as a selenocysteine codon in type I deiodinase requires sequences in the 3' untranslated region. *Nature*, 353(6341):273–6.

[Birringer et al., 2002] Birringer, M., Pilawa, S., and Flohé, L. (2002). Trends in selenium biochemistry. *Natural product reports*, 19(6):693–718.

[Castellano et al., 2005] Castellano, S., Lobanov, A. V., Chapple, C., Novoselov, S. V., Albrecht, M., Hua, D., Lescure, A., Lengauer, T., Krol, A., Gladyshev, V. N., and Guigó, R. (2005). Diversity and functional plasticity of eukaryotic selenoproteins: identification and characterization of the SelJ family. *Proceedings of the National Academy of Sciences of the United States of America*, 102(45):16188–16193.

[Castellano et al., 2001] Castellano, S., Morozova, N., Morey, M., Berry, M. J., Serras, F., Corominas, M., and Guigó, R. (2001). In silico identification of novel selenoproteins in the Drosophila melanogaster genome. *EMBO reports*, 2(8):697–702.

[Castellano et al., 2004] Castellano, S., Novoselov, S. V., Kryukov, G. V., Lescure, A., Blanco, E., Krol, A., Gladyshev, V. N., and Guigó, R. (2004). Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian

family with scattered phylogenetic distribution. *EMBO reports*, 5(December 2003):71–77.

[Chapple and Guigó, 2008] Chapple, C. E. and Guigó, R. (2008). Relaxation of selective constraints causes independent selenoprotein extinction in insect genomes. *PloS one*, 3(8):e2968.

[Chiba et al., 2010] Chiba, S., Itoh, Y., Sekine, S.-i., and Yokoyama, S. (2010). Structural basis for the major role of O-phosphoseryl-tRNA kinase in the UGA-specific encoding of selenocysteine. *Molecular cell*, 39(3):410–20.

[Chivers et al., 1997] Chivers, P. T., Prehoda, K. E., and Raines, R. T. (1997). The CXXC motif: A rheostat in the active site. *Biochemistry*, 36(14):4061–4066.

[Clark et al., 2007] Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., Pollard, D. A., Sackton, T. B., Larracuente, A. M., Singh, N. D., Abad, J. P., Abt, D. N., Adryan, B., Aguade, M., Akashi, H., Anderson, W. W., Aquadro, C. F., Ardell, D. H., Arguello, R., Artieri, C. G., Barbash, D. A., Barker, D., Barsanti, P., Batterham, P., Batzoglou, S., Begun, D., Bhutkar, A., Blanco, E., Bosak, S. A., Bradley, R. K., Brand, A. D., Brent, M. R., Brooks, A. N., Brown, R. H., Butlin, R. K., Caggese, C., Calvi, B. R., Bernardo de Carvalho, A., Caspi, A., Castrezana, S., Celniker, S. E., Chang, J. L., Chapple, C., Chatterji, S., Chinwalla, A., Civetta, A., Clifton, S. W., Comeron, J. M., Costello, J. C., Coyne, J. A., Daub, J., David, R. G., Delcher, A. L., Delehaunty, K., Do, C. B., Ebling, H., Edwards, K., Eickbush, T., Evans, J. D., Filipski, A., Findeiss, S., Freyhult, E., Fulton, L., Fulton, R., Garcia, A. C. L., Gardiner, A., Garfield, D. A., Garvin, B. E., Gibson, G., Gilbert, D., Gnerre, S., Godfrey, J., Good, R., Gotea, V., Gravely, B., Greenberg, A. J., Griffiths-Jones, S., Gross, S., Guigo, R., Gustafson, E. A., Haerty, W., Hahn, M. W., Halligan, D. L., Halpern, A. L., Halter, G. M., Han, M. V., Heger, A., Hillier, L., Hinrichs, A. S., Holmes, I., Hoskins, R. A., Hubisz, M. J., Hultmark, D., Huntley, M. A., Jaffe, D. B., Jagadeeshan, S., Jeck, W. R., Johnson, J., Jones, C. D., Jordan, W. C., Karpen, G. H., Kataoka, E., Keightley, P. D., Kheradpour, P., Kirkness, E. F., Koerich, L. B., Kristiansen, K., Kudrna, D., Kulathinal, R. J., Kumar, S., Kwok, R., Lander, E., Langley, C. H., Lapoint, R., Lazzaro,

B. P., Lee, S.-J., Levesque, L., Li, R., Lin, C.-F., Lin, M. F., Lindblad-Toh, K., Llopart, A., Long, M., Low, L., Lozovsky, E., Lu, J., Luo, M., Machado, C. A., Makalowski, W., Marzo, M., Matsuda, M., Matzkin, L., McAllister, B., McBride, C. S., McKernan, B., McKernan, K., Mendez-Lago, M., Minx, P., Mollenhauer, M. U., Montooth, K., Mount, S. M., Mu, X., Myers, E., Negre, B., Newfeld, S., Nielsen, R., Noor, M. A. F., O'Grady, P., Pachter, L., Papaceit, M., Parisi, M. J., Parisi, M., Parts, L., Pedersen, J. S., Pesole, G., Phillippy, A. M., Ponting, C. P., Pop, M., Porcelli, D., Powell, J. R., Prohaska, S., Pruitt, K., Puig, M., Quesneville, H., Ram, K. R., Rand, D., Rasmussen, M. D., Reed, L. K., Reenan, R., Reily, A., Remington, K. A., Rieger, T. T., Ritchie, M. G., Robin, C., Rogers, Y.-H., Rohde, C., Rozas, J., Rubenfield, M. J., Ruiz, A., Russo, S., Salzberg, S. L., Sanchez-Gracia, A., Saranga, D. J., Sato, H., Schaeffer, S. W., Schatz, M. C., Schlenke, T., Schwartz, R., Segarra, C., Singh, R. S., Sirot, L., Sirota, M., Sisneros, N. B., Smith, C. D., Smith, T. F., Spieth, J., Stage, D. E., Stark, A., Stephan, W., Strausberg, R. L., Strempel, S., Sturgill, D., Sutton, G., Sutton, G. G., Tao, W., Teichmann, S., Tobari, Y. N., Tomimura, Y., Tsolas, J. M., Valente, V. L. S., Venter, E., Venter, J. C., Vicario, S., Vieira, F. G., Vilella, A. J., Villasante, A., Walenz, B., Wang, J., Wasserman, M., Watts, T., Wilson, D., Wilson, R. K., Wing, R. A., Wolfner, M. F., Wong, A., Wong, G. K.-S., Wu, C.-I., Wu, G., Yamamoto, D., Yang, H.-P., Yang, S.-P., Yorke, J. A., Yoshida, K., Zdobnov, E., Zhang, P., Zhang, Y., Zimin, A. V., Baldwin, J., Abdouelleil, A., Abdulkadir, J., Abebe, A., Abera, B., Abreu, J., Acer, S. C., Aftuck, L., Alexander, A., An, P., Anderson, E., Anderson, S., Arachi, H., Azer, M., Bachantsang, P., Barry, A., Bayul, T., Berlin, A., Bessette, D., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Bourzgui, I., Brown, A., Cahill, P., Channer, S., Cheshatsang, Y., Chuda, L., Citroen, M., Collymore, A., Cooke, P., Costello, M., D'Aco, K., Daza, R., De Haan, G., DeGray, S., DeMaso, C., Dhargay, N., Dooley, K., Dooley, E., Doricent, M., Dorje, P., Dorjee, K., Dupes, A., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Fisher, S., Foley, C. D., Franke, A., Friedrich, D., Gadbois, L., Gearin, G., Gearin, C. R., Giannoukos, G., Goode, T., Graham, J., Grandbois, E., Grewal, S., Gyaltsen, K., Hafez, N., Hagos, B., Hall, J., Henson, C., Hollinger, A., Honan, T., Huard, M. D., Hughes, L., Hurhula, B., Husby, M. E., Kamat, A., Kanga, B., Kashin, S., Khazanovich, D., Kisner, P., Lance,

K., Lara, M., Lee, W., Lennon, N., Letendre, F., LeVine, R., Lipovsky, A., Liu, X., Liu, J., Liu, S., Lokyitsang, T., Lokyitsang, Y., Lubonja, R., Lui, A., MacDonald, P., Magnisalis, V., Maru, K., Matthews, C., McCusker, W., McDonough, S., Mehta, T., Meldrim, J., Meneus, L., Mihai, O., Mihalev, A., Mihova, T., Mittelman, R., Mlenga, V., Montmayeur, A., Mulrain, L., Navidi, A., Naylor, J., Negash, T., Nguyen, T., Nguyen, N., Nicol, R., Norbu, C., Norbu, N., Novod, N., O'Neill, B., Osman, S., Markiewicz, E., Oyono, O. L., Patti, C., Phunkhang, P., Pierre, F., Priest, M., Raghuraman, S., Rege, F., Reyes, R., Rise, C., Rogov, P., Ross, K., Ryan, E., Settipalli, S., Shea, T., Sherpa, N., Shi, L., Shih, D., Sparrow, T., Spaulding, J., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Strader, C., Tesfaye, S., Thomson, T., Thoulutsang, Y., Thoulutsang, D., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Young, G., Yu, Q., Zembek, L., Zhong, D., Zimmer, A., Zwirko, Z., Alvarez, P., Brockman, W., Butler, J., Chin, C., Grabherr, M., Kleber, M., Mauceli, E., and MacCallum, I. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–18.

[Cone et al., 1976] Cone, J. E., Del Río, R. M., Davis, J. N., and Stadtman, T. C. (1976). Chemical characterization of the selenoprotein component of clostridial glycine reductase: identification of selenocysteine as the organoselenium moiety. *Proceedings of the National Academy of Sciences of the United States of America*, 73(8):2659–2663.

[Copeland et al., 2000] Copeland, P. R., Fletcher, J. E., Carlson, B. A., Hatfield, D. L., and Driscoll, D. M. (2000). A novel RNA binding protein, SBP2, is required for the translation of mammalian selenoprotein mRNAs. *The EMBO journal*, 19(2):306–14.

[Dayer et al., 2008] Dayer, R., Fischer, B. B., Eggen, R. I. L., and Lemaire, S. D. (2008). The peroxiredoxin and glutathione peroxidase families in Chlamydomonas reinhardtii. *Genetics*, 179(1):41–57.

[Dias et al., 2016] Dias, F. A., Gandara, A. C. P., Perdomo, H. D., Gonçalves, R. S., Oliveira, C. R., Oliveira, R. L. L., Citelli, M., Polycarpo, C. R., Santesmasses, D., Mariotti, M., Guigó, R., Braz, G. R., Missirlis, F., and Oliveira,

P. L. (2016). Identification of a selenium-dependent glutathione peroxidase in the blood-sucking insect Rhodnius prolixus. *Insect biochemistry and molecular biology*, 69:105–14.

[Dikiy et al., 2007] Dikiy, A., Novoselov, S. V., Fomenko, D. E., Sengupta, A., Carlson, B. A., Cerny, R. L., Ginalski, K., Grishin, N. V., Hatfield, D. L., and Gladyshev, V. N. (2007). SelT, SelW, SelH, and Rdx12: genomics and molecular insights into the functions of selenoproteins of a novel thioredoxin-like family. *Biochemistry*, 46(23):6871–82.

[Eddy and Durbin, 1994] Eddy, S. R. and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088.

[Ellis and Brown, 2009] Ellis, J. C. and Brown, J. W. (2009). The RNase P family. *RNA biology*, 6(4):362–369.

[Fagegaltier et al., 2000] Fagegaltier, D., Hubert, N., Yamada, K., Mizutani, T., Carbon, P., and Krol, A. (2000). Characterization of mSelB, a novel mammalian elongation factor for selenoprotein translation. *The EMBO journal*, 19(17):4796–805.

[Fletcher et al., 2001] Fletcher, J. E., Copeland, P. R., Driscoll, D. M., and Krol, A. (2001). The selenocysteine incorporation machinery: interactions between the SECIS RNA and the SECIS-binding protein SBP2. *RNA (New York, N.Y.)*, 7(10):1442–53.

[Flohe et al., 1973] Flohe, L., Günzler, W. A., and Schock, H. H. (1973). Glutathione peroxidase: a selenoenzyme. *FEBS letters*, 32(1):132–134.

[Forchhammer et al., 1989] Forchhammer, K., Leinfelder, W., and Böck, A. (1989). Identification of a novel translation factor necessary for the incorporation of selenocysteine into protein. *Nature*, 342(6248):453–6.

[Forchhammer et al., 1990] Forchhammer, K., Rucknagel, K. P., and Bock, A. (1990). Purification and biochemical characterization of SELB, a translation factor involved in selenoprotein synthesis. *Journal of Biological Chemistry*, 265(16):9346–9350.

[Glass et al., 1993] Glass, R. S., Singh, W. P., Jung, W., Veres, Z., Scholz, T. D., and Stadtman, T. C. (1993). Monoselenophosphate: synthesis, characterization, and identity with the prokaryotic biological selenium donor, compound SePX. *Biochemistry*, 32(47):12555–9.

[Gobler et al., 2011] Gobler, C. J., Berry, D. L., Dyhrman, S. T., Wilhelm, S. W., Salamov, A., Lobanov, A. V., Zhang, Y., Collier, J. L., Wurch, L. L., Kustka, A. B., Dill, B. D., Shah, M., VerBerkmoes, N. C., Kuo, A., Terry, A., Pangilinan, J., Lindquist, E. A., Lucas, S., Paulsen, I. T., Hattenrath-Lehmann, T. K., Talmage, S. C., Walker, E. A., Koch, F., Burson, A. M., Marcoval, M. A., Tang, Y.-Z., Lecleir, G. R., Coyne, K. J., Berg, G. M., Bertrand, E. M., Saito, M. A., Gladyshev, V. N., and Grigoriev, I. V. (2011). Niche of harmful alga Aureococcus anophagefferens revealed through ecogenomics. *Proceedings of the National Academy of Sciences of the United States of America*, 108(11):4352–7.

[Gonzalez-Flores et al., 2012] Gonzalez-Flores, J. N., Gupta, N., DeMong, L. W., and Copeland, P. R. (2012). The selenocysteine-specific elongation factor contains a novel and multi-functional domain. *Journal of Biological Chemistry*, 287(46):38936–38945.

[Griffiths-Jones, 2005] Griffiths-Jones, S. (2005). RALEE–RNA ALignment editor in Emacs. *Bioinformatics (Oxford, England)*, 21(2):257–9.

[Gromer et al., 2005] Gromer, S., Eubel, J. K., Lee, B. L., and Jacob, J. (2005). Human selenoproteins at a glance. *Cell. Mol. Life Sci*, 6205(21):2414–2437.

[Guigó et al., 1992] Guigó, R., Knudsen, S., Drake, N., and Smith, T. (1992). Prediction of gene structure. *Journal of Molecular Biology*, 226(1):141–157.

[Gupta et al., 2013] Gupta, N., Demong, L. W., Banda, S., and Copeland, P. R. (2013). Reconstitution of selenocysteine incorporation reveals intrinsic regulation by SECIS elements. *Journal of Molecular Biology*, 425(14):2415–2422.

[Hüttenhofer and Böck, 1998] Hüttenhofer, A. and Böck, A. (1998). RNA Structures Involved in Selenoprotein Synthesis. *Cold Spring Harbor Monograph Archive*, 35(0).

[International Aphid Genomics Consortium, 2010] International Aphid Genomics Consortium (2010). Genome sequence of the pea aphid Acyrthosiphon pisum. *PLoS biology*, 8(2):e1000313.

[Itoh et al., 2009] Itoh, Y., Chiba, S., Sekine, S.-I., and Yokoyama, S. (2009). Crystal structure of human selenocysteine tRNA. *Nucleic acids research*, 37(18):6259–68.

[Itoh et al., 2013] Itoh, Y., Sekine, S.-i., Suetsugu, S., and Yokoyama, S. (2013). Tertiary structure of bacterial selenocysteine tRNA. *Nucleic acids research*, 41(13):6729–38.

[Katoh et al., 2002] Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14):3059–3066.

[Kim, 2013] Kim, H.-Y. (2013). The methionine sulfoxide reduction system: selenium utilization and methionine sulfoxide reductase enzymes and their functions. *Antioxidants & redox signaling*, 19(9):958–69.

[Kim and Gladyshev, 2007] Kim, H.-Y. and Gladyshev, V. N. (2007). Methionine sulfoxide reductases: selenoprotein forms and roles in antioxidant protein repair in mammals. *The Biochemical journal*, 407(3):321–9.

[Krol, 2002] Krol, A. (2002). Evolutionarily different RNA motifs and RNA protein complexes to achieve selenoprotein synthesis. *Biochimie*, 84(8):765–774.

[Kromayer et al., 1996] Kromayer, M., Wilting, R., Tormay, P., and Böck, A. (1996). Domain structure of the prokaryotic selenocysteine-specific elongation factor SelB. *Journal of molecular biology*, 262(4):413–420.

[Kryukov et al., 2003] Kryukov, G. V., Castellano, S., Novoselov, S. V., Lobanov, A. V., Zehtab, O., Guigó, R., and Gladyshev, V. N. (2003). Characterization of mammalian selenoproteomes. *Science (New York, N.Y.)*, 300(5624):1439–43.

[Kryukov and Gladyshev, 2004] Kryukov, G. V. and Gladyshev, V. N. (2004). The prokaryotic selenoproteome. *EMBO reports*, 5(5):538–43.

[Kryukov et al., 1999] Kryukov, G. V., Kryukov, V. M., and Gladyshev, V. N. (1999). New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *Journal of Biological Chemistry*, 274(48):33888–33897.

[Labunskyy et al., 2014] Labunskyy, V. M., Hatfield, D. L., and Gladyshev, V. N. (2014). Selenoproteins: molecular pathways and physiological roles. *Physiological reviews*, 94(3):739–77.

[Laslett and Canback, 2004] Laslett, D. and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, 32(1):11–16.

[Latrèche et al., 2009] Latrèche, L., Jean-Jean, O., Driscoll, D. M., and Chavatte, L. (2009). Novel structural determinants in human SECIS elements modulate the translational recoding of UGA as selenocysteine. *Nucleic Acids Research*, 37(17):5868–5880.

[Lescure et al., 1999] Lescure, A., Gautheret, D., Carbon, P., and Krol, A. (1999). Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif. *Journal of Biological Chemistry*, 274(53):38147–38154.

[Lesoon et al., 1997] Lesoon, A., Mehta, A., Singh, R., Chisolm, G. M., and Driscoll, D. M. (1997). An RNA-binding protein recognizes a mammalian selenocysteine insertion sequence element required for cotranslational incorporation of selenocysteine. *Molecular and cellular biology*, 17(4):1977–1985.

[Lewis et al., 2002] Lewis, S. E., Searle, S. M. J., Harris, N., Gibson, M., Lyer, V., Richter, J., Wiel, C., Bayraktaroglir, L., Birney, E., Crosby, M. A., Kaminker, J. S., Matthews, B. B., Prochnik, S. E., Smithy, C. D., Tupy, J. L., Rubin, G. M., Misra, S., Mungall, C. J., and Clamp, M. E. (2002). Apollo: a sequence annotation editor. *Genome biology*, 3(12):RESEARCH0082.

[Li and Godzik, 2006] Li, W. and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

[Lillig et al., 2008] Lillig, C. H., Berndt, C., and Holmgren, A. (2008). Glutaredoxin systems.

[Lin et al., 2015] Lin, J., Peng, T., Jiang, L., Ni, J.-Z., Liu, Q., Chen, L., and Zhang, Y. (2015). Comparative genomics reveals new candidate genes involved in selenium metabolism in prokaryotes. *Genome biology and evolution*, 7(3):664–76.

[Lobanov et al., 2006] Lobanov, A. V., Delgado, C., Rahlfs, S., Novoselov, S. V., Kryukov, G. V., Gromer, S., Hatfield, D. L., Becker, K., and Gladyshev, V. N. (2006). The Plasmodium selenoproteome. *Nucleic acids research*, 34(2):496–505.

[Lobanov et al., 2007] Lobanov, A. V., Fomenko, D. E., Zhang, Y., Sengupta, A., Hatfield, D. L., and Gladyshev, V. N. (2007). Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life. *Genome biology*, 8(9):R198.

[Lobanov et al., 2008] Lobanov, A. V., Hatfield, D. L., and Gladyshev, V. N. (2008). Selenoproteinless animals: selenophosphate synthetase SPS1 functions in a pathway unrelated to selenocysteine biosynthesis. *Protein science : a publication of the Protein Society*, 17(1):176–82.

[Lobanov et al., 2009] Lobanov, A. V., Hatfield, D. L., and Gladyshev, V. N. (2009). Eukaryotic selenoproteins and selenoproteomes. *Biochimica et biophysica acta*, 1790(11):1424–8.

[Lorenz et al., 2011] Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB*, 6(1):26.

[Lowe and Eddy, 1997] Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research*, 25(5):955–64.

[Mariotti, 2013] Mariotti, M. (2013). *Computational genomics of selenoproteins*. PhD thesis.

[Mariotti and Guigó, 2010] Mariotti, M. and Guigó, R. (2010). Selenoprofiles: profile-based scanning of eukaryotic genome sequences for selenoprotein genes. *Bioinformatics (Oxford, England)*, 26(21):2656–63.

[Mariotti et al., 2013] Mariotti, M., Lobanov, A. V., Guigo, R., and Gladyshev, V. N. (2013). SECISearch3 and Seblastian: new tools for prediction of SECIS elements and selenoproteins. *Nucleic acids research*, 41(15):e149.

[Mariotti et al., 2016] Mariotti, M., Lobanov, A. V., Manta, B., Santesmasses, D., Bofill, A., Gladyshev, V. N., and Programme, G. (2016). Lokiarchaeota marks the transition between the archaeal and eukaryotic selenocysteine encoding systems. *Molecular Biology and Evolution*.

[Mariotti et al., 2012] Mariotti, M., Ridge, P. G., Zhang, Y., Lobanov, A. V., Pringle, T. H., Guigo, R., Hatfield, D. L., and Gladyshev, V. N. (2012). Composition and evolution of the vertebrate and mammalian selenoproteomes. *PloS one*, 7(3):e33066.

[Mariotti et al., 2015] Mariotti, M., Santesmasses, D., Capella-Gutierrez, S., Mateo, A., Arnan, C., Johnson, R., D'Aniello, S., Yim, S. H., Gladyshev, V. N., Serras, F., Corominas, M., Gabaldón, T., Guigó, R., Gabaldon, T., and Guigo, R. (2015). Evolution of selenophosphate synthetases: emergence and relocation of function through independent duplications and recurrent subfunctionalization. *Genome research*, 25(9):1256–67.

[Mesquita et al., 2015] Mesquita, R. D., Vionette-Amaral, R. J., Lowenberger, C., Rivera-Pomar, R., Monteiro, F. A., Minx, P., Spieth, J., Carvalho, A. B., Panzera, F., Lawson, D., Torres, A. Q., Ribeiro, J. M. C., Sorgine, M. H. F., Waterhouse, R. M., Montague, M. J., Abad-Franch, F., Alves-Bezerra, M., Amaral, L. R., Araujo, H. M., Araujo, R. N., Aravind, L., Atella, G. C., Azambuja, P., Berni, M., Bittencourt-Cunha, P. R., Braz, G. R. C., Calderon-Fernandez, G., Carareto, C. M. A., Christensen, M. B., Costa, I. R., Costa, S. G., Dansa, M., Daumas-Filho, C. R. O., De-Paula, I. F., Dias, F. A., Dimopoulos, G., Emrich, S. J., Esponda-Behrens, N., Fampa, P., Fernandez-Medina, R. D., da Fonseca, R. N., Fontenele, M., Fronick, C., Fulton, L. A., Gandara, A. C., Garcia, E. S., Genta, F. A., Giraldo-Calderon, G. I., Gomes, B.,

Gondim, K. C., Granzotto, A., Guarneri, A. A., Guigo, R., Harry, M., Hughes, D. S. T., Jablonka, W., Jacquin-Joly, E., Juarez, M. P., Koerich, L. B., Latorre-Estivalis, J. M., Lavore, A., Lawrence, G. G., Lazoski, C., Lazzari, C. R., Lopes, R. R., Lorenzo, M. G., Lugon, M. D., Majerowicz, D., Marcet, P. L., Mariotti, M., Masuda, H., Megy, K., Melo, A. C. A., Missirlis, F., Mota, T., Noriega, F. G., Nouzova, M., Nunes, R. D., Oliveira, R. L. L., Oliveira-Silveira, G., Ons, S., Pagola, L., Paiva-Silva, G. O., Pascual, A., Pavan, M. G., Pedrini, N., Peixoto, A. A., Pereira, M. H., Pike, A., Polycarpo, C., Prosdocimi, F., Ribeiro-Rodrigues, R., Robertson, H. M., Salerno, A. P., Salmon, D., Santesmasses, D., Schama, R., Seabra-Junior, E. S., Silva-Cardoso, L., Silva-Neto, M. A. C., Souza-Gomes, M., Sterkel, M., Taracena, M. L., Tojo, M., Tu, Z. J., Tubio, J. M. C., Ursic-Bedoya, R., Venancio, T. M., Walter-Nuno, A. B., Wilson, D., Warren, W. C., Wilson, R. K., Huebner, E., Dotson, E. M., and Oliveira, P. L. (2015). Genome of Rhodnius prolixus, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. *Proceedings of the National Academy of Sciences*, 112(48):14936–14941.

[Morgenstern et al., 1996] Morgenstern, B., Dress, A., and Werner, T. (1996). Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 93(22):12098–12103.

[Mukai et al., 2016] Mukai, T., Englert, M., Tripp, H. J., Miller, C., Ivanova, N. N., Rubin, E. M., Kyrpides, N. C., and Söll, D. (2016). Facile Recoding of Selenocysteine in Nature. *Angewandte Chemie (International ed. in English)*.

[Nawrocki and Eddy, 2013] Nawrocki, E. P. and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics (Oxford, England)*, 29(22):2933–5.

[Novoselov et al., 2006] Novoselov, S. V., Hua, D., Lobanov, A. V., and Gladyshev, V. N. (2006). Identification and characterization of Fep15, a new selenocysteine-containing member of the Sep15 protein family. *The Biochemical journal*, 394(Pt 3):575–579.

[Novoselov et al., 2002] Novoselov, S. V., Rao, M., Onoshko, N. V., Zhi, H., Kryukov, G. V., Xiang, Y., Weeks, D. P., Hatfield, D. L., and Gladyshev, V. N.

(2002). Selenoproteins and selenocysteine insertion system in the model plant cell system, Chlamydomonas reinhardtii. *EMBO Journal*, 21(14):3681–3693.

[Otero et al., 2014] Otero, L., Romanelli-Cedrez, L., Turanov, A. A., Gladyshev, V. N., Miranda-Vizuete, A., and Salinas, G. (2014). Adjustments, extinction, and remains of selenocysteine incorporation machinery in the nematode lineage. *RNA (New York, N.Y.)*, 20(7):1023–34.

[Palioura et al., 2009] Palioura, S., Sherrer, R. L., Steitz, T. A., Söll, D., and Simonovic, M. (2009). The human SepSecS-tRNASec complex reveals the mechanism of selenocysteine formation. *Science (New York, N.Y.)*, 325(5938):321–5.

[Pejchal and Ludwig, 2005] Pejchal, R. and Ludwig, M. L. (2005). Cobalamin-independent methionine synthase (MetE): a face-to-face double barrel that evolved by gene duplication. *PLoS biology*, 3(2):e31.

[Peng et al., 2016] Peng, T., Lin, J., Xu, Y.-Z., and Zhang, Y. (2016). Comparative genomics reveals new evolutionary and ecological patterns of selenium utilization in bacteria. *The ISME journal*.

[Powell et al., 2003] Powell, J. R., Sezzi, E., Moriyama, E. N., Gleason, J. M., and Caccone, A. (2003). Analysis of a shift in codon usage in Drosophila. *Journal of molecular evolution*, 57 Suppl 1:S214–25.

[Proctor, 2011] Proctor, L. M. (2011). The Human Microbiome Project in 2011 and beyond. *Cell host & microbe*, 10(4):287–91.

[Rho et al., 2010] Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic acids research*, 38(20):e191.

[Romero et al., 2005] Romero, H., Zhang, Y., Gladyshev, V. N., and Salinas, G. (2005). Evolution of selenium utilization traits. *Genome biology*, 6(8):R66.

[Rother et al., 2000] Rother, M., Wilting, R., Commans, S., and Böck, A. (2000). Identification and characterisation of the selenocysteine-specific translation factor SelB from the archaeon Methanococcus jannaschii. *Journal of molecular biology*, 299(2):351–8.

[Saito and Takahashi, 2002] Saito, Y. and Takahashi, K. (2002). Characterization of selenoprotein P as a selenium supply protein. *European Journal of Biochemistry*, 269(22):5746–5751.

[Shchedrina et al., 2011] Shchedrina, V. A., Kabil, H., Vorbruggen, G., Lee, B. C., Turanov, A. A., Hirosawa-Takamori, M., Kim, H.-Y., Harshman, L. G., Hatfield, D. L., and Gladyshev, V. N. (2011). Analyses of fruit flies that do not express selenoproteins or express the mouse selenoprotein, methionine sulfoxide reductase B1, reveal a role of selenoproteins in stress resistance. *The Journal of biological chemistry*, 286(34):29449–61.

[Shchedrina et al., 2007] Shchedrina, V. A., Novoselov, S. V., Malinouski, M. Y., and Gladyshev, V. N. (2007). Identification and characterization of a selenoprotein family containing a diselenide bond in a redox motif. *Proceedings of the National Academy of Sciences of the United States of America*, 104(35):13919–13924.

[Sheppard et al., 2008] Sheppard, K., Yuan, J., Hohn, M. J., Jester, B., Devine, K. M., and Söll, D. (2008). From one amino acid to another: tRNA-dependent amino acid biosynthesis. *Nucleic acids research*, 36(6):1813–25.

[Sherrer et al., 2011] Sherrer, R. L., Araiso, Y., Aldag, C., Ishitani, R., Ho, J. M. L., Söll, D., and Nureki, O. (2011). C-terminal domain of archaeal O-phosphoseryl-tRNA kinase displays large-scale motion to bind the 7-bp D-stem of archaeal tRNA(Sec). *Nucleic acids research*, 39(3):1034–41.

[Shetty et al., 2014] Shetty, S. P., Shah, R., and Copeland, P. R. (2014). Regulation of selenocysteine incorporation into the selenium transport protein, selenoprotein P. *Journal of Biological Chemistry*, 289(36):25317–25326.

[Stenvall et al., 2011] Stenvall, J., Fierro-González, J. C., Swoboda, P., Saamarthy, K., Cheng, Q., Cacho-Valadez, B., Arnér, E. S. J., Persson, O. P., Miranda-Vizuete, A., and Tuck, S. (2011). Selenoprotein TRXR-1 and GSR-1 are essential for removal of old cuticle during molting in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America*, 108(3):1064–9.

[Stock and Rother, 2009] Stock, T. and Rother, M. (2009). Selenoproteins in Archaea and Gram-positive bacteria. *Biochimica et biophysica acta*, 1790(11):1520–32.

[Taskov et al., 2005] Taskov, K., Chapple, C., Kryukov, G. V., Castellano, S., Lobanov, A. V., Korotkov, K. V., Guigó, R., and Gladyshev, V. N. (2005). Nematode selenoproteome: the use of the selenocysteine insertion system to decode one codon in an animal genome? *Nucleic acids research*, 33(7):2227–38.

[The Human Microbiome Project Consortium, 2012] The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–14.

[Toppo et al., 2008] Toppo, S., Vanin, S., Bosello, V., and Tosatto, S. C. E. (2008). Evolutionary and structural insights into the multifaceted glutathione peroxidase (Gpx) superfamily. *Antioxidants & redox signaling*, 10(9):1501–14.

[Tujebajeva et al., 2000] Tujebajeva, R. M., Copeland, P. R., Xu, X. M., Carlson, B. A., Harney, J. W., Driscoll, D. M., Hatfield, D. L., and Berry, M. J. (2000). Decoding apparatus for eukaryotic selenocysteine insertion. *EMBO reports*, 1(2):158–163.

[Wilting et al., 1997] Wilting, R., Schorling, S., Persson, B. C., and Böck, A. (1997). Selenoprotein synthesis in archaea: identification of an mRNA element of Methanococcus jannaschii probably directing selenocysteine insertion. *Journal of molecular biology*, 266(4):637–641.

[Wright, 1990] Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene*, 87(1):23–9.

[Xu et al., 2007] Xu, X.-M., Carlson, B. A., Mix, H., Zhang, Y., Saira, K., Glass, R. S., Berry, M. J., Gladyshev, V. N., and Hatfield, D. L. (2007). Biosynthesis of selenocysteine on its tRNA in eukaryotes. *PLoS biology*, 5(1):e4.

[Zhang and Gladyshev, 2005] Zhang, Y. and Gladyshev, V. N. (2005). An algorithm for identification of bacterial selenocysteine insertion sequence elements and selenoprotein genes. *Bioinformatics (Oxford, England)*, 21(11):2580–9.

[Zhang and Gladyshev, 2008] Zhang, Y. and Gladyshev, V. N. (2008). Trends in selenium utilization in marine microbial world revealed through the analysis of the global ocean sampling (GOS) project. *PLoS genetics*, 4(6):e1000095.

[Zhang et al., 2006] Zhang, Y., Romero, H., Salinas, G., and Gladyshev, V. N. (2006). Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues. *Genome biology*, 7(10):R94.

[Zinoni et al., 1986] Zinoni, F., Birkmann, A., Stadtman, T. C., and Böck, A. (1986). Nucleotide sequence and expression of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 83(13):4650–4.

Appendix A

GGSUNBURST

Introduction

Many analyses produce data that can be organized into hierarchies. Biological data, from gene clustering to classification of organisms, is often represented in phylogenetic trees, a type of node-link diagrams. Other visualization techniques exist to leverage hierarchical structures. Adjacency diagrams are a space filling variant of node-link diagrams, where rather than drawing a link between parent and child in the hierarchy, nodes are drawn as solid areas (either arcs or bars), and their placement relative to adjacent nodes reveals their position in the hierarchy. Because the nodes are space-filling, they reveal an additional dimension that would be difficult to show in a node-link diagram.

`ggsunburst` is an R package for visualization of hierarchical data using adjacency diagrams (sunburst and icicle) and node-link diagrams (tree). It can be considered as an extension of ggplot (a popular R package for data visualization based on the Grammar of Graphics). This extension adds the functionality for working with phylogenetic trees, or other types of hierarchical data, and allows to annotate and visualize multiple attributes in the tree structure. ggsunburst can be installed following the instructions in http://genome.crg.cat/∼dsantesmasses/ggsunburst/.

ggsunburst

`ggsunburst` provides a set of tools for visualization of hierarchical data. The input is a tree structure in newick or NHX format, but it also accepts other types of input, from which the tree structure can be obtained.

Types of input are accepted:

- a tree in newick format;
- New Hampshire eXtended format (NHX). NHX is based on the New Hampshire (NH) standard (also called "Newick tree format");
- a delimiter-separated format file.

In order to load the input structure, the function `sunburst_data` runs python code under the hood. The function uses ETE, a python environment for tree exploration. The tree structure is traversed and their nodes are mapped into a cartesian coordinate system. The tree becomes a collection of coordinates that describe each of the nodes, stored as an R object, which is the output of the function `sunburst_data`. This function accepts the parameter `node_attributes` that can be used to add attributes to the nodes of the tree, to be used in the visualization.

Once the `sunburst_data` object is obtained, it can be passed to one of the three layout functions that are implemented for the visualization of the structure:

- `ggtree`: links or branches between parent and child nodes.
- `icicle`: the root node appears at the top, with child nodes underneath. Because the nodes are space-filling, it reveals an additional dimension that would be difficult to show in a node-link diagram
- `sunburst`: equivalent to the `icicle` layout, but in polar coordinates

The delimiter-separated format

The idea of this input format is to have an alternative to the newick format. The file is a flat text file where each row is a record, and each record contains several

fields separated by a delimiter. The tree structure is build from the information contained in this file (see the documentation below for details).

Node annotation

The easiest way to annotate the tree structure is to include the attributes in the input file. The NHX format accepts multiple attributes for the nodes of the tree, using the format of "`name=value`". The ETE package can be used to add attributes to a tree and get them in the NHX format (see node annotation section in http://etetoolkit.org/docs/latest/tutorial/tutorial_trees.html).

If the delimiter-separated file is used, attributes can also be added for any of the nodes (see the documentation below for details).

The list of attribute names to be displayed in the plot can be passed to the `node_attributes` parameter of `sunburst_data`. For example by using `sunburst_data('tree.nw', node_attributes=c('attr1', 'attr2'))`, `sunburst_data` will look for the value for each of this attribute names (`attr1` and `attr2`) in the nodes of the tree, and it will include the corresponding values in its output. `attr1` and `attr2` can now be directly accessed by the layout functions.

In fact, the plot generated by the layout functions is a ggplot object. That means the additional layers can be created using ggplot plethora of functions. The `sunburst_data` object contains all necessary coordinates that can be passed to the ggplot functions.

Documentation

The following pages include the ggsunburst documentation. The documentation is being maintained and updated at http://genome.crg.cat/∼dsantesmasses/ggsunburst/

ggsunburst

`ggsunburst` is an `R` package that offers a set of tools to plot **adjacency diagrams** and **trees** using `ggplot2`.

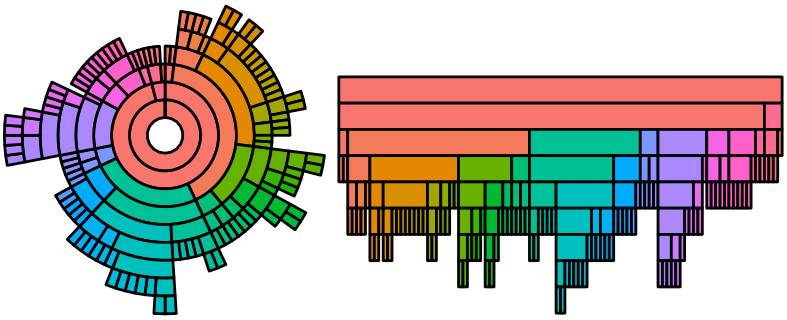
Adjacency diagrams are space-filling variants of node-link diagrams; rather than drawing a link between parent and child in the hierarchy, nodes are drawn as solid areas (either arcs or bars), and their placement relative to adjacent nodes reveals their position in the hierarchy.

`ggsunburst` uses ete2, a Python Environment for (phylogenetic) Tree Exploration, as built-in part of the package.

Overview

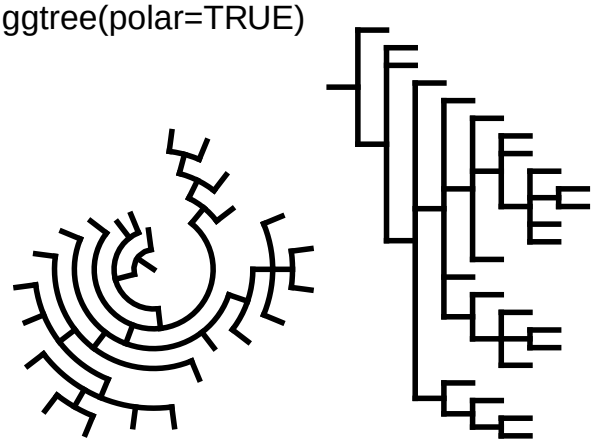
sunburst()

icicle()



ggtree()

ggtree(polar=TRUE)



Install

`ggsunburst` depends on the packages `ggplot2` and `rPython`. You will need to install them before using `ggsunburst`. Start a R session an type:

```
if (!require("ggplot2")) install.packages("ggplot2")
if (!require("rPython")) install.packages("rPython")
```

You can install the latest version of **ggsunburst** with

```
install.packages("http://genome.crg.es/~didac/ggsunburst/ggsunburst_0.0.6.tar.gz",
                 repos=NULL, type="source")
```

Once installed, you just need to load **ggsunburst** from your library:

```
library(ggsunburst)
```

Input

ggsunburst accepts two different formats as input:

- tree in newick format. It can be either a string or a file. The newick format is parsed by ete2, and many variants can be read: see reading-and-writing-newick-trees from ete2 documentation.

```
# newick format string
nw <- "(((D,F)B,(E,H)C)A);"
nw_print(nw)
```

```
##
##             /-D
##          /B|
##         |   \-F
## -NoNameA|
##         |   /-E
##          \C|
##             \-H
```

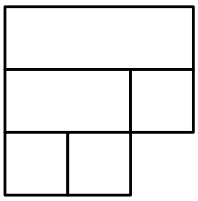
- delimiter-separated format. A flat file from which the tree structure can be obtained. See details

The function **sunburst_data** extracts relevant information from the underlying tree structure in the input and returns a list of data.frames

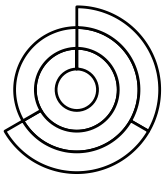
```
# extract data from the newick string defined above
sb <- sunburst_data(nw)
```

Layouts

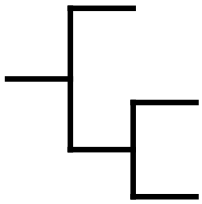
- `icicle`
  the root node appears at the top, with child nodes underneath. Because the nodes are space-filling, it reveals an additional dimension that would be difficult to show in a node-link diagram



- `sunburst`
  equivalent to the “icicle” layout, but in polar coordinates



- `ggtree`
  rectangular or circular tree



last update: Sun Jul 24 13:28:03 2016

Package developed at CRG, Barcelona.
Page build with knitr.
Didac Santesmasses, 2016.

delimiter-separated format

The idea of this input format is to have an alternative to the newick format. The delimiter-separated file is a flat text file where each row is a record, and each record contains several fields separated by a delimiter. This file contains the underlying tree structure that you want to represent. Two types are accepted:

node_parent

In the `type = "node_parent"`, each row contains one node and its parent. Consider the following tree

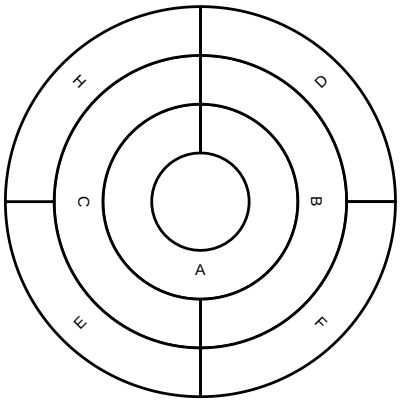
```
      /-D
   /B|
  |   \-F
-A|
  |   /-E
   \C|
      \-H
```

it can be defined by its nodes and the corresponding parent. The first line indicates what corresponds each column.

```
node,parent
D,B
F,B
B,A
E,C
H,C
C,A
```

Assuming the file "node_parent.csv" contains the lines above, it can be loaded with

```
sb <- sunburst_data("node_parent.csv", type = "node_parent", sep=",")
sunburst(sb, node_labels = T)
```



You can easily assign attributes to each of the nodes just by adding additional columns. Let's add a third column to assign the attribute "level" to each of the nodes

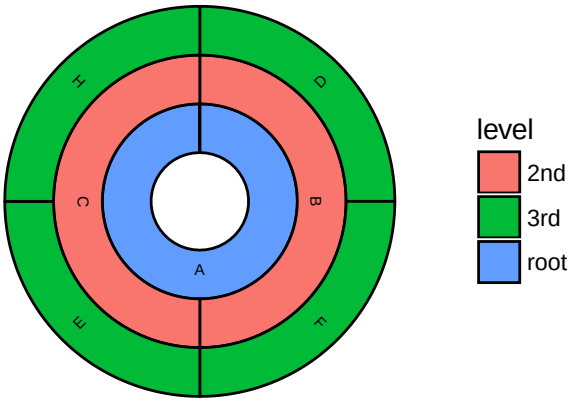
```
node,parent,level
D,B,3rd
F,B,3rd
B,A,2nd
E,C,3rd
H,C,3rd
C,A,2nd
A,,root
```

note that we have added an additional last line to assign a "level" to the node "A", the root of the tree. The root has no parent, and this is why the second column is empty. All rows must have the same number of columns. Now, these attributes can be included in the tree, you just need to specify them with the `node_attributes` parameter.

```
sb <- sunburst_data("node_parent.csv", type = "node_parent", sep=",", node_attributes = "level")
```

Let's call the `sunburst` funcion using the "level" information to assign a color to each of the nodes with the `rects.fill.aes` parameter

```
sunburst(sb, node_labels = T, rects.fill.aes = "level")
```



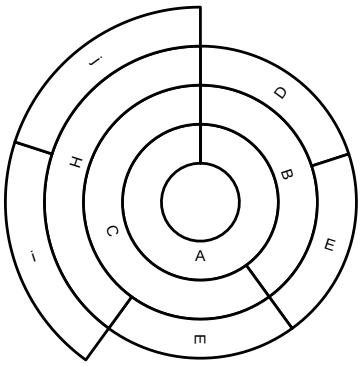
There is no limitation in the number of columns, you can add as many attributes as you need, but the fields `node` and `parent` are required.

lineage

In the `type = "lineage"`, each row in the input file represents a complete lineage, from root to terminal node

```
A,B,D
A,B,E
A,C,E
A,C,H,i
A,C,H,j
```

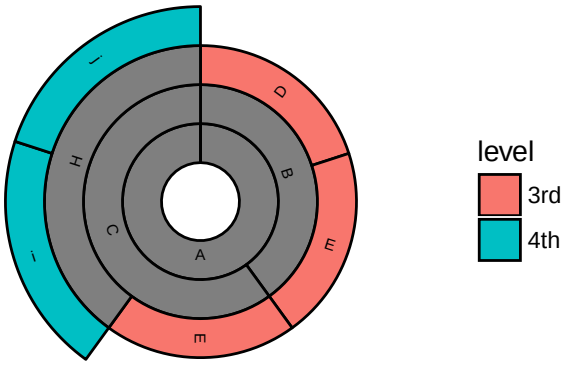
```
sb <- sunburst_data("lineage.csv", type = "lineage", sep=",")
sunburst(sb, node_labels = T)
```



You can also add attributes to the terminal nodes. The special delimiter `->` is used to separate the lineage and the attributes. The attributes have a `key:value` format

```
A,B,D->level:3rd
A,B,E->level:3rd
A,C,E->level:3rd
A,C,H,i->level:4th
A,C,H,j->level:4th
```

```
sb <- sunburst_data("lineage.csv", type = "lineage", sep=",", node_attributes = "level")
sunburst(sb, node_labels = T, rects.fill.aes = "level")
```



last update: Wed Jul 20 23:26:51 2016

Package developed at CRG, Barcelona.
Page build with knitr.
Didac Santesmasses, 2016.

ggsunburst Overview Layouts Introduction

sunburst_data

Usage:

```
sunburst_data(newick, ladderize = F, ultrametric = F, type='sunburst',
              xlim=360, rot=0, node_attributes='')
```

```
# demonstrate the use of sunburst_data output with ggplot
require(ggplot2)
ggplot() +
  geom_rect(data=sb$rects,
            aes(xmin=xmin, xmax=xmax, ymin=ymin, ymax=ymax), color="white") +
  geom_text(data=sb$leaf_labels, aes(x=x, y=y, label=label), size=3, color="white")
```

last update: Wed Jul 20 23:26:26 2016

Package developed at CRG, Barcelona.
Page build with knitr.
Didac Santesmasses, 2016.

ggsunburst

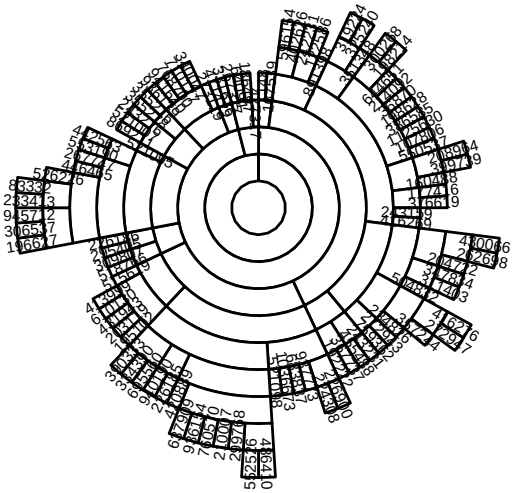
sunburst

Usage:

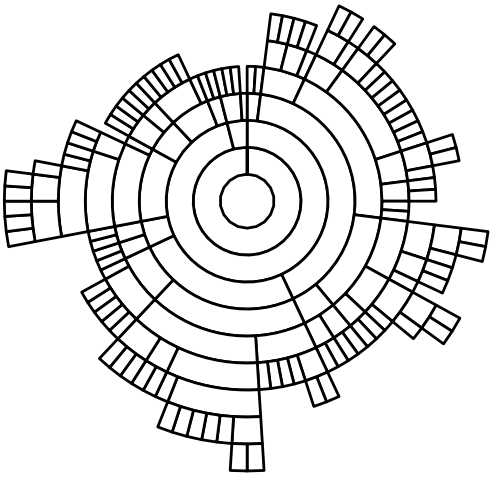
```
sunburst( data, rects.fill="white", rects.fill.aes=0, rects.color="black", rects.size=.5,
          blank=T, leaf_labels=T, leaf_labels.size = 2, leaf_labels.color = "black",
          node_labels = F, node_labels.size = 2, node_labels.color = "black", node_labels.min = 90)
```

- data object obtained using the `sunburst_data` function
- rects.fill color of space-filled nodes
- rects.fill.aes color of space-filled nodes mapped to a variable
- rects.color color of line delimiter between partitions
- rects.size size of line delimiter between partitions
- blank if TRUE, a blank theme is applied
- leaf_labels if TRUE, shows leaf labels
- text.size size for text of leaf labels
- text.color color for text of leaf labels
- node_labels if TRUE, shows node labels
- node_labels.size size for text of node labels
- node_labels.color color for text of node labels
- node_labels.min sets the minimum size in angles for a internal node to display the label

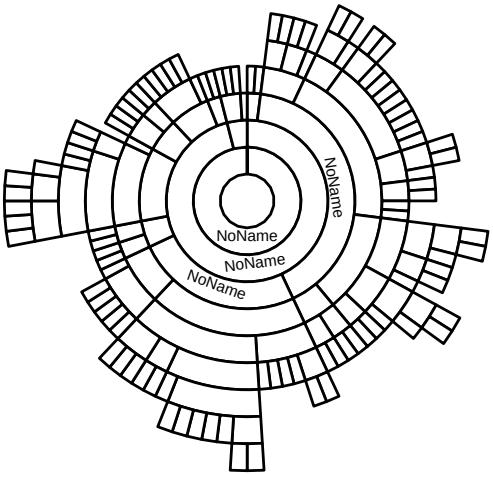
```
sunburst(sb)
```



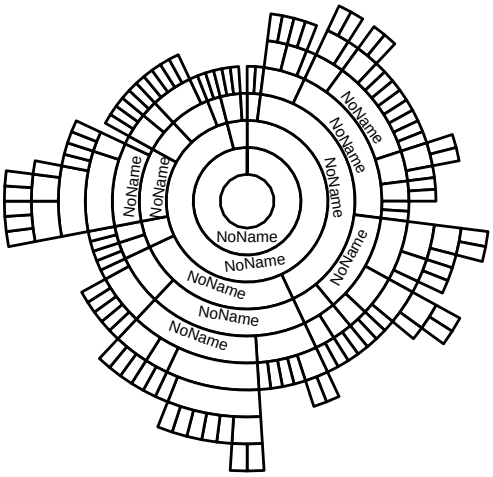
```
# hide leaf labels (terminal nodes)
sunburst(sb, leaf_labels=F )
```



```
# show node labels (internal nodes)
sunburst(sb, leaf_labels=F, node_labels=T )
```



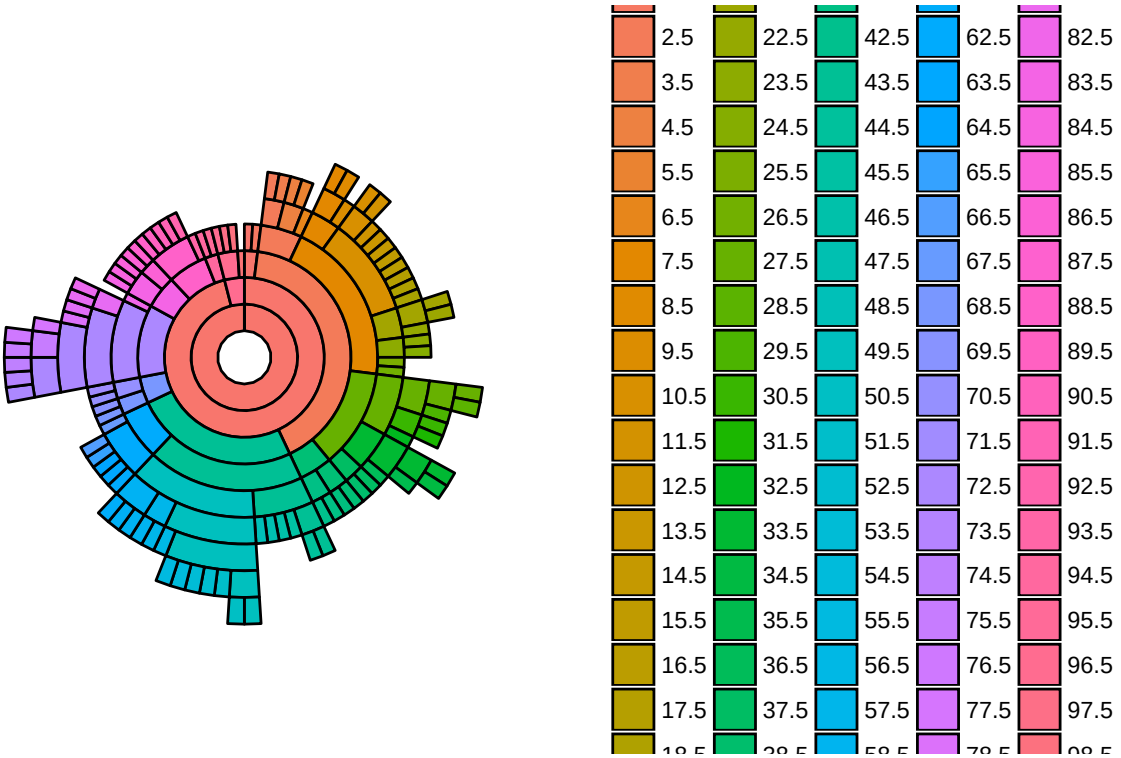

```
# modifying in which internal nodes the label is shown, by the node size in angles
# node_labels.min sets the minimum size in angles for a node to display the label, 90 degrees by defaul
# Here, nodes of size 30 degrees or larger will display the label
sunburst(sb, leaf_labels=F, node_labels=T, node_labels.min = 30 )
```




```
# fill and color
sunburst(sb, leaf_labels=F, rects.fill="grey80", rects.color="white", node_labels=T,
         node_labels.color="white", node_labels.size = 2.5)
```

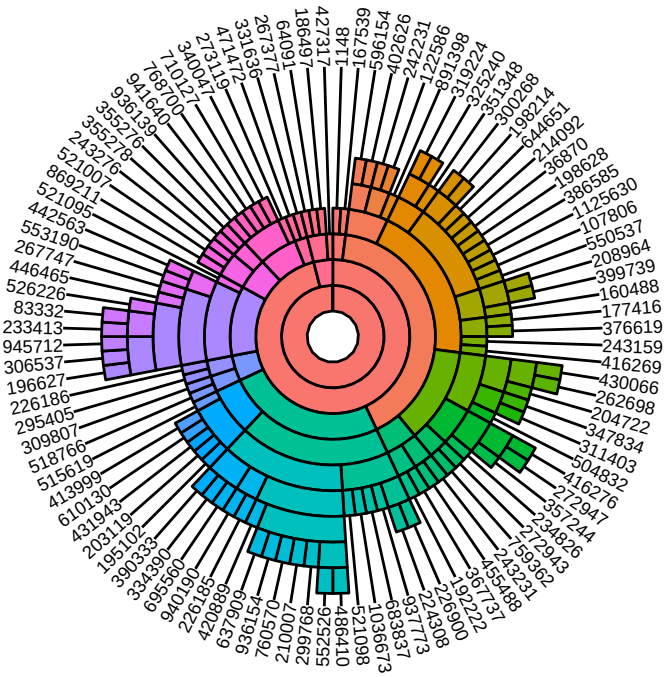


```
# mapping fill to a variable.
# For example the xmin value: sb$leaf_labels has a variable called xmin, let's map it to the fill aesth
sunburst(sb, leaf_labels=F, rects.fill.aes="factor(xmin)")
```




```
# the sunburst function returns a ggplot object. This allows additional layers to be added by calling o

sunburst(sb, leaf_labels=F, rects.fill.aes="factor(xmin)") +
  geom_text(data=sb$leaf_labels, aes(x=x, y=max(y)+1, label=label, angle=angle,
                                     hjust=hjust), size=2.5) +
  geom_segment(data=sb$leaf_labels, aes(x=x, xend=x, y=y+.5, yend=max(y)+1)) +
  theme(legend.position="none")
```



last update: Thu Jul 21 17:26:33 2016

Package developed at CRG, Barcelona.
Page build with knitr.
Didac Santesmasses, 2016.

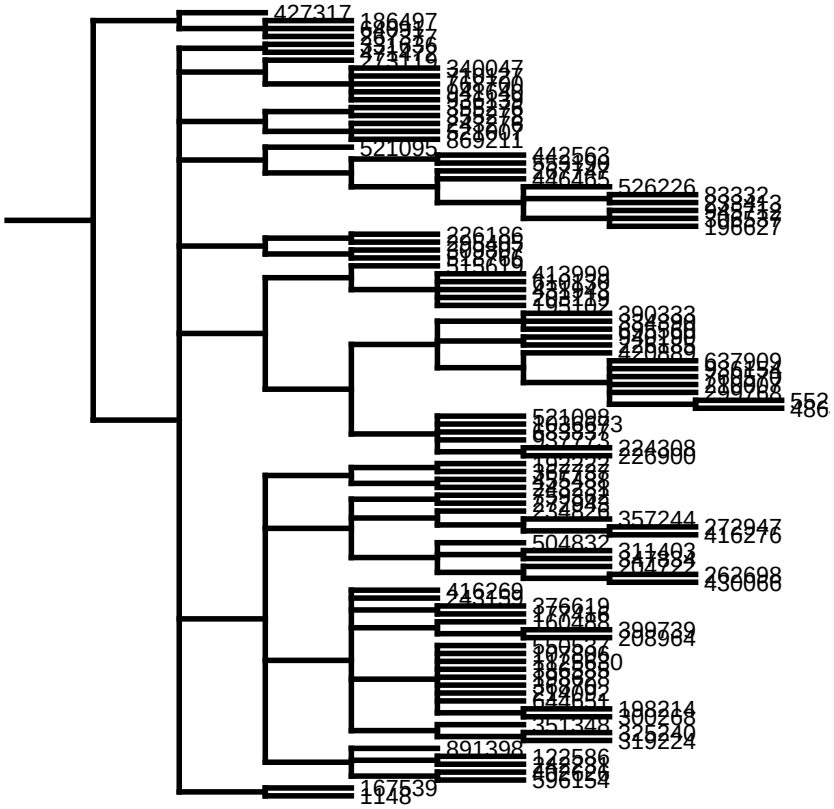
ggsunburst

ggtree

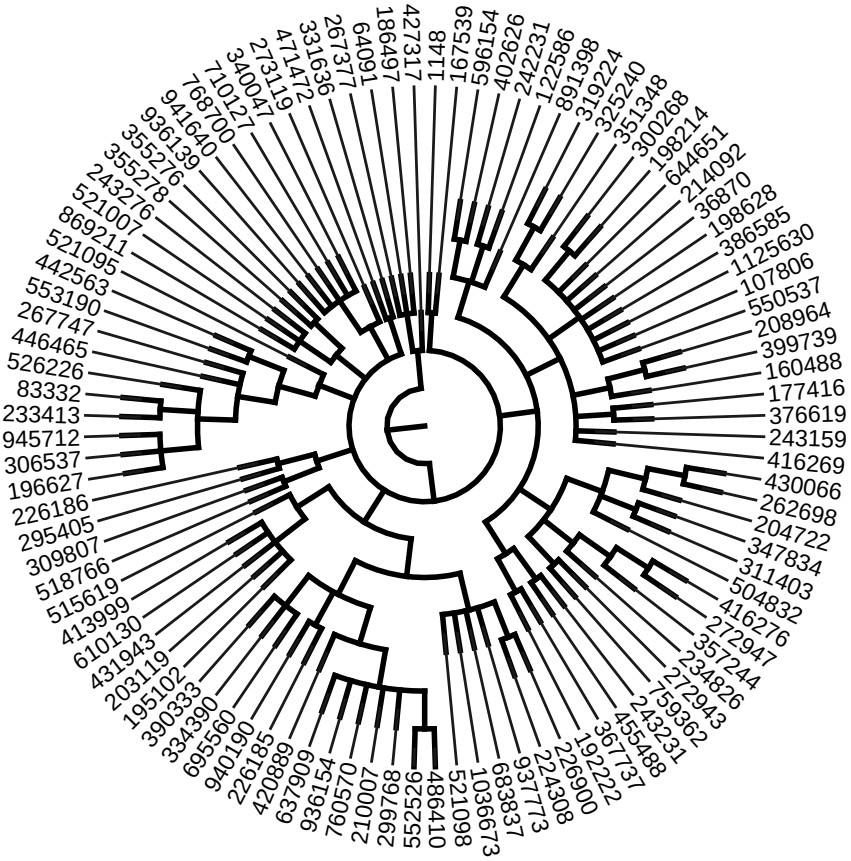
Usage:

```
ggtree(data=0, plot=0, rotate=T, color="black", size=1, blank=T,
       labels=T, text.size=3, text.color="black", polar=F)
```

```
ggtree(a)
```




```
ggtree(a, polar=T)
```



last update: Wed Jul 20 23:24:21 2016

Package developed at CRG, Barcelona.
Page build with knitr.
Didac Santesmasses, 2016.