TESI DOCTORAL UPF 2016

Origin and evolution of eukaryotic compartmentalization

Alexandros Pittis

Director

Dr. Toni Gabaldón

Comparative Genomics Group
Bioinformatics and Genomics Department
Centre for Genomic Regulation (CRG)

To my father Stavros for the PhD he never did

Acknowledgments

Looking back at the years of my PhD in the Comparative Genomics group at CRG, I feel it was equally a process of scientific, as well as personal development. All this time I was very privileged to be surrounded by people that offered me their generous support in both fronts. And they were available in the very moments that I was fighting more myself than the unsolvable (anyway) comparative genomics puzzles. I hope that in the future I will have many times the chance to express them my appreciation, way beyond these few words.

First, to *Toni*, my supervisor, for all his support, and patience, and confidence, and respect, especially at the moments that things did not seem that promising. He offered me freedom, to try, to think, to fail, to learn, to achieve that few enjoy during their PhD, and I am very thankful to him.

Then, to my good friends and colleagues, those that I found in the group already, and others that joined after me. A very special thanks to *Marinita* and *Jaime*, for all their valuable time, and guidance and paradigm, which nevertheless I never managed to follow. They have both marked my phylogenomics path so far and I cannot escape. To *Les* and *Salvi*, my fellow students and pals at the time for all that we shared; to *Gab* and *Fran* and *Dam*, for all their direct and indirect inspiration and valuable friendship, important influence for my beginning in the field; to *Fran* especially for all these little chats on statistics that so much meant to me; to *Ester*, for all the invaluable uplifting moments and determination and generosity and affection just before I would explode; to *Ewa* and *Damian*, for the space trips and kitten smiles and laugh and easiness; *Laia* for the number one laughter in PRBB and *Ernesto* for his sharpness except for the wildest look and most ultra-noble outfit around, so unfortunate to have me as the "senior"

PhD when they started; *Cindy* for all her kindness and selflessness and equilibrium and input; *Radi* for her freshness and smile; *Miguel*, for his eukaryotic endless geekiness, *Irene* for her cheerful company at the corner; *Jesse* for all the passes once upon a time; and *Su* por *su* carácter and her endless electric energy; and the newcomers, Veronica and Grant for the new air that they bring in this ever renewing environment. Toni now knows, one non latin alphabet is highly recommended!

I am very grateful for the time I spent with all of them during the PhD, and I sincerely hope that our personal and scientific lives will keep crossing in the years to come.

And to all the other *friends* and *colleagues* in PRBB that we shared thoughts and breaks and worries and cigarettes.

Finally, to *Dimi* and my *brothers* and *sisters* (after my beloved sister *Kat*) here in Bcn and Athens and elsewhere, they know, for all the days and nights they had - and will have - to stand me, and deal with my tensions (and tendons), and lift me up, and listen to my spooky - (very) late night tales about exotic early-earth creatures and groovy endosymbionts. And to all my *family*, for their limitless support and trust always.

Alexandros Pittis

Barcelona, July 2016

Abstract

The origin of eukaryotic compartmentalization stands as a major conundrum in biology. Current evidence indicates that the last eukaryotic common ancestor (LECA) already possessed many eukaryotic hallmarks, including a complex subcellular organization. The lack of evolutionary intermediates challenges the elucidation of the relative order of emergence of eukaryotic traits. Central in the discussion is the exogenous origin of mitochondria, ubiquitous eukaryotic organelles derived from an α-proteobacterial endosymbiont. Different hypotheses disagree on whether mitochondria were acquired early or late during eukaryogenesis. Similarly, the nature and complexity of the receiving host are debated, with models ranging from a simple prokaryotic host to an already complex proto-eukaryote. In this thesis, I have used phylogenomic methods to address different questions on the origin and evolution of subcellular compartmentalization in Eukaryotes. We provide evidence for extensive retargeting of proteins between the different compartments, and suggest an evolutionary link between mitochondria and peroxisomes. We focus on the evolution of calcium homeostasis in mitochondria and reveal strong co-evolution patterns among the components of the recently identified mitochondrial calcium uniporter complex. Through alternative methodologies we analyze the phylogenetic signal carried by LECA-inferred gene families. Our analyses indicate that the ancestral eukaryotic proteome is a composite of genes originating from different prokaryotic sources. Finally, our work provides strong support for the late acquisition of mitochondria by a complex host. Altogether, our findings shed light on long-standing questions on the origin of Eukaryotes and provide new grounds for further advancements, as new data become available.

Keywords: Evolution, Eukaryotes, Organelles, Endosymbiosis

Resumen

El origen de la compartimentación celular en Eucariotas se presenta como uno de los enigmas mas importantes de la biología. Las evidencias actuales indican que el último ancestro común eucariota (LECA) ya poseía muchas de sus características avanzadas, incluyendo una organización subcelular compleja. Además, la falta de intermediarios evolutivos desafía la elucidación del orden en el que las características eucariotas aparecieron. En el centro de la discusión está el origen exógeno de las mitocondrias, orgánulos eucariotas derivados de α-proteobacteria vía endosimbiosis. Las diferentes hipótesis discrepan sobre si las mitocondrias fueron adquiridas al principio o al final durante el proceso de eucariogénesis. Del mismo modo, se debate la naturaleza y complejidad del hospedador, con modelos que van desde un simple hospedador procariota hasta un proto-eucariota dotado de cierta complejidad. En esta tesis, se han utilizado métodos filogenómicos para contestar a diferentes preguntas sobre la evolución de la compartimentación eucariota. Proporcionamos evidencia de una amplia relocalización de proteínas entre los diferentes compartimentos y sugerimos un vínculo evolutivo entre las mitocondrias y los peroxisomas. Nos centramos en la evolución de la homeostasis del calcio en las mitocondrias y observamos patrones de coevolución entre los componentes del sistema transportador mitocondrial de calcio. A través de metodologías diferentes se analiza la señal filogenética de familias de genes del ancestro común de Eucariotas. Nuestros análisis demuestran que el proteoma ancestral eucariota es un mosaico de genes de diferentes fuentes procariotas. Por último, nuestro trabajo proporciona un fuerte soporte a la hipótesis que la adquisición de la mitochondria tuvo lugar hacia el final de la eucariogénesis por parte de un hospedador complejo. En conjunto, nuestros resultados aclaran cuestiones que llevaban mucho tiempo abiertas sobre el origen de los Eucariotas y proporcionan nuevas bases para avances adicionales.

Palabras Clave: Evolución, Eucariotas, Orgánulos, Endosimbiosis

Thesis overview

This PhD thesis focuses on the study of the origin of Eukaryotes and the evolution of eukaryotic organelles, through the use of comparative genomic approaches. The work is divided into different chapters, which I briefly introduce here.

Chapter 1 provides an introduction to the question of the origin of the eukaryotic cell. An overview of the evolution of cellular life on earth is presented, and the main hypotheses on the origin of Eukaryotes are discussed within a historical context.

Chapter 2 presents the main objectives of this thesis.

Chapter 3 is a review on the "Origin and evolution of metabolic sub-cellular compartmentalization in Eukaryotes". The theories on the origin of organellar proteomes are discussed, with a special focus on the evolution of protein retargeting. An novel analysis on the retargeting patterns across different eukaryotic compartments is presented.

Chapter 4 describes the evolutionary analysis of the main molecular components of calcium homeostasis in mitochondria. It provides insights into the origin of the molecular machineries of mitochondrial calcium uptake in Eukaryotes and examines co-evolution patterns between proteins that are known to interact.

Chapter 5 describes the development of a novel methodology for the evaluation of the phylogenetic signal carried by protein families. We analyze the genomic component inferred to the Last Eukaryotic Common Ancestor (LECA), and we show that the signal consistently points to contributions from various prokaryotic sources.

Chapter 6 presents our analysis on the relative timing of the acquisition of mitochondria during the emergence of the eukaryotic cell. Using phylogenomics, we analyze the phylogenetic signal carried by gene families

considered to be present in the LECA. Our results suggest that mitochondria were acquired relatively late by a host that already possessed a certain degree of genomic complexity.

Chapter 7 is our response to criticisms related to our work in the previous chapter. We get the opportunity to re-assess the use of branch length distribution from phylogenetic trees for testing hypothesis on the evolution of genes and lineages, and provide additional support to our work through new data and analyses.

Chapter 8 is the general discussion of the thesis, where I present some thoughts on the importance of comparative genomics in the genomics era to address key biological questions. The role of comparative genomic analysis in the problem of eukaryotic origins is discussed, together with some perspectives on the future of the research in the topic.

Finally, the **Appendix** compiles a list of studies in which I have participated during my PhD.

Contents

Part I

Introduction

1 The evolutionary origin of eukaryotic cells

1.1 Evolution of cellular life on Earth

The question of the origin of Eukaryotes has been puzzling evolutionary biologists for years (López-García and Moreira, 2015). On the one hand, as eukaryotic organisms, our own lineage's deep roots is a major piece in the history of our species. On the other, the emergence of Eukaryotes on Earth is arguably the most important transition towards complexity in the evolution of life after the origin of life itself. Whether or not complex life is expected or even inevitable, given enough time, is an enigma with many implications for our understanding of biology and the world.

According to current estimates, the age of the Universe is about 13.8 billion years ago (Gya), and our solar system's formation dates back to 4.54 Gya, which is also the upper limit estimation for the formation of Earth (Dalrymple, 2001). Examination of the microfossil, chemofossil and rock records for direct or indirect traces of life based on evidence for biological activity, suggests that the first living organisms appeared in our planet 3.5-4.1 Gya (Bell et al., 2015), if not earlier. Thus, while the Earth is the only known planet to harbor life and we have no evidence of any earlier biological activity anywhere in the Universe, it seems to have appeared rapidly once the planet Earth was formed. This realization has led some people to believe that if simple life can be rapidly form given the right conditions, then it might be a common phenomenon in the Universe (Sagan et al., 2013). Undoubtedly, because of its many implications and its great difficulty, the origin of life remains one of the greatest unsolved scientific problems (Chyba and Sagan, 1992; Kauffman, 2011).

Looking at the overall picture of the evolution of life on Earth (Figure

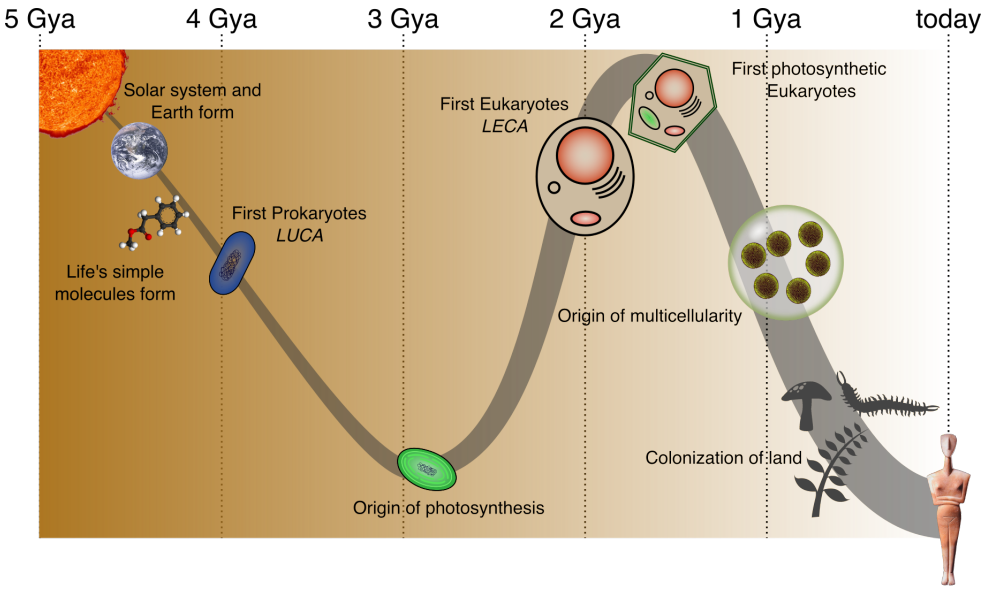


Figure 1.1: Timeline of the evolution of life on Earth. Schematic representation of major events in the evolution of cellular life on Earth. Half of the time elapsed since the earliest indications for biological activity (~4 billion years ago (Gya)), prokaryotic cells are the only inhabitants of the planet. Eukaryotic cells emerge around 2 Gya and multicellular organisms appear another billion years later (see also text).

1.1), it is clear that there has been an increase in complexity, while by no means overlooking that evolution towards simplicity has equally proved to be adaptive in various lineages (Wolf and Koonin, 2013; O'Malley et al., 2016). The first organisms on Earth were small, single-celled and simple in terms of cellular organization, resembling probably some present-day bacterial cells. Cells of this kind are classified as Prokaryotes (coming from *pro* "before" + *karyon* "nut or kernel") due to the absence of a nucleus, the distinct membrane-bound compartment of Eukaryotes (*eu* "well"), which contains most of the genetic material. Thanks to their ability to evolve and adapt, Prokaryotes diversified and inhabited almost any possible niche on Earth. If not for the emergence of Eukaryotes, through a process called eukaryogenesis, and their subsequent diversification into animals, plants, and fungi, among other groups, Prokaryotes would likely still be the only type of organisms inhabiting our planet. The fossil record suggests that the first Eukaryotes appeared 1.6-2.1 Gya (Knoll et al., 2006;

Han and Runnegar, 1992; Bengtson et al., 2009), while through indirect evidence, traces of eukaryotic-specific biomarkers, ages up to 2.7 Gya have been suggested (Brocks et al., 1999). Convincing eukaryotic fossils, with detectable structures suggesting the presence of cytoskeleton like those of *Tappania plana*, are dated to 1.5 Gya, but are commonly interpreted as stem group Eukaryotes: intermediate steps, lineages that branched off during eukaryogenesis, before the diversification of the main "crown" groups (Porter, 2004). However at these time scales, the record is too sparse and inconclusive, and the identification of major cellular features from such ancient rocks, which would give valuable information on the process, is highly problematic if not impossible. In the absence of reliable information on intermediate forms in the fossil record, and since, as far as we know, no such organisms survive today, we are left with the comparison of extant eukaryotic diversity as our only tool to infer the steps that gradually led to Eukaryotes. The transition from a prokaryotic to a eukaryotic cell brought highly significant changes at all levels. The distinguishing structure of Eukaryotes and the deep fundamental differences with Prokaryotes were described by Roger Stanier, Michael Douderoff, and Edward Adelberg in early 1960s as "the greatest single evolutionary discontinuity to be found in the present day world" (Stanier, 1963). Since then, and most particularly in the current genomics era, the comparison of molecular sequences has played a significant role in the study of the evolution of the cell. Phylogenomics has revolutionized our understanding of the evolutionary relationships among eukaryotic lineages and has provided significant answers on the origin of Eukaryotes, but it has also brought new questions and various, often conflicting, interpretations of the genomic data. Despite great advances in recent years the Eukaryote-to-Prokaryote transition still remains one of the biggest questions in evolutionary biology.

1.2 Endosymbiosis and the evolutionary origin of Eukaryotes

The emergence of Eukaryotes has long been considered a major transition in the evolution of life on Earth (Szathmary and Smith, 2000). This transition brought not only a highly sophisticated subcellular compartmentalization

and a tight control of metabolic compartmentalization, but also a high level of control in gene expression with the separation of transcription and translation (Martin and Koonin, 2006) and a general organizational complexity associated with many innovations at the structural and molecular level (Table 1.1). For most of the features that are ubiquitous in Eukaryotes, there are no direct counterparts in Bacteria or Archaea, which leaves us with the question of how they arose in the first place. All extant eukaryotic lineages (Figure 1.2) share the main features of cellular architecture and molecular regulatory circuits. Reconstruction of the genetic repertoire of the Last Eukaryotic Common Ancestor (LECA) using comparative genomics has shown that this ancestor already possessed the main features associated to eukaryotic complexity. Without any (known) intermediate between such complex LECA and the more simple Prokaryotes, the possible steps in which this complexity may have been built remain mostly in the area of speculative ideas. Deciphering this gap is one of the greatest challenges of evolutionary biology today (Koonin, 2010).

Feature	Eukaryotic cells	Prokaryotic cells
Cell size	10-100μm	1-10μm
Nucleus	Membrane-enclosed nucleus & nucleolus	No
Chromosomes	Multiple linear & histones	Single circular, lacking histones
Membrane-bound organelles	Yes	No
Cytoskeleton	Yes	No
Flagella	Tubulin microtubules (9+2)	Flagellin filaments
Ribosomes	Large (80S)	Small (70S)
Cell division	Mitotic division	Binary fission
Sexual reproduction	Meiosis	No
Introns	Yes	No
Transcription-Translation	Uncoupled	Coupled

Table 1.1: Main differences between eukaryotic and prokaryotic cells.

The endosymbiotic origin of mitochondria and plastids has a prominent position in the discussion of eukaryotic origins. The modern era of the research on the origin of Eukaryotes begins with the influential paper of Lynn Margulis (Lynn Sagan at the time) "On the origin of mitosing cells" in 1967 (Sagan, 1967). It was her that set the ground for the establishment of the modern endosymbiotic theory, proposing that mitochondria were once free-living bacteria that were transformed into organelles. However

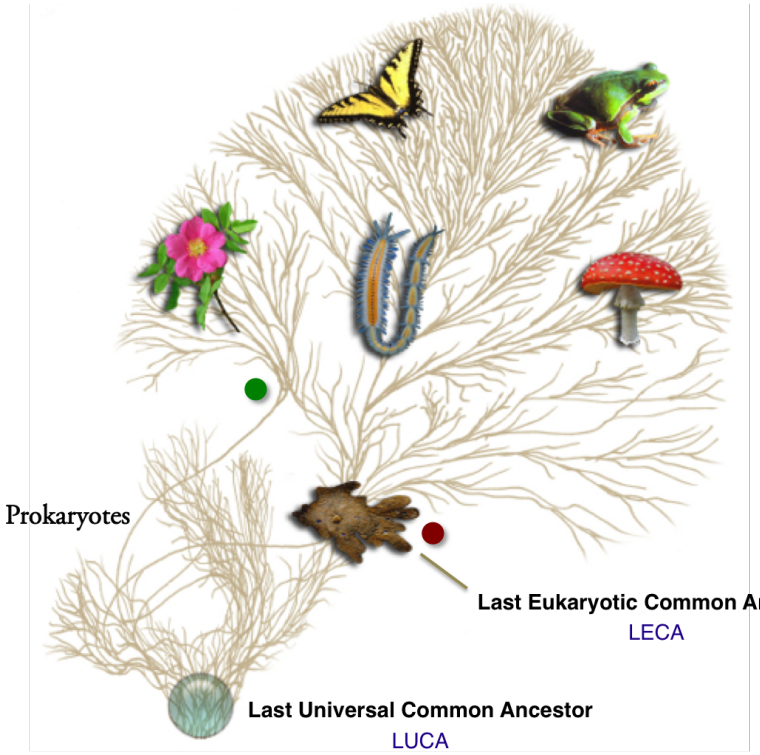


Figure 1.2: Tree of Life. A simplified representation of the Tree of Life, with emphasis in the eukaryotic domain. The origin of life is represented as the root of the tree at the bottom (LUCA). The last common ancestor of all extant Eukaryotes (LECA) is represented graphically at the base of the clade. The red and green circles represent the endosymbiotic origin of mitochondria (base of Eukaryotes) and plastids (base of Viridiplantae) accordingly. Image adapted from *http://tolweb.org*.

the idea of an organism being engulfed by another to explain the origin of organelles was not completely new, but based on earlier ideas, as Margulis always acknowledged. Endosymbiotic theories for organelles were proposed already in the beginning of the 20th century. Already back in 1905, the Russian botanist Constantin Mereschkowsky, considered the "founding father" of endosymbiotic theory (Archibald, 2015) proposed an endosymbiotic origin for the plastids of photosynthetic eukaryotes. With other fellow "symbiogeneticists" in Russia at the time, Andrey Famintsyn and Boris Kozo-Polyansky, he argued that linear "darwinian" evolution

could not account for large-scale biological changes, instead such changes could result from different organisms combining to create a new one, a process that he named "symbiogenesis". In his view, the mechanism of biogenesis of plastids from pre-existing organelles was suggesting their free-living ancestry (Martin and Kowallik, 1999). Independently, the American researcher Ivan Wallin suggested an endosymbiotic origin for mitochondria, and he even argued for a change in the hereditary patterns of the host organism through the acquisition of genes from the symbiont (Wallin et al., 1927), a process known today to have significantly contributed to shaping eukaryotic genomes (Timmis et al., 2004). Nevertheless, at the time all theses ideas were treated with contempt and were mostly ignored by the scientific establishment. Lynn Margulis revived and popularized symbiogenesis ideas and advocated that symbiosis and endosymbiosis particularly, had played a central role in the evolution of cellular life. The discovery that mitochondria (Nass and Nass, 1963) and plastids (Ris and Plaut, 1962) contain their own DNA few years earlier, and finally the development of molecular and sequence comparison techniques led to the universal acceptance of the endosymbiotic origin of the two DNA-bearing organelles (Gray and Doolittle, 1982).

Nowadays we know that the intracellular symbiotic association involving a host cell and a symbiont has been one of the important evolutionary forces in evolution, and has been key in major innovations in Eukaryotes (Archibald, 2015). The origin of mitochondria and plastids, which introduced aerobic respiration and photosynthesis to Eukaryotes, respectively, are the two most significant examples, but there are many other endosymbioses that occured more recently across the Eukaryotic Tree of Life (Nowack and Melkonian, 2010). The conversion of a free-living organism to an organelle, involves a high degree of metabolic integration, the transfer of genes from the symbiont to the host, a process called endosymbiotic gene transfer (EGT), and targeting and transport systems for re-localizing the proteins back to the endosymbiotic organelle (Cavalier-Smith and Lee, 1985). How this integration happens is an active research topic, with major implications for our understanding of eukaryogenesis.

1.3 Three vs two domains

Sequence-based phylogenies, apart from providing the final proof for the bacterial nature of mitochondria and plastids (Gray, 1992), have been in the core of the discussion of all other aspects of the question of eukaryogenesis over the last few decades. The theoretical frameworks for molecular phylogenetics were already set in the early 1060s by Emile Zuckerkandl and Linus Pauling, who first started exploring the use of sequences as taxonomic characters (Zuckerkandl and Pauling, 1965). They were at first working on the few sequences available at this time, mainly globins and hemoglobins. A decade later, Carl Woese and George Fox introduced the use of small subunit ribosomal RNA (SSU rRNA) genes, based on the same principles, to establish a phylogeny for the history of life. The rRNA molecules being a fundamental component of the ribosome, the major component of the translational apparatus, are ubiquitously distributed and highly conserved at the sequence level, hence they were considered to carry strong signal for resolving phylogenetic relationships between species (Olsen and Woese, 1993). And indeed, rRNA molecules are still today, almost 40 years later, used as standard phylogenetic markers. Traditionally, the diversity of cellular life on Earth had been divided into Prokaryotes and Eukaryotes. Based on the comparison of the morphological, physiological and biochemical characteristics of organisms, the negative definition of anucleate microbes as Prokaryotes was dichotomously dividing life in these two domains (Stanier and Niel, 1962). When Carl Woese and his colleagues reconstructed a phylogenetic tree using the SSU rRNA genes (Figure 1.3), strikingly they found Prokaryotes to be separated into two distinct phylogenetic groups and Eukaryotes to form a third monophyletic clade. In their influential work published in 1977, they suggested the existence of three "domains" of life or superkingdoms, classifying them into Bacteria, Eucarya, and the newly defined group of Archaea (Woese and Fox, 1977; Woese et al., 1990). The "universal tree of life" (Pace, 2006, 2009) had a huge impact in our understanding of biological diversity, and the discovery of Archaea (archaeBacteria) as a distinct domain has been one of the most important discoveries in the history of microbiology.

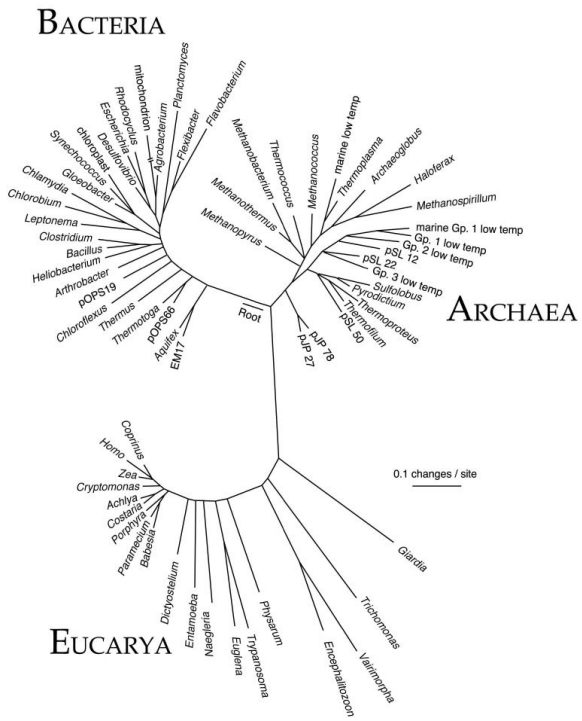


Figure 1.3: The 3 domain Tree of Life. The traditional 3 domain tree, based on SSU rRNA sequences, divides the diversity of organisms into Bacteria, Archaea and Eukaryotes. For this tree 64 rRNA sequences of all three domains were used for a phylogenetic reconstruction. Image obtained from Pace, 1997.

The unrooted three-domain tree, apart from the remarkable discovery of Archaea, had also major implications for the origin of Eukaryotes, as it seemed to suggest that Eukaryotes resulted from progressive evolution towards complexity, in parallel to the two prokaryotic lineages. It was also soon realized the existence of an evolutionary relatedness between Eukaryotes and Archaea, which suggested a sister relationship between these two groups. In the genomics era the idea of a sister relationship between Eukaryotes and Archaea has been conceptually built upon two sources of data, 1) various attempts to root the universal tree of life (ToL) based on different molecular criteria, have placed the root along the bacterial root or within Bacteria, thus having Archaea and Eukaryotes as sister groups within a monophyletic clade (Gogarten et al., 1989; Brown and Doolittle, 1995; Baldauf et al., 1996; Zhaxybayeva and Gogarten, 2007) and 2) the

analysis of eukaryotic genomic components has revealed that genes involved in informational processes (replication, transcription, translation) are closely related to Archaea (Rivera et al., 1998; Yutin et al., 2008). The topology of the universal ToL (Figure 1.3) indicates monophyly of the three domains, each domain forms a monophyletic group and thus Archaea and Eukaryotes (assuming a rooting in the bacterial branch) share a common ancestor and both together share a common ancestor with Bacteria, the Last Universal Common Ancestor (LUCA), back in the origins of cellular life. However, with the accumulation of additional data in the decades that followed the publication of the three-domains tree, it became clear that the topologies obtained using the sequences of different genes often provided contradictory results. In 1984 James A. Lake and colleagues (Lake et al., 1984) put forward the "Eocyte hypothesis". Comparing the structural patterns of the ribosomal large and small subunits, they presented an alternative view according to which Eukaryotes emerged from within Archaea, from a specific group of thermophilic prokaryotes, the "eocyte" archaebacteria. The idea of a sister-group relationship between Eukaryotes and Crenarchaeota (eocytes), one of the major archaeal divisions, gained further support in the genomic era, when many fully-sequenced genomes become available, and multi-gene phylogenies were explored (Embley and Martin, 2006; Cox et al., 2008; Williams et al., 2012, 2013).

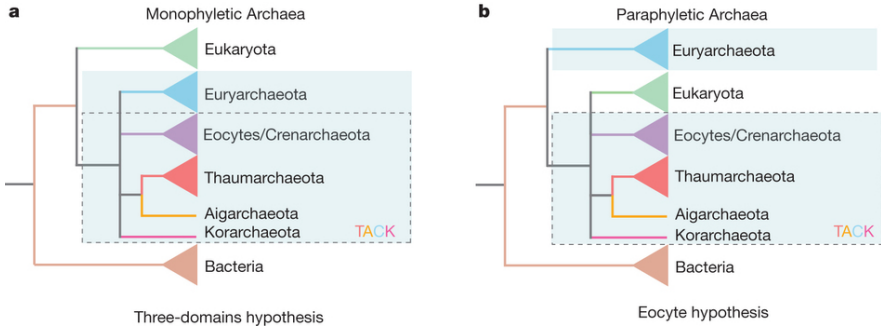


Figure 1.4: Competing hypothesis for the archaeal roots of Eukaryotes. a, The rooted three-domains tree depicts Archaea as a monophyletic group, where the TACK superphylum (Crenarchaeota and their relatives) groups together with with Euryarchaeota, in the exclusion of Eukaryotes. **b,** The rooted eocyte tree depicts Eukaryotes emerging within Archaea, with a sister-group relationship to TACK. Image obtained from Williams et al., 2013

Despite the availability of abundant data, and better and more complex evolutionary models which leads to more accurate phylogenetic reconstructions, the discussion is still not settled. The three-domains (3D) vs the eocyte (2D) debate is one of the most controversial in the study of the origins of Eukaryotes, largely focused on the exploration of the phylogenetic signal carried by sequences and on whether it can reliably provide information on very ancient relationships. As in most other aspects of the question of the origins of Eukaryotes, different studies have provided for and against arguments for each model, and even different methods applied to the same data have provided opposite results (for a review see Gribaldo et al., 2010). The reason why the monophyly of Archaea has been so important, is connected to its implications on the nature of eukaryotic cells and how eukaryotic complexity arose. The 3D model implies a gradual increase in complexity, with Eukaryotes diverging from the base of the archaeal clade. On the other hand, a major implicaton of the 2D eocyte model is that Archaea is a polyphyletic group or, putting this in an alternative manner, Euakryotes are Archaea. This model, at least its initial formulations, postulates a radical shift caused by the interaction of a member of Crenarchaeota or a related group (within the TACK superphylum) with a bacterium, particularly an α-proteobacterium that gave rise to mitochondria. As it will be discussed later, the difficulty to give a definitive answer to this question, among others, is due to the expectedly weak phylogenetic signal in such deep evolutionary relationships, but also the fact that different components of the eukaryotic genome point to different prokaryotic ancestors, even within archaeal diversity. Significant progress has been achieved during the last years by new techniques and further sampling of previously unexplored microbial diversity (Spang et al., 2015) and there is no doubt that over the following years new discoveries will shed further light on this important question.

1.4 Timing of mitochondrial endosymbiosis

The timing of the mitochondrial acquisition through endosymbiosis (Poole and Gribaldo, 2014) occupies a central position in the discussion on eukaryotic origins. The nature of the receiving host, the mechanism of

acquisition, the impact that it had in the evolutionary success of Eukaryotes, are all aspects that cannot be disentangled from the timing in which the endosymbiosis occured, relative to the process of eukaryogenesis. It has been generally accepted that modern mitochondria and related organelles, mitosomes and hydrogenosomes, all trace back to this ancestral endosymbiont (Müller et al., 2012), and that the mitochondrial ancestor was a bacterium related to α-proteobacteria (Gray, 1992), even if the precise nature of it is still uncertain (Gabaldón and Huynen, 2004; Rodríguez-Ezpeleta and Embley, 2012; Thrash et al., 2011; Wang and Wu, 2015). Until recently, with the discovery of the oxymonad Monocercomonoides sp., the first Eukaryote discovered lacking all hallmark mitochondrial proteins (Karnkowska et al., 2016), there was no known Eukaryote without a mitochondrion or a mitochondrial-derived organelle, and some of its functions were considered indispensable for eukaryotic cells. However in the early steps of the research on eukaryotic origins using molecular sequences, the first phylogenetic trees based on rRNA (Woese et al., 1990) and elongation factor sequences (Hashimoto and Hasegawa, 1996), placed some protozoa that lack mitochondria at the base of the eukaryotic tree. These organisms, were thought to represent early diverging lineages, preceding the mitochondrial acquisition, and were named "Archezoa" (Cavalier-Smith, 1983, 1989), relics of the first amitochondriate proto-eukaryotic lineages.

Based on implied assumptions on the relative timing of mitochondrial endosymbiosis, most eukaryogenesis models can be classified into mitochondria-early (mito-early) and mitochondria-late (mito-late) (Figure 1.5). The Archezoa hypothesis, the classical mito-late scenario, served as a main research framework for many years after it was first proposed by Thomas Cavalier-Smith in 1983. As others did previously, Cavalier-Smith considered the presence of a cytoskeleton, and phagocytotic capibilities as a prerequisite for the uptake of the bacterial endosymbiont for mitochondrial origins. Anaerobic single-celled amitochondriate organisms, such as diplomonads, parabasalids and microsporidians were thought to belong to lineages that had diverged after the formation of the nucleus and cytoskeleton and before mitochondrial acquisition, were thus intermediate steps in the

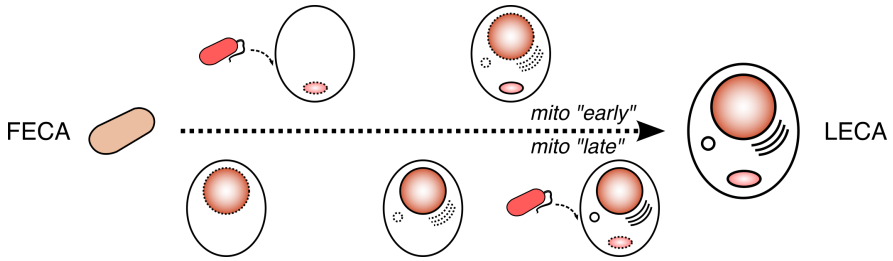


Figure 1.5: The timing of mitochondrial acquisition in the eukaryotic stem phase. Competing hypotheses on eukaryogenesis can be roughly grouped into mito-early (top) and mito-late (bottom) models. The former consider the α-proteobacterial endosymbiosis as the driving force of eukaryogenesis. The latter assume a certain degree of cellular complexity predating the acquisition of mitochondria. In both models there is an intermediate phase (the stem phase of eukaryotic evolution, Koonin, 2010) between the First Eukaryotic Common Ancestor (FECA), commonly an Archaeon or Archaea-related organism, and the Last Eukaryotic Common Ancestor (LECA) before the diversification of Eukaryotes.

Prokaryote-Eukaryote transition. This hypothesis was seriously challenged by two findings: 1) It was realized that their monophyly and basal position relative to other eukaryotic groups was an long brach attraction artefact due to their lifestyle adaptations and fast evolving rates (Hirt et al., 1999; Philippe, 2000) and 2) Members of the proposed Archezoa were found to possess mitochondria-derived anaerobic organelles and some of the nuclear-encoded, hallmark mitochondrial genes (HSP70, HSP60) were retained in their nuclear genome (Embley et al., 2003; Tovar, 2007). As mentioned previously, mitochondria are now considered to be an ancestral eukaryotic feature. These results led to the conclusion that Archezoa was not a true clade but rather an assemblage of, phylogenetically unrelated, highly diverged organisms descending from mitochondria-bearing ancestors. After the rejection of the Archezoa hypothesis, numerous new models emerged, without the assumption of a complex host for the acquisition of mitochondria. The most classic of the first mito-early scenarios, was proposed by William Martin and Mikloś Müller in 1998, and is known as the"hydrogen hypothesis" Martin and Müller (1998). As all mito-early models, it hypothesizes that there was never an amitochondriate phase, and that the advance in complexity came after the symbiosis between two simple prokaryotes, an archaeon

and a bacterium. In these models, the mitochondrial acquisition is considered an extremely rare event that triggered eukaryogenesis as a response. It is important to note that the fact that Archezoa do not represent "missing links" on the way to LECA, does not make it impossible that such links do exist or existed in the past and are now extinct. Undoubtedly, the increase in complexity throughout eukaryogenesis must have occurred progressively during the phase between the First Eukaryotic Common Ancestor (FECA) to the LECA, the stem phase. We do not know how long the stem phase lasted but we can imagine that there were lineages that diverged before LECA that got extinct or simply remain to be discovered. The fact that primary amitochondriate Eukaryotes have not been discovered is not sufficient to support that such never existed (Poole and Penny, 2007). Interpretation of these data remains a highly controversial issue (Booth and Doolittle, 2015) and the timing of the acquisition of mitochondria lies in its core. Some of the main controversies in the field are presented in the following sections.

1.5 The mosaic eukaryotic genome

The analysis of the first available eukaryotic genome sequences revealed that they are composed of genes of different evolutionary origin, some of bacterial descent, others of archaeal, and others are eukaryotic specific, lacking any detectable prokaryotic homologs (Martin et al., 1999; Katz, 2002). Among the prokaryotes-derived genes, the largest fraction originates from bacteria, in a 3-4:1 ratio with respect to archaeal derived genes (Koonin et al., 2004; Esser et al., 2004; Rivera and Lake, 2004). As more sequenced genomes became available, analyses strikingly demonstrated that these different gene subsets are enriched in different functions, informational genes, involved in replication, transcription and translation, are predominantly of archaeal origin, whereas operational genes, involved in various metabolic processes, are generally of bacterial origin (Ribeiro and Golding, 1998; Rivera et al., 1998). How this genomic mosaic formed is fundamental to any theory trying to model the origin and early evolution of Eukaryotes. To explain this genomic chimerism, most models assume a genomic fusion between at least two ancestral genomes, one archaeal and one bacterial, the latter one usually

being the mitochondrial ancestor, given the well-established mitochondrial ancestry from α-proteobacteria (Gray et al., 1999). However early attempts to ascertain the nature of the fused ancestors revealed a very complicated picture. The phylogenetic signal, rather than pinpointing two or a few more specific taxonomic groups, is distributed across various groups in both bacterial and archaeal domains (Figure 1.6). Although many of the bacteria-derived eukaryotic genes show an affiliation to α-proteobacteria, these are far from dominant, with the remainder being affiliated to a variety of other bacterial groups. Similarly, for archaea-derived genes, some are more similar to Euryarchaeota and others to members of the TACK superphylum. Being one of the most crucial points in the discussion during the last years, several scenarios have been proposed to explain this dispersion in the phylogenetic signal. Horizontal gene transfer (HGT) among prokaryotes is a well-established process, which contributes significantly in shaping prokaryotic genomes (Ochman et al., 2000). HGT results in discordant signals among gene trees and species trees and may explain disparate phylogenetic sources in the genome of an organism. Extensive HGT to the protomitochondrial ancestor has been hypothesized as a possible explanation for the dispersion of origins in the LECA repertoire (Esser et al., 2007). Furthermore, for the large evolutionary distances considered the signal is expected to be weak, if not lost in some cases. Thus resolving such ancient evolutionary relationships can be extremely challenging for current phylogenetic methods (Mossel, 2003) and the existence of noise in the data is to be expected. An alternative possibility is that (at least partially) the observed signal reflects the acquisition of genes from various sources by a eukaryotic ancestor, prone to HGT (slow-drip hypothesis, Lester et al., 2006) and/or able to phagocytize (phagotrophic proto-eukaryote, Doolittle, 1998), several bacterial sources, other than the α-proteobacterial ancestor of the mitochondrion, could have contributed genes to the early eukaryotic lineage.

1.6 Open questions in eukaryotic origins

In the past decades there has been considerable progress in our understanding of the early steps in eukaryotic evolution. Since the late 1960s, and

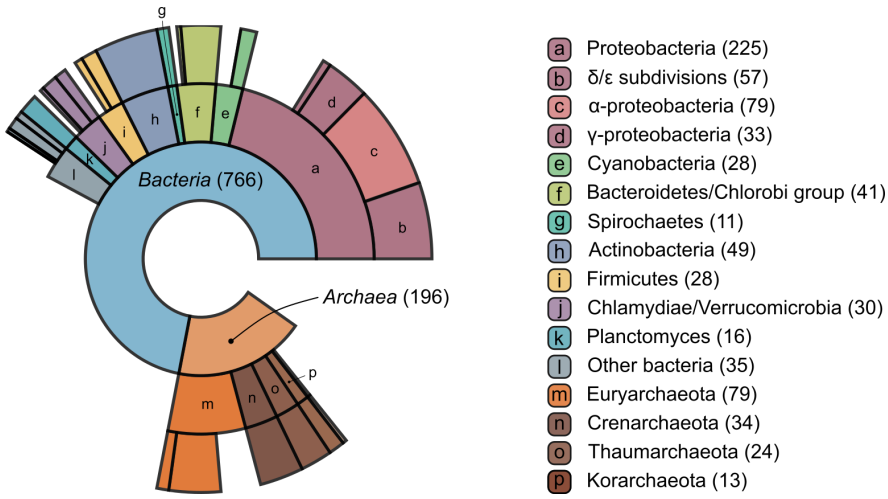


Figure 1.6: The mosaic eukaryotic genome. Ring plot showing the distribution of evolutionary affinities for the eukaryotic gene families assigned to LECA. The taxonomic affiliation of the sister clade in a ML phylogenetic reconstruction was assumed the putative origin of each family. Inner layers represent hierarchically lower (broader) taxonomic levels. The number of LECA families assigned to each group is indicated in parentheses next to the corresponding level in the ring plot or in the boxes below. Image adapted from Pittis and Gabaldón, 2016.

the modern endosymbiotic theory, advances in cell biology, comparative genomics and molecular phylogenetics methods, as well as increased sampling of microbial diversity in numerous environments have led to significant advances. A consensus has been reached in several aspects of the question of eukaryotic origins. It has been widely accepted that the process of endosymbiosis played a major role in shaping the eukaryotic genome (Timmis et al., 2004). It cannot be ruled out the possibility that other symbiotic interactions happened in the course of eukaryogenesis, but the α-proteobacterial ancestry of mitochondria is nowadays widely accepted and firmly established as a working hypothesis in practically all studies. The mixed archaeal and bacterial ancestry of Eukaryotes has also been confirmed with a variety of data and approaches, while there are strong hints to support a close link of Eukaryotes to Archaea (Yutin et al., 2008). Reconstructions of the ancestral genomic repertoire of LECA has unequivocally demonstrated that it was already a very complex organism. Most molecular machineries were already present, as well as all main eukaryotic features, such as nucleus, in-

trons, endoplasmic reticulum, peroxisomes, cytoskeleton and mitochondria (Koumandou et al., 2013). Many of the details of all these matters remain open and there is still much work to be done, but the most controversial questions at the center of the discussion relate to the nature of the host of the mitochondrial endosymbiont and the timing of the acquisition, the order of events that led to LECA (Poole and Gribaldo, 2014; López-García and Moreira, 2015). Regarding this aspect, as explained in section 1.4, many eukaryogenesis models that have been proposed could be generally grouped under two opposing views, the mito-late on the one side, and the mito-early on the other. Many models could hardly fall in any of the two categories, there is however an important conceptual point where they disagree. The mito-late view focuses on the gradual advancement in complexity, with sequential changes driving the process, while it considers that most part of the subcellular compartmentalization and organelles were already formed before the acquisition of mitochondria. Mitochondrial endosymbiosis is commonly explained mechanistically by phagocytotic capabilities of the proto-eukaryotic host. The mito-early view considers the mitochondrial endosymbiosis as the driving force of eukaryogenesis. For these models the increase in complexity is triggered by the endosymbiont, and all structures and organelles result as a response (Figure 1.5). An answer to this controversy is tightly related to all other aspects of the problem of eukaryogenesis, all the open issues are interconnected. Such as a late acquisition of mitochondria would point to a complex host and would suggest a phagocytotic mechanism for endosymbiosis, a complex host would similarly suggest a late mitochondrial acquisition. Advances in specific research lines that are related to the origin of Eukaryotes have an effect to all others, as they all form part of the same puzzle. The contribution of phylogenomics in all analyses has been instrumental, especially over the last years. Bridging the enormous amount of information generated as more genomic data become available, to knowledge of the biology of living organisms in the huge diversity that is being gradually revealed, is key to understand the origin of Eukaryotes. There are exciting perspectives that are currently unfolded, and comparative genomics and comparative biology generally, will certainly be at the core of them in the following years.

2 Objectives

- Estimate the degree of retargeting in eukaryotic protein families across different cell compartments.
- The comprehensive evolutionary analysis of the main molecular components of calcium homeostasis in mitochondria, and the detection of co-evolution patterns among them.
- Explore alternative methodologies for the analysis of the phylogenetic signal of protein sequences.
- Assess the mosaic phylogenetic signal carried by protein families inferred to the Last Eukaryotic Common Ancestor (LECA).
- Contrast the ancestry of LECA protein families to main hypotheses on the origin of eukaryotic complexity.
- Analyze the distribution of branch lengths from phylogenetic trees within a large-scale phylogenomic context, and evaluate it as a source of information for the understanding of the evolution of genes and lineages.

Part II

Results

3 Origin and evolution of metabolic subcellular compartmentalization in eukaryotes.

Gabaldón, T. & Pittis, A. A. (2015). Origin and evolution of metabolic subcellular compartmentalization in eukaryotes. Biochimie, 119, 262-268.

3.1 Abstract

A high level of subcellular compartmentalization is a hallmark of eukaryotic cells. This intricate internal organization was present already in the common ancestor of all extant Eukaryotes, and the determination of the origins and early evolution of the different organelles remains largely elusive. Organellar proteomes are determined through regulated pathways that target proteins produced in the cytosol to their final subcellular destinations. This internal sorting of proteins can vary across different physiological conditions, cell types and lineages. Evolutionary retargeting – the alteration of a subcellular localization of a protein in the course of evolution – has been rampant in Eukaryotes and involves any possible combination of organelles. This fact adds another layer of difficulty to the reconstruction of the origins and evolution of organelles. In this review we discuss current themes in relation to the origin and evolution of organellar proteomes. Throughout the text, a special focus is set on the evolution of mitochondrial and peroxisomal proteomes, which are two organelles for which extensive proteomic and evolutionary studies have been performed.

Keywords: Evolution, Eukaryotes, Compartments, Organelles

Gabaldón T, Pittis AA. Origin and evolution of metabolic sub-cellular compartmentalization in eukaryotes. Biochimie. 2015 Dec;119:262-8. doi:10.1016/j.biochi.2015.03.021

4 Co-evolution of mitochondrial Ca^{2+} uptake components

Pittis A. A., Perocchi F. & Gabaldón, T. (in preparation). Phylogenomics of mitochondrial calcium homeostasis.

4.1 Abstract

Calcium uptake by mitochondria is a key regulatory mechanism in Eukaryotes, which is relevant for a wide range of cellular functions, including cell death, bioenergetics and signalling. The recent identification of the mammalian mitochondrial calcium uniporter (MCU) and its regulator (MICU1), has provided long-awaited insights into the molecular bases of calcium homeostasis in mitochondria. Initial comparative analyses across 138 sequenced eukaryotes have revealed a broad taxonomic distribution of the proteins of the complex, suggestive of an ancient origin, but have uncovered some organismal groups where one or both components are missing. Here, we perform a comprehensive phylogenomics survey across an expanded dataset of 243 fully-sequenced eukaryotes of these two components as well as of the essential MCU regulator (EMRE) and the mitochondrial Sodium/Calcium exchanger. By complementing similarity searches with exhaustive phylogenetic analyses, our analyses uncover duplication events and properly delineate orthology relationships, as well as strong co-evolution patterns among interacting components of the complex. Our results show a patchy distribution of MCU and MICU1 in certain eukaryotic groups and indicate that the evolution of the mitochondrial calcium uniporter and the Na/Ca exchanger are largely uncoupled. In contrast, in Opisthokonts the distribution of MCU and MICU1 largely overlaps with that of EMRE, which suggests a close functional link in this group.

Keywords: Ca^{2+}, mitochondria, transporters, phylogenomics

4.2 Introduction

Calcium homeostasis is known to regulate diverse cellular functions (Berridge et al., 2003; Clapham, 2007). Besides their key role in energy metabolism, mitochondria participate in many signaling pathways, and serve as major regulators and targets of calcium signaling. Already more than 50 years ago, early studies suggested a close connection between mitochondria and calcium signaling, by showing that activated mitochondria take up large amounts of Ca^{2+} (Deluca and Engstrom, 1961; Vasington and Murphy, 1962). In the following years, the properties of mitochondrial calcium uptake were studied exhaustively. Transport across the inner mitochondrial membrane occurs via a "uniporter" and is driven by the membrane potential generated by the respiratory chain (Gunter and Pfeiffer, 1990). However, while the physiological role of the uniporter was studied extensively for years, its molecular identity remained elusive. Only recently, the gene coding for the calcium uniporter MCU and its regulator, MICU1, were identified (Perocchi et al., 2010; Baughman et al., 2011; Bick et al., 2012). Further analyses revealed that, in mammalian cells, paralogues of MCU and MICU1 are involved in the uniport complex (Plovanich et al., 2013; Raffaello et al., 2013). In addition, EMRE was identified as another essential component of this complex in human cells, and is required for the interaction of MCU and MICU1 (Sancak et al., 2013). The transport function of the uniporter is mediated by microdomains of high concentrations of Ca^{2+} in close junctions between the ER and mitochondria, the mitochondria-associated ER membrane (MAM) (Rizzuto et al., 1998; Szalai et al., 2000; Csordás et al., 1999). The activation of the inositol-1,4,5-trisphosphate-sensitive channels (IP3R) releases Ca^{2+}, which accumulation promotes the mitochondrial uptake of this cation through the low affinity MCU pore-forming protein. The main efflux path in mitochondria to the cytoplasm is thought to be NCLX, the mitochondrial sodium-calcium exchange protein.

Assessing the evolutionary patterns of the various components of a functional complex can help to understand their interactions. Initial experiments in fungal model species (Carafoli and Lehninger, 1971) indicated a lack of mitochondrial uniporter activity. More recently, comparative analyses across

138 sequenced eukaryotic genomes suggested an ancient eukaryotic origin for the main components of the complex (MCU and MICU1 families), based on their widespread distribution across Eukaryotes, and confirmed the loss of MCU in the whole Saccharomycotina clade. However MCU homologs were detected in other filamentous Ascomycota which were shown to also lack classical uniporter activity. The distribution of MICU1 homologs was shown to largely overlap with that of MCU across many groups, pointing to correlated evolutionary histories between the uniporter and its regulator, which supports the existence of a functional link between them. However, the absence of MICU1 homologs in Fungi constitutes a notable exception to this overlap.

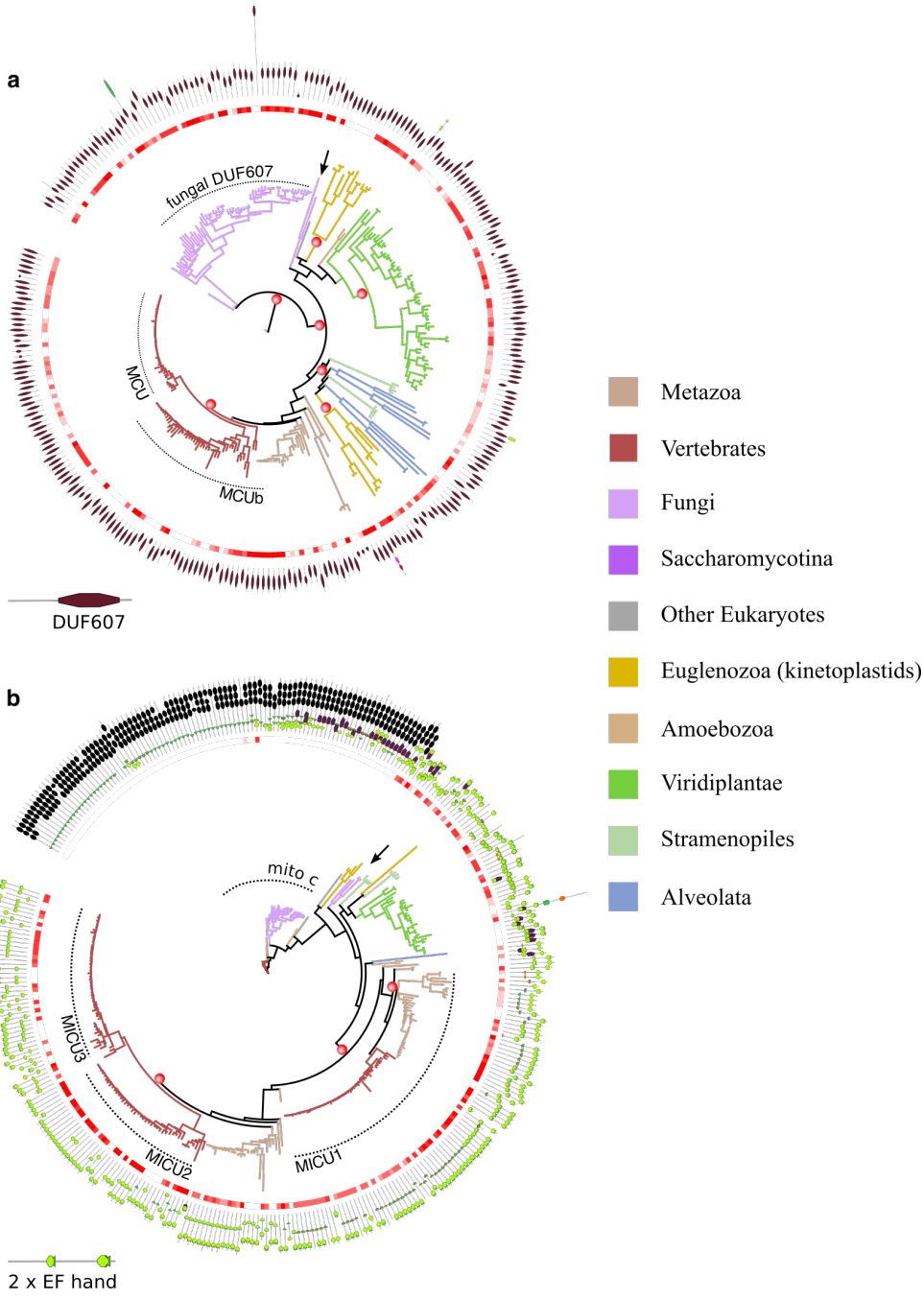
Co-evolution patterns can be very useful in predicting molecular interactions on the basis of the inter-dependence of proteins that function in cooperation (methods reviewed in de Juan et al., 2013). For instance, parallel events of gene duplication or loss can indicate strong functional associations, as they may result from evolutionary constrains to keep coordinated absence of presence in equal ammounts of both encoded proteins. The molecular characterization of members of the mitochondrial Ca^{2+} uniporter complex is just starting to provide answers to the long lasting question of its molecular identity. Our study provides novel insights into the evolution of the complex across a wide range of eukaryotic diversity.

4.3 Results and discussion

4.3.1 Phylogenomics survey across 243 fully-sequenced eukaryotes

Evolution of the uniporter complex components

Using a combination of profile-based sequence searches, assessment of protein domain composition, and phylogenetic analysis (see Material and Methods 4.4), we inferred the evolutionary history of the components of the mitochondrial calcium signalling pathway and their homologs encoded in the genomes of 243 fully-sequenced eukaryotes. We first focused on the two main components, known to be part of the uniporter complex across Eukaryotes, MCU and MICU1. Gene duplications and losses in the various lineages appear to have driven the evolution of these two protein families. In the case of the MCU family (Figure 4.1a), our phylogenetic analysis indicates that an initial duplication in the last common ancestor of eukaryotes (LECA) was followed by two main differential losses that distinguish fungi from the rest of eukaryotes. As a result most fungi retain one of the two paralogous subfamilies, while, conversely, only the other subfamily is retained in all other eukaryotes. In our dataset, only two early-diverging Fungi, the blastocladiomycete *Allomyces macrogynus* and the chytrid *Spizellomyces punctatus*, retained both copies. Given that many of the species that have the subfamily that is widespread in fungi are shown to lack uniporter activity, we will refer to the alternative subfamily, widespread in all other eukaryotic groups as the "true" MCU. Subsequent duplications, at the base of Kinetoplastids, Streptophyta, Chromalveolates and Vertebrates, as well as in other recent lineages, have created additional multiple paralogous copies of the protein in certain species. For instance, in *Trypanosoma brucei* we detected six sequences, whereas we found two in most plant species and in some Chromalveolates. Particularly the duplication of MCU at the base of vertebrates created the two paralogous genes, MCU and MCUb, which have been shown to interact (Raffaello et al., 2013; Sancak et al., 2013), although the regulatory mechanism used by MCUb remains uknown. Similarly, the MICU family, has diversified through consecutive gene duplications (Figure 4.1b). MICU members contain at least two EF-


a
fungal DUF607
MCU
MCUb
DUF607
b
mito c
MICU3
MICU2
MICU1
2 x EF hand
Metazoa
Vertebrates
Fungi
Saccharomycotina
Other Eukaryotes
Euglenozoa (kinetoplastids)
Amoebozoa
Viridiplantae
Stramenopiles
Alveolata

hand domains, which are very common in calcium-binding proteins. Within the mitochondria-localized orthologs of MICU1, an ancestral gene was duplicated once in the ancestor of Metazoa and further subsequent times at the bases of arthropods and vertebrates, resulting in three copies in most Metazoan species. In vertebrates, MICU2 was confirmed to be involved in the protein complex and to physically interact with other components, while MICU3 was shown to exhibit tissue specificity, being preferentially expressed in the central nervous system (Plovanich et al., 2013). Phylogenies of the other studied protein families also showed extensive gene duplications and losses (results not shown). Interestingly, the IP3R channels, which function as Ca^{2+} exporters from the ER towards the ER-mitochondria junctions, duplicated twice in the vertebrates ancestor, raising the possibility of correlated evolution between functionally inter-dependent proteins.

Distribution of the different components in Eukaryotes

The overall distributions of MCU and MICU1 are largely congruent, for the 138 shared species, with that of the above-mentioned previous genomics survey (Bick et al., 2012). We confirm the presence of MCU in at least some species of all major eukaryotic groups sequenced so far (Unikonts, Chromalveolates, Plants and Euglenozoa), and confirm the absence in all Apicomplexans, irrespectively of the presence of mitochondria, yeasts (Saccharomycotina and *Schizosaccharomyces* clades), Microsporidia, *Trichomonas* and *Giardia*. Moreover, our expanded dataset serves to detect additional tax-

Figure 4.1 *(preceding page)*: Maximum Likelihood phylogenetic trees of MCU (a) and MICU (b) families. The circular rooted trees of the main components of Ca^{2+} uniporter complex, DUF607/MCU family (**a**) and MICU family (**b**). The two families have been expanded through duplication rounds in various independent lineages. Major duplication events are indicated with a red sphere on the relevant tree node. The taxonomic group of the sequences is according to the color code in the legend (right). The red aligned boxes in different shades of red in the outer part, indicate "reliability class" or strength of the mitochondrial localization prediction, as reported by TargetP. Absence of a red box indicates a predicted localization in other compartments rather than mitochondria. In the outer layer, the domain architecture according to the HMMER prediction based on Pfam database is shown. On the bottom-left part of each tree there is a representative architecture of each family.

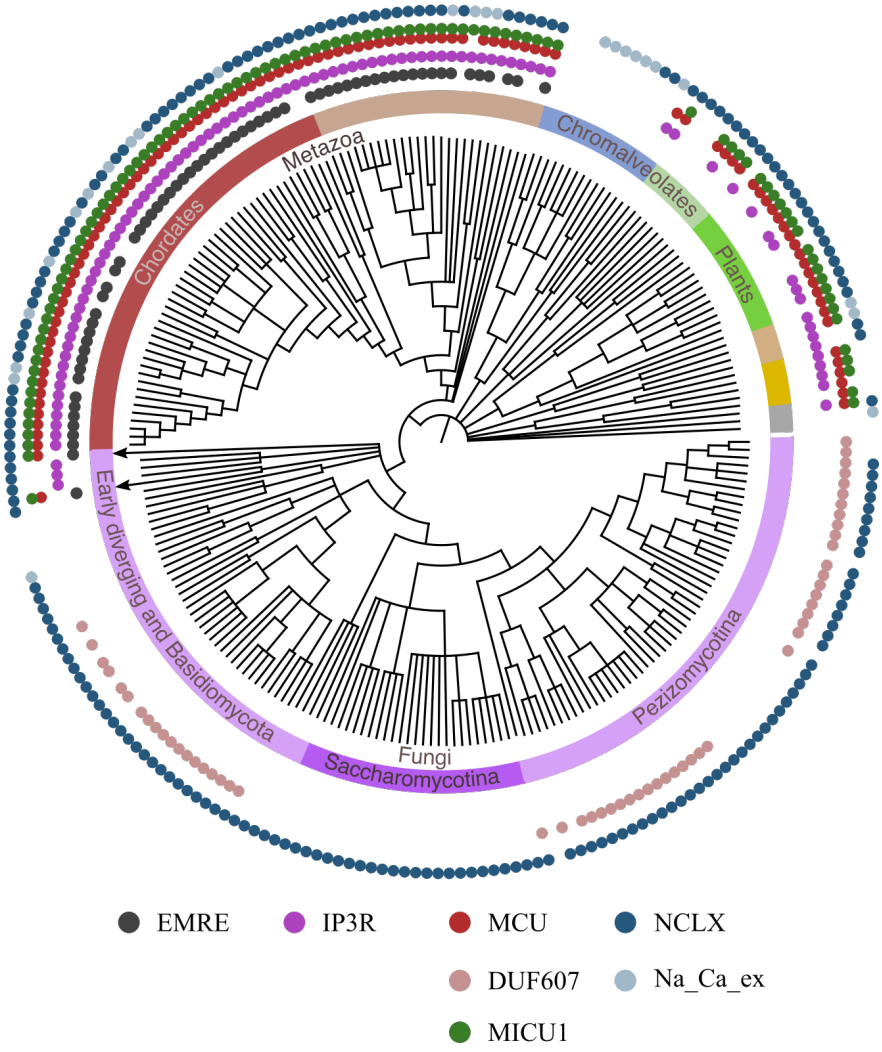


Figure 4.2: Phylogenetic distribution of Ca^{2+} transporters across Eukaryotes. The phylogenetic distribution across 243 eukaryotic genomes is shown on the NCBI taxonomy tree. The detection of one or more sequences in the genome is illustrated by a circle in the corresponding component/color. The components are color coded as shown in the bottom legend. With an arrow in the terminal leaf are indicated the only Fungi in which a "true" MCU, a MICU and an EMRE homologs were detected, *Allomyces macrogynus* (up) and *Spizellomyces punctatus* (down). From inner to outer components, EMRE : Essential MCU regulator, IP3R : IP3 (inositol 1,4,5-trisphosphate) receptor, MCU : "True" Mitochondrial Calcium Uniporter, DUF607 : DUF607-containing homologs of MCU, MICU1 : mitochondrial calcium uptake 1, NCLX : Mitochondrial sodium/potassium/calcium exchanger 6, Na_Ca_ex : Paralogous sodium/calcium exchangers.

onomic groups that do not present MCU, like diatoms. However, as we mentioned before, our analysis reveals a striking co-evolution pattern within Fungi, for which our dataset is particularly rich - with 121 species as compared to 50 in Bick et al., 2012. The only two early-diverging fungi with detectable MICU1 homologs, *Allomyces macrogynus* and *Spizellomyces punctatus*, are the ones with "true" MCU orthologs, while the rest retain only a fungal-specific copy (DUF607 domain containing protein, see also Figure 4.2a). Notably, this provides an explanation to earlier observations which had been said to represent an evolutionary paradox Kamer and Mootha (2015): the presence of uniporter homologs in species with no detectable mitochondrial Ca^{2+} uptake, all species lacking MICU1 homologs have a distant homolog to the uniporter of unknown function. Surprisingly, an additional chytrid species included only in our dataset (*Batrachochytrium dendrobatidis*), lacks both proteins, indicating that their presence is not widespread among basal fungi. Thus the two species mentioned above remain the only fungal species with the presence of both the uniporter and its regulator. The distribution of the DUF607-containing proteins is rather patchy with at least six entire fungal clades lacking any sequence, namely Microsporidia, Saccharomycotina, Onygenales, Zygomycotina, Taphrinomycotina, and Pucciniomycotina. Agaromycotina show a variable presence of this protein. The function of the DUF607-containing homologs of MCU in Fungi is unknown, thus the possibility that proteins of this clade are involved in Ca^{2+} transport through alternative mechanisms, cannot be excluded.

As to the profiles of the other proteins studied, EMRE, which was shown to be essential for the uniporter's functionality in human cells and ubiquitously expressed in mammals (Sancak et al., 2013), was found to be specific to Opisthokonts, widespread in Metazoa and present in Fungi only in the very same species containing MCU and MICU, *Allomyces macrogynus* and *Spizellomyces punctatus*. This striking, previously undetected, co-evolution pattern in Fungi between MCU, MICU and EMRE suggests a strong interdependence, especially given the fact that Allomyces and Spizellomyces are quite distantly related (Torruella et al., 2012). We complemented this analysis with manual searches in public databases of EMRE homologs in Eukaryotes, which confirmed its unique presence in these two species

among fungi. In addition, these searches also revealed the presence of EMRE in Choanoflagellates (*Monosiga brevicollis*) and Ichthyosporea (*Capsaspora owczarzaki*) pointing to an opisthokonta-specific origin, rather than metazoa-specific as previously thought. IP3R, whereas fully detectable in all Metazoa and generally overlapping with MCU and MICU1 in other species (undetectable cases could also be due to methodological reasons), in Fungi it is present only in Mucoromycotina, which lack other uniporter components. The observation that the presence patterns of the two proteins are mutually exclusive in early-diverging Fungi, suggests the existence of alternative calcium pathways at the fungal base. Finally the mitochondrial Na/Ca exchanger (NCLX) shows a patchy distribution, largely uncoupled to the presence of MCU and MICU genes. Moreover its presence in some mitochondria-lacking lineages, like *Cyptosporidium* and *Entamoeba*, and its predicted non-mitochondrial localization out of Fungi, questions the hypothesis that it plays a universal role in mitochondrial calcium homeostasis.

4.3.2 Mitochondrial calcium signalling is an ancestral eukaryotic feature

Homologs of the mitochondrial calcium uniporter are found in all major eukaryotic groups sequenced so far: namely Unikonts, Chromalveolates, Viridiplantae, and Excavata. The current phylogenetic distribution strongly suggests that calcium signalling was present in the mitochondrion of the last common ancestor of eukaryotes (LECA), the so-called protomitochondrion (Gabaldón and Huynen, 2004), and that mitochondrial calcium signalling is an ancient hallmark of the eukaryotic cell. We could only detect the presence of the MCU (DUF607 domain) in 3 out of 1,100 bacterial genomes analyzed. Thus MCU is most probably an early eukaryotic invention. This eukaryotic signature of a key regulatory mechanism in mitochondria is consistent with earlier reports, showing that pathways controlling mitochondrial biogenesis and function are generally of eukaryotic descent, whether acquired *de novo* or by replacement of the bacterial-derived counterpart (Gabaldón and Huynen, 2007). Thus, in principle one cannot discard that a similar system was already present in the alpha-proteobacterial ancestor. The existence of

calcium signalling in bacteria is a matter of discussion, but there is evidence that calcium levels are tightly regulated in bacterial cells and that this ion may play a role in cell morphogenesis and division (Dominguez, 2004). Thus replacement of the calcium import mechanism would have enabled the host to take control over any calcium-dependent mechanism already present in the endosymbiont. Alternatively, a somewhat less parsimonious scenario would involve the early re-targeting to the organelle of a calcium-dependent pathway coupled to the emergence of the uniporter system. Note that these two alternative hypotheses have fundamental implications on the predicted evolutionary origin (α-proteobacterial or eukaryotic) of the originally targeted pathway. We hypothesize that, as seems to be the case in extant prokaryotes, the ancestral α-proteobacterial endosymbiont already displayed calcium-dependent processes, perhaps already coupled to fluctuations in the hosts' cytoplasm, and that the acquisition of MCU enabled full control by the host.

On the other hand, calcium signalling appears to have been lost many times independently during the evolution of eukaryotes. Importantly, a significant number of these losses correlate with extreme streamlining of mitochondrial metabolism. Indeed, all five lineages encompassing relict forms of anaerobic mitochondria such as mitosomes (Microsporidians, *Entamoeba*, *Giardia*, *Cryptosporidium*) or hydrogenosomes (*Trichomonas*) seem to have lost the mitochondrial Ca^{2+} uniporter and its main regulator MICU. In all these lineages, with the unexpected exceptions of *Cryptosporidium* and *Entamoeba* mitosomes, loss of MCU is coupled to the loss of a mitochondrion-targeted Na/Ca exchanger (NCLX). This suggests that calcium signalling in the organelle is completely lost and thus it is unnecessary to regulate the few mitosome-harbored metabolic pathways (e.g Fe-S cluster assembly), and raises the question of the function of Na/Ca exchanger in the mitosomes of these two species. At least seven other lineages have lost MCU or DUF607-containing homologs without evolving mitochondria into mitosomes or hydrogenosomes: namely diatomes, Plasmodium, Toxoplasma, and several groups of fungi including all sequenced Onygenales, Saccharomycetales, Schizosaccharomycetes and Zygomycotina. The patchy distribution in fungi is remarkable with most non-filamentous, and few filamentous fungi

having lost them. Further work will be needed to clarify 1) what is the function of the fungal DUF607-containing proteins and 2) what physiological adaptations have driven their loss in such diverse phyla. All these lineages generally retain the mitochondrion-targetted Na/Ca exchanger, suggesting that Calcium extrusion is important even in the absence of active import by MCU, perhaps to compensate calcium entry from alternative importers or leakage from the cytoplasm. Conversely, some lineages such as Euglenozoa (e.g trypanosomes) harbor MCU in the absence of a mitochondrially-targeted Na/Ca exchanger, raising the question of what drives Ca^{2+} extrusion in these organisms. Altogether our results show that the evolution of the mitochondrial calcium uniporter and the Na/Ca exchanger are largely uncoupled.

4.3.3 Concluding remarks

The evolution of the MCU and MICU families are highly correlated, which highlights their importance and inter-dependence as parts of the uniporter complex. Their presence in all extant eukaryotic groups points to an ancestral origin of the complex, already in the Last Eukaryotic Common Ancestor (LECA), and that additional regulatory components, like EMRE with no detectable homology out of Metazoa, appear to be later lineage-specific innovations. Our phylogenetic analysis indicates that the MCU homologs in Fungi, apart from some specific early-diverging species, are distant paralogs originating from an ancient duplication in LECA, and that all other eukaryotic lineages lost this copy. This, together with the absence of MICU in most fungal lineages, and the previously undetectable uniporter activity in some of them, suggests that Ca^{2+} uptake in most Fungi, if any, is not mediated by the same pathway as in other Eukaryotes. Streamlining of mitochondria as a result of extreme adaptation to anaerobic environments is always associated to loss of MCU homologs in multiple lineages. Remarkably all sequenced species harboring mitosomes or hydrogenomes have lost MCU. All except *Cryptosporidium* and *Entamoeba* mitosomes have also lost NCLX. The profiles of the main components of the complex and NCLX are largely uncoupled and there are organism with solely the presence of either of these. This raises the question of what drives Ca^{2+} extrusion

in organisms with MCU such as Euglenozoans (e.g. Trypanosomes). On the opposite side, organisms that lost any member of the MCU family while generally retaining a mitochondrial NCLX are Diatomes, *Plasmodium*, *Toxoplasma*, and several groups of fungi including all sequenced Onygenales, Saccharomycetales, Schizosaccharomycetes, and Zygomycotina.

After decades of research on the mitochondrial Ca^{2+} uniporter, the discovery of the molecular identity of of its main components has opened new research avenues. By using comparative genomics techniques, predictions can be made and future experiments can be driven by predictions made taking into account the evolutionary context of the pathway. Our phylogenomic analysis indicates that mitochondrial calcium signaling, and its strongly interconnected protein components, is a eukaryotic hallmark. It was invented early in the evolution of Eukaryotes and being involved in so diverse important functions, it probably played a prominent role in their evolutionary success. Disentangling the evolution of the uniporter's components can provide significant information and mark novel paths in its research.

4.4 Methods

Sequence data

The protein sequences encoded in 243 completely sequenced eukaryotic genomes were retrieved from various database sources (see Appendix Table A1). The human sequences for the proteins under study were retrieved from Uniprot. These were used for the protein domain identification or as queries in BLAST similarity searches, for the detection of homologues across the 243 genomes. TargetP 1.1 (Emanuelsson et al., 2000) was used for the prediction of sub-cellular localization of proteins.

Homology determination and annotation

For each of the five protein families studied, homologs were selected on the basis of sequence similarity and phylogenetic analysis. HMMER searches were performed using HMMER 3.0 (Eddy, 2011) and using the Gathering Cut-Off threshold (–cut_ga). To minimise the number of false positives and BLAST searches were performed by filtering out low complexity regions in the query sequence (default parameter) and an E-value threshold of 10^{-5}. Domains in all retrieved sequences, were annotated using the HMM profiles of Pfam release 26.0.

MCU Proteins in our database containing at least one DUF607/MCU (PF04678) Pfam domain were detected using HMMER 3.0. The DUF607 domain has two transmembrane regions with two coiled coil motifs. 286 protein sequences were selected for subsequent analysis.

MICU1 348 protein sequences were retrieved from a BLAST search, using the above parameters. HMMER 3.0 was used to search for additional domains in the retrieved sequences using the all the Pfam domain profiles and all the sequences with at least one Mito_carr domain were excluded, as members of the mitochondrial carrier family (slc25a12-Aralar homologs). The filtering was confirmed through the phylogenetic tree where all the Mito_carr containing protein clustered together.

NCKX6 Na_Ca_ex Pfam domain characterizes an ubiquitous superfamily of sodium/calcium exchangers that regulate intracellular Ca^{2+} concentrations in many cell types. Therefore the selection on that element of the NCKX6 (NCLX) homologs with HMMER leads to the identification of 1,694 Na_Ca_ex possessing sequences in the 243 species. To narrow down this number and extract only the closer to NCKX6 homologs, we applied a coverage threshold of 50% over the query sequence. By that we selected 719 sequences that were further evaluated.

ITPR3 Homologs of ITPR3 were defined based on the two domains characterizing the inositol-1,4,5- trisphosphate receptors and the ryanodine receptors as well, which are the RIF domain (RYDR_ITPR - PF01365) and the 1,4,5-trisphosphate/ryanodine receptor domain (Ins145_P3_rec - PF08709). Sequences in which at least one of the two domains was present were selected for the analysis.

EMRE EMRE sequences are characterized by a DDDD (PF10161) Pfam domain. HMMER searches with the previous thresholds were performed, and all detected homologs were retrieved for subsequent analysis.

Phylogenetic analysis

We reconstructed phylogenetic trees for all four protein families described above. The selected homologous proteins were aligned with MAFFT v6.901b (E-INS-i and G-INS-i options) (Katoh and Toh, 2008) and positions in the alignment with gaps in more than 90% of the sequences were removed with trimAl (Capella-Gutiérrez et al., 2009). ProtTest 3.2 (Darriba et al., 2011) was used for the selection of the best-fit evolutionary model among the three tested (LG, WAG, JTT) comparing the likelihood values. RAxML v7.2.8 (Stamatakis, 2014) and the rapid hill climbing algorithm was used to derive Maximum Likelihood (ML) trees. Branch support was calculated using the rapid bootstrap algorithm of the same program. A discrete gamma-distribution model with four rate categories plus invariant positions was used, estimating the gamma parameter and the fraction of invariant positions from the data.

5 Analysis of LECA repertoire based on sequence similarity profiles

Pittis A. A. & Gabaldón, T. (in preparation). Assessing the origins of the genetic repertoire of the Last Eukaryotic Common Ancestor through the analysis of similarity distributions.

5.1 Abstract

The question of the origin of eukaryotes has long been one of the most difficult problems in evolutionary biology. Previous studies have inferred the origins and genome composition of the Last Eukaryotic Common Ancestor (LECA) by examining the evolutionary signal carried by modern genomes. However, such inference is challenged by a number of conceptual and technical difficulties. Phylogenetic analyses can be problematic when considering very ancient relationships, as much of the signal is saturated and prone to phylogenetic artefacts. Moreover, widespread horizontal gene transfer (HGT) among prokaryotes can make the interpretation of conflicting gene trees very complex. Similarity searches have been used as an alternative approach to infer phylogenetic origins, but these can also suffer from artefacts, especially when only the first hits in a search are considered. We here attempt to circumvent these problems and propose a new statistical framework to analyze the whole range of hits from a similarity (Blast or HMM-profile) search in a taxonomic context. We apply this framework to the exploration of prokaryotic origins of widespread eukaryotic protein families and we compare with phylogeny-based methods. We detect a diversity of prokaryotic signals in LECA which points to complex scenarios of eukaryogenesis and supports genetic contributions from different prokaryotic sources. Besides investigations on the origin of eukaryotes, our methodology is suited to other applications

such as the detection of horizontally transferred genes or the taxonomic classification of metagenomic data.

5.2 Introduction

The problem of the origin of Eukaryotes has been puzzling evolutionary biologists for decades. The revival of the endosymbiotic theories for eukaryogenesis from Lynn Margulis in the late 1960s (Sagan, 1967), brought the topic back to the scientific mainstream discussion, and highlighted the importance of symbiosis in processes of evolutionary innovation (Latorre and Moya, 2013). More recently, advances and methodological throughputs in genomics, computing, and statistical analyses, have paved the way to study the evolutionary history of life through the analysis of completely-sequenced genomes. Soon after the establishment of phylogenetic techniques as a powerful framework for inferring the history of genes and species, it became clear that different approaches can often provide contradicting results (Jeffroy et al., 2006). The evolutionary signal carried by sequences can be blurry, methodological artefacts pose a challenge in very ancient relationships (Philippe et al., 2011), and the complexity revealed in genomic information is often difficult to interpret. For instance, the misplacement of fast evolving sequences together, known as the "long branch attraction" (LBA) artefact has been well documented (Bergsten, 2005), and some authors have expressed doubts on whether sequences retain valid phylogenetic information on events that date back to 2 billion years ago or more (Gribaldo et al., 2010; Philippe et al., 2011). Importantly, extensive Horizontal Gene Transfer (HGT), the lateral transmission of genes among distinct organisms, especially common in bacteria, makes significantly more difficult to disentangle the relationships between species (Ochman et al., 2000). Inferences on the origin of genes based on sequence similarity searches, have been used as an alternative to the sophisticated and computationally intense phylogenetic reconstruction methods, and have been shown to be congruent to some extent (Esser et al., 2004). However, they are also generally considered to be only roughly approximating the phylogenetic position of sequences, and the best hit in such searches is often not the nearest neighbor in phylogenetic reconstructions (Esser et al., 2004).

Despite the observed incongruences among the different methodologies that have been used to assess the origin of the eukaryotic genome, all previous

genome-wide analyses consistently suggested that eukaryotic genomes are mosaics composed of genes of different evolutionary origins (Ribeiro and Golding, 1998; Rivera et al., 1998). Inferences on the ancestral genomic composition of LECA (gene families that are assumed to derive from the common ancestor of Eukaryotes due to their widespread distribution) have shown that three major components can be generally distinguished: a) a eukaryotic-specific subset with no detectable prokaryotic counterparts detected; b) a set of genes with bacterial sequences as the closest prokaryotic homologs; and c), a subset of genes with archaeal homologs as the closest prokaryotic sequences. Moreover, it has been observed that these subsets are not functionally equivalent, as the bacterial component is strongly enriched in genes of metabolic functions, whereas the archaeal one, which is two to three times smaller, mostly consists of genes related to informational processes, meaning replication, transcription and translation. Such mosaic nature of the ancestral eukaryotic genome pointed to genome fusion scenarios, portraying a merging between archaeal and bacterial partners for the formation of a hybrid cell that gave rise to LECA. However the quest for characterizing the partners of such a fusion event revealed an unexpected complexity, the signal of the inferred ancestries is dispersed across bacterial and archaeal taxonomic groups, rather than suggesting specific prokaryotic partners (Koonin, 2010). Several proposals have been put forward to explain this paradox, some that consider that this complexity reflects, at least to some extend, biological complexity, and others that have interpreted it as a result of methodological and biological ”noise”. The former models attribute the complicated signal to sequential endosymbiotic events or HGT to the eukaryotic lineage, due to an assumed phagotrophic lifestyle of proto-eukaryotic cells (Doolittle, 1998; Lester et al., 2006; Ku et al., 2015). The latter ones invoke widespread HGT among prokaryotes and limitations in phylogenetic methods, and suggest that most genes originate from the α-proteobacterial endosymbiont and an archaeal host, in its simplest formulations.

Here, we use a novel approach based on the taxonomic-aware analysis of similarity distributions, to assess the phylogenetic origin of eukaryotic protein families inferred to be present in LECA and we contrast these results

with existing hypotheses on the origin of eukaryotes. Our analysis supports a multitude of signals and points to a complex composite origin of the LECA gene reportoire. The proposed algorithm will be implemented as a bioinformatics tool that will enable the phylogenetic profiling of proteins and protein families by comprehensive exploration of similarity searches on the broadest taxonomic context. Besides investigations on the origin of eukaryotes, our methodology is suited to other applications such as the detection of horizontally transferred genes or the taxonomic classification of metagenomic data.

5.3 Results and discussion

5.3.1 Phylogenetic profiling of proteins through a novel statistical framework

We developed a novel algorithm to analyze, in a taxonomic context, the whole range of hits from a similarity (Blast or HMM-profile search) search. Briefly, the algorithm proceeds as follows: starting from a query sequence (Blast) or a set of homologous sequences (HMM-profile), a sequence database is queried for the retrieval of homologous hits. Rather than taking into account only for the first hit, or for an arbitrary number of the first top scoring hits, the whole distribution of scores is considered. This is done by accounting their relative position in the NCBI taxonomy structure. To do this, the NCBI taxonomic tree structure is first pruned to only contain the species where significant hits were obtained, and the tree is annotated with the scores obtained within the different taxonomic groups. In order to find the taxonomic level with the highest scores, the tree is traversed from root to tips, iteratively. At each internal node (taxonomic level) the scores in the two or more taxonomic partitions defined by that taxonomic level (e.g. Ascomycetes and Basidiomycetes are the two groups contained by Dikarya) are compared in an statistical framework (see Methods). If there is a group with statistically significant higher scores as compared to the rest, this group is selected for the next iteration and the process continues within that clade, iteratively, until there is no significant difference detected between the groups or the number of hits is too small for a statistical test. This procedure ensures the detection of the the most specific taxonomic group which shows a statistically significant higher degree of similarity as compared to other taxonomic groups of the same level, which we use as the best proxy for phylogenetic origin. Our procedure is expected to produce broader, less-specific, assignments, rather than incorrect ones, in the presence of stochastic noise or blurred signal.

5.3.2 LECA proteome shows complex origin from various prokaryotic sources

Out of the 238 eukaryotic species present in eggNOG v 4.0 (Powell et al., 2013), 46 species were selected (Table 5.1), representing the broadest possible taxonomic representation available in the database, namely seven major groups, Opisthokonta, Stramenopiles, Viridiplantae, Amoebozoa, Alveolata, Euglenozoa and Parabasalia. Among all the eukaryotic clusters of homologous groups (euNOGs), we selected only those 1,181 which contained at least 10 eukaryotic proteins, and at least one protein from each of the major groups. This very strict LECA definition was chosen in order to narrow our focus to the widely-distributed protein families. The sequences of each NOG were aligned and an HMM profile was built. This HMM profile was used to search for potential homologs against the prokaryotic sequences of Uniprot database. Families with less than 10 detected prokaryotic homologs were characterized as eukaryotic-specific and for the remaining, considered of prokaryotic origin, we used our phylogenetic profiling methodology to infer their most specific prokaryotic ancestry.

Overall, our result (Figure 5.1) is in agreement with previous studies and reflects the hybrid nature of the ancestral eukaryotic proteome, the 1,181 families are divided into eukaryotic-specific (269), bacterial (515) and archaeal (359) ancestry, with 38 families without specific prokaryotic assignment. The analysis of the functions assigned to the different family subsets, showed that the bacterial component is enriched in metabolic processes, whereas the archaeal component is enriched in informational processes. Finally, functions related to ”cell motility” are over-represented among the eukaryotic-specific families. Given our overly stringent thresholds, the LECA families of bacterial ancestry are probably under-represented in our dataset, mainly due to the loss of mitochondria-associated genes involved in respiration, in groups of anaerobic organisms that secondarily lost mitochondria. Interestingly, this criteria seem to affect families of different bacterial origin differently. We detect only 19 families with inferred α-proteobacterial origin, while families of different bacterial-inferred origin appear to be more conserved

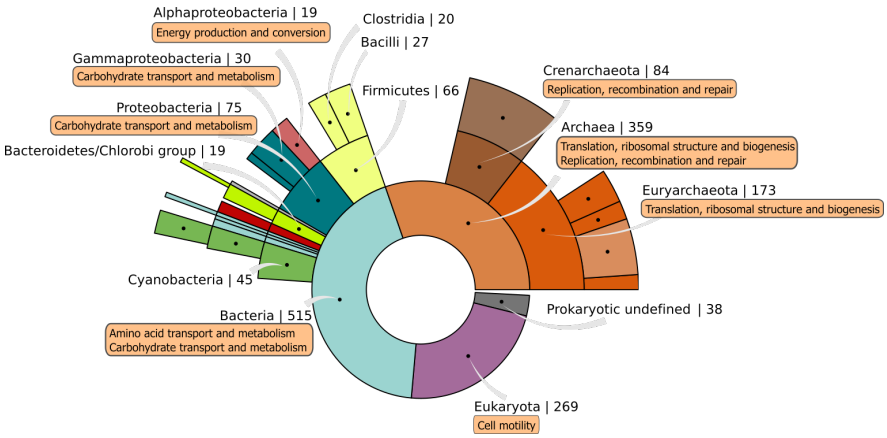


Figure 5.1: Phylogenetic origin of eukaryotic gene families. The distribution of ancestries of the analyzed families is shown in a ringchart. The layers in the plot follow the hierarchical structure of the NCBI taxonomy, with the inner sections representing higher taxonomic levels and containing the subsequent lower levels. The number of families mapped to some representative groups is shown next to the group name, and KOG functional categories enriched in a given group (if any) is shown at the top.

across all Eukaryotes, including species that lack mitochondria, possibly indicating a more ancient origin of this proteome. The archaeal component is distributed among Crenarchaeota and Euryarchaeota, with the latter Archaeal clade being twice more numerous (84:173). All in all, we detect a diversity of prokaryotic signals in LECA, including three major bacterial and two main archaeal groups, which suggests genetic contributions from different prokaryotic sources during eukaryogenesis.

5.3.3 Using the signal in plants and rickettsia as positive controls

To validate our approach, we applied our pipeline in cases with an expected distribution of gene origins. First we analyzed 172 families, eukaryotic NOGs specific to Viridiplantae among Eukaryotes (Figure 5.2a). Roughly half of the families were found to be lineage specific, representing putative plant innovations with no homologs in any other group. The prokaryotic component, as expected, was found to be largely of cyanobacterial origin (60 out of 76), presumably consisting of families originating from the ancestor

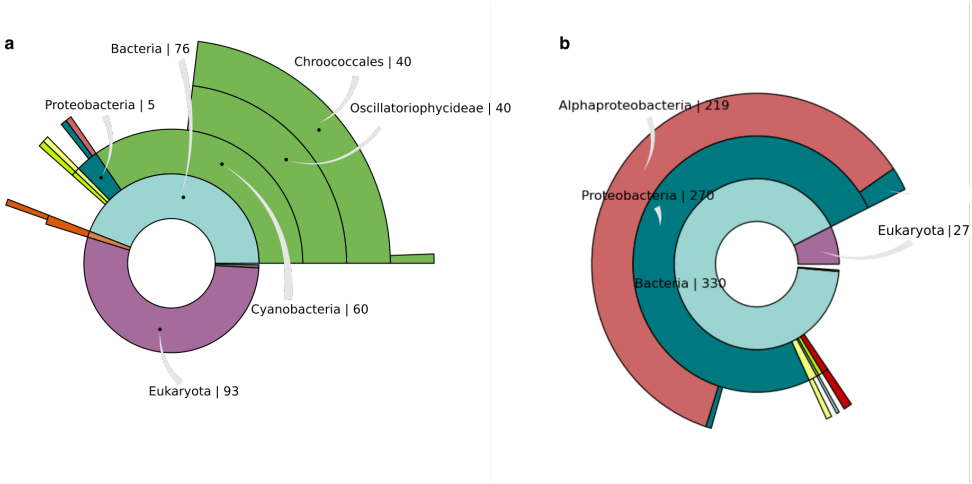


Figure 5.2: Profiles of plant specific (a) and Rickettsiales (b) families. a, The profile of 172 families, specific to Viridiplantae, indicates that their expected cyanobacterial ancestry is correctly inferred, irrespectively of the HGT that might have happened between cyanobacteria and other bacterial lineages, since the plastid endosymbiosis occured. **b,** Familes that are widespread in Rickettsiales are predominantly correctly inferred to α-proteobacterial ancestry in the absence of sequences from Rickettsiales in the database. Both (**a**) and (**b**) serve as positive controls for the validation of the method.

of plastids that were not found in other lineages that acquired plastids through secondary endosymbiosis. Few other groups are inferred as source of other families, something expected in the context of lineage specific adaptation, but the cyanobacterial component was clearly the dominant signal. Secondly, we assessed the signal carried by bacterial families, to see to what extend our method can correctly predict the phylogenetic affiliation of genes, irrespective of the potential conflict of signal caused by HGT. We selected the 361 families that were present in 80% of the Rickettsiales species available in the database. We extracted the corresponding Ricketsiales sequences, and built an HMM profile for each of the families. We used these profiles to query a sequence database in which all the sequences belonging to Rickettsiales had been removed. The α-proteobacterial nature of the families was correctly inferred (Figure 5.2b), showing that HGT can blur the signal in particular cases, but not to the extent that has been assumed by some authors (Thiergart et al., 2012). Finally, similar results were obtained

when we applied the method to mitochondria- and chloroplast-encoded gene families (results not shown), as they were vastly predicted to be of bacterial/α-proteobacterial (79.1%) and bacterial/cyanobacterial (94.4%) origin, corresponding to the known evolutionary origin of these organelles (Gray, 2012; de Alda et al., 2014).

5.3.4 Concluding remarks

We have developed a novel methodology to infer the evolutionary origin of genes, through the analysis of similarity search results under a taxonomic context. We use this methodology as an alternative to standard phylogenomic methods, to assess the ancestry of the genomic repertoire of the LECA. Our method is fast and efficient compared to the computationally demanding phylogenetic reconstruction pipelines, as it depends mostly on fast similarity searches (HMMsearch in this study). In addition, this approach is less prone to some common causes of phylogenetic conflict, as LBA and HGT, as it relies on the general signal of whole taxonomic groups. In that sense, we expect that our approach provides less resolution as compared to more sophisticated evolutionary modelling, but fewer specific erroneous inferences. We detect a large diversity of evolutionary signals in the reconstructed LECA proteome, especially in the bacterial derived component, groups other than α-proteobacteria seem to have a large contribution in our specific dataset. This is to some extend due to our strict LECA definition – which requires gene families to be present in all main eukaryotic clades, including mitochondria-lacking lineages. This result, suggests that genes that came with the mitochondrial ancestor are tightly associated to the presence of this organelle, whereas others appear to be involved in more general functionalities. Our positive controls indicate that our method successfully detects a dominant origin, in cases where such would be expected, which argues for the signal diversity in LECA being, at least for a major part, valid and a result of evolutionary ancestry, rather than an artefact resulting from HGT, or methodological inaccuracies. Altogether our results provide support to complex eukaryogenesis scenarios, where the lineage leading to LECA acquired genes from various prokaryotic sources in addition to the mitochondrial ancestor.

5.4 Methods

Method description

In brief, the algorithm proceed as follows: from a seed sequence or a seed group of sequences (e.g. a cluster of orthologous genes), homologs are retrieved with a sequence or profile-based similarity search; then, hits are mapped onto the NCBI taxonomy tree, according to their annotated organism source. We implemented Blast (Altschul et al., 1990) as the sequence-based and HMMER 3.0 (Eddy, 2011) as the profile-based similarity search engines, but in principle other search algorithms could be used. Hits can be filtered based on their coverage and/or E-value. Then taxonomic affiliations in forms of NCBI taxid are retrieved from the annotation of the hits in order to form a table of hit name, e-value, and taxid for all hits. A NCBI taxonomy tree containing as terminal nodes only those taxa present in the hit set is built using ETE3 Huerta-Cepas et al. (2016). This structure, which effectively provides a tree representation of similarity degrees of the query with hits across the entire taxonomic scope, is used in downstream analyses.

To determine the closest relative among a given number of clades (e.g. the closest prokaryotic clade to a given eukaryotic gene family), the algorithm starts traversing the tree structure, starting from the root. At each bi- or multifuraction, it compares the two (or more) populations of hits defined by each branch (clade), in terms of their similarity scores (e-value or scores). This is done using the non parametric Mann–Whitney U test. If the hits within one of the clades are found to have significantly (depending on the P-value threshold set) lower e-values (or higher scores) than the other(s), then that clade is chosen for the next traversing step. The first bi- or multifurcation within that clade is tested as described before and the procedure continues until one depth is found that contains no sub-clades having significantly lower e-values than the rest, or the number of hits in the sub-clades is less than 10 and therefore the test cannot be performed. Then that clade is selected as the best inference for the significantly closest clade to the query sequence.

Sequence data

405,196 protein sequences in 4851 eukaryotic orthologous groups from 46 species (see Table 5.1) present in eggNOG v 4.0 (Powell et al., 2013) database, were selected to represent eukaryotic diversity. The Uniprot database - release 2013_12, was used for homology searching, and sequences belonging to specific groups were excluded, as needed for each analysis.

Code availability

All the code used in the analysis was implemented in python and is available upon request. Functions from the SciPy python package (Oliphant, 2007) were used for the statistical tests, and the matplotlib python library (Hunter, 2007) was used to produce the plots.

Taxid	Species name	Eukaryotic group
10116	*Rattus norvegicus*	Opisthokonta
10090	*Mus musculus*	Opisthokonta
9606	*Homo sapiens*	Opisthokonta
9823	*Sus scrofa*	Opisthokonta
9031	*Gallus gallus*	Opisthokonta
8364	*Xenopus (Silurana) tropicalis*	Opisthokonta
7955	*Danio rerio*	Opisthokonta
7739	*Branchiostoma floridae*	Opisthokonta
7719	*Ciona intestinalis*	Opisthokonta
121224	*Pediculus humanus corporis*	Opisthokonta
7165	*Anopheles gambiae*	Opisthokonta
7227	*Drosophila melanogaster*	Opisthokonta
6239	*Caenorhabditis elegans*	Opisthokonta
6182	*Schistosoma japonicum*	Opisthokonta
45351	*Nematostella vectensis*	Opisthokonta
10228	*Trichoplax adhaerens*	Opisthokonta
5141	*Neurospora crassa*	Opisthokonta
5061	*Aspergillus niger*	Opisthokonta
162425	*Aspergillus nidulans*	Opisthokonta
4932	*Saccharomyces cerevisiae*	Opisthokonta
559307	*Zygosaccharomyces rouxii CBS 732*	Opisthokonta
5476	*Candida albicans*	Opisthokonta
4896	*Schizosaccharomyces pombe*	Opisthokonta
81824	*Monosiga brevicollis*	Opisthokonta
67593	*Phytophthora sojae*	Stramenopiles
164328	*Phytophthora ramorum*	Stramenopiles
2850	*Phaeodactylum tricornutum*	Stramenopiles
39947	*Oryza sativa Japonica Group*	Viridiplantae
15368	*Brachypodium distachyon*	Viridiplantae
4558	*Sorghum bicolor*	Viridiplantae
29760	*Vitis vinifera*	Viridiplantae
3702	*Arabidopsis thaliana*	Viridiplantae
3218	*Physcomitrella patens*	Viridiplantae
70448	*Ostreococcus tauri*	Viridiplantae
3055	*Chlamydomonas reinhardtii*	Viridiplantae
44689	*Dictyostelium discoideum*	Amoebozoa
294381	*Entamoeba histolytica HM-1:IMSS*	Amoebozoa
5911	*Tetrahymena thermophila*	Alveolata
5807	*Cryptosporidium parvum*	Alveolata
5833	*Plasmodium falciparum*	Alveolata
5855	*Plasmodium vivax*	Alveolata
5874	*Theileria annulata*	Alveolata
5875	*Theileria parva*	Alveolata
5691	*Trypanosoma brucei*	Euglenozoa
5671	*Leishmania infantum*	Euglenozoa
5722	*Trichomonas vaginalis*	Parabasalia

Table 5.1: The selected 46 representative eukaryotic species used in the study.

6 Relative timing of mitochondrial acquisition

Pittis A. A. & Gabaldón, T. (2016). Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature, 531(7592), 101-104.

6.1 Abstract

The origin of eukaryotes stands as a major conundrum in biology (Koonin, 2010). Current evidence indicates that the last eukaryotic common ancestor already possessed many eukaryotic hallmarks, including a complex subcellular organization (Koonin, 2010; Embley and Martin, 2006; Koumandou et al., 2013). In addition, the lack of evolutionary intermediates challenges the elucidation of the relative order of emergence of eukaryotic traits. Mitochondria are ubiquitous organelles derived from an alphaproteobacterial endosymbiont (Gray et al., 1999). Different hypotheses disagree on whether mitochondria were acquired early or late during eukaryogenesis (Poole and Gribaldo, 2014). Similarly, the nature and complexity of the receiving host are debated, with models ranging from a simple prokaryotic host to an already complex proto-eukaryote (Koonin, 2010; Koumandou et al., 2013; Martijn and Ettema, 2013; Lester et al., 2006). Most competing scenarios can be roughly grouped into either mito-early, which consider the driving force of eukaryogenesis to be mitochondrial endosymbiosis into a simple host, or mito-late, which postulate that a significant complexity predated mitochondrial endosymbiosis (Koumandou et al., 2013). Here we provide evidence for late mitochondrial endosymbiosis. We use phylogenomics to directly test whether proto-mitochondrial proteins were acquired earlier or later than other proteins of the last eukaryotic common ancestor. We find that last eukaryotic common ancestor protein families of alphaproteobacterial ancestry and of mitochondrial localization show the shortest phylogenetic distances

to their closest prokaryotic relatives, compared with proteins of different prokaryotic origin or cellular localization. Altogether, our results shed new light on a long-standing question and provide compelling support for the late acquisition of mitochondria into a host that already had a proteome of chimaeric phylogenetic origin. We argue that mitochondrial endosymbiosis was one of the ultimate steps in eukaryogenesis and that it provided the definitive selective advantage to mitochondria-bearing eukaryotes over less complex forms.

Subject terms: Evolution, Evolutionary theory, Organelles

Pittis AA, Gabaldón T. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature. 2016 Mar 3;531(7592):101-4. doi: 10.1038/nature16941.

7 Branch length distributions

Pittis A. A. & Gabaldón, T. (2016). Response to "Late mitochondrial acquisition is pure artefact".

7.1 Prologue

Shortly after the publication of our article "Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry." (Pittis and Gabaldón, 2016) we received a harsh critical comment from William Martin, the main supporter of the mito-late model, and colleagues (Martin et al., 2016). Given the importance of the discussion for researchers in the field, we decided to publish our response. The analysis of branch length distributions from phylogenetic trees is indeed a biologically relevant source of information, as we explain in this chapter.

7.2 Abstract

In their paper Martin et al., 2016 criticize several methodological aspects of our previous work (Pittis and Gabaldón, 2016). None of the points raised affect the core of our conclusions -i.e. that differences in stem lengths relate to phylogenetic origin of LECA families so that they are shorter in bacterial, and particularly α-proteobacterial derived families- because the observed relationships i) are independent of the clustering performed in Figure 1 of Pittis and Gabaldón, 2016, and ii) their criticism focuses on one single comparison of a single dataset but the differences are present across several datasets and approaches, including the very same dataset from the authors mentioned in their letter (Ku et al., 2015), as we show below.

Furthermore, their interpretation of our stem length measurement and how they extrapolate to branches sub-tending eukaryotic clades is conceptually flawed, as we also demonstrate below. Thus none of their arguments compromise at any rate the main conclusions of our article. Finally, the new dataset and analyses brought about by this discussion further supports he validity of our results across methods and datasets. As the relative timing of the mitochondrial endosymbiosis is a central point to the question of eukaryogenesis, we want to discuss their points.

Gabaldón T. Response to Late mitochondrial origin is pure artifact. Treevolution : Biology through the evolutionary lens. 31 maig 2016. [Data consulta: 21 nov 2016]

7.3 Main

Contrary to what Martin et al., 2016 claim we do not assume a normal distribution of the global distribution of stem lengths. The claim that our statistical analyses are inappropriate is simply not true, we clearly explain all the methods used, and the tests performed to support observed differences are all nonparametric, without any assumption of normality. In Figure 6.1 we did use a probabilistic clustering method that fits a Gaussian mixture model, a mixture of normal distributions, assuming multimodality in the data. Martin et al., 2016 show that a unimodal log-normal distribution would better fit the data when the number of parameters is penalized. Does this demonstrates that the underlying distribution is not a composite of five gaussians? No, because when data are drawn from a five gaussian distributions with the obtained parameters, in 81% of the cases a log-normal distribution would be (wrongly) preferred using the BIC criterion (Figure 7.1). Also, the fact that any randomly sampled log-normal distribution could be fitted by a mixture model is by no means a surprise. In fact any distribution of data could be fitted by a finite number of mixture components, and this is precisely why these mixture models are commonly used as universal function approximators and as a tool to partition various kinds of data. Finally the definition of overfitting is not BIC inflation but the lack of predictive power. Thus other parameters have to be considered when assessing whether a model provides a reasonable representation of the data. The use of the EM algorithm is justified as a method for partitioning the data because i) we may expect composite of signals from a proteome (LECA) with at least two ancestral components (Archaeal host, and bacterial endosymbiont), and ii) prior studies have suggested that normalized branch lengths measurements as the ones used here to be approximately normal (Rasmussen and Kellis, 2007). The assumption of a unimodal distribution such as the one proposed by Martin et al., 2016 does not capture the expected mixture origins for a chimeric proteome and does not fit with the observation that differences in stem lengths relate to non-homogeneous phylogenetic origins. In any case our results are independent of this clustering exercise as the differences in stem lengths are apparent when simply grouping the LECA families according to their sister clades (Figure

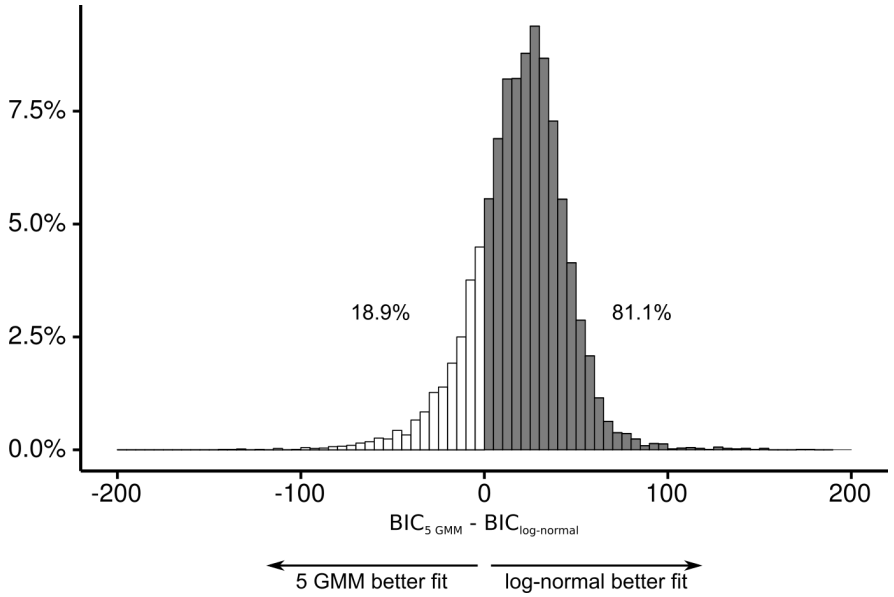


Figure 7.1: BIC difference of 5 GMM vs log-normal fit of randomly sampled 5 GMM with the *sl* estimated parameters. A log-normal distribution is a better fit according to BIC more than 80% of the times, when the data are randomly generated with 5 gaussians mixture (5 GMM) parameters.

6.2) and Supplementary Figure 6.4b of Pittis and Gabaldón, 2016), or when using other forms of clustering the data such as equal binning (results not shown).

Their purported extrapolation of our analyses to eukaryotic clades and their derived dates is totally flawed and misleading. First of all, we explicitly say that we do not assume constant rates (i.e. molecular clock), and our normalized branch length is a measurement that is proportional to time but multiplied by a ratio between the rate preceding and postdating LECA, so their timing exercise, providing date estimates, is completely ungrounded. Secondly, Martin et al., 2016 consider the normalized sl to yield arbitrary values, resulting in a log-normal distribution. This openly contradicts the observation that families of different prokaryotic origins show significant differences in sl and also rsl values. All our analyses robustly prove the opposite, there are differences and these differences reflect the relative divergence times. The cases of the cyanobacterial signal in Archaeplastida (Supplementary Figure 6.6) and of Lokiarchaeota

signal in LECA (Supplementary Figure 6.10) nicely indicate the validity of the measurement. Expecting some extreme ebl values to reflect radical adaptations and fast rates of some lineages, we used the median because of its robustness with respect to extreme outliers (see Methods). We also tried not accounting for fast evolving taxonomic groups in the calculations, without any change in our main results. All these observations are not explained by the interpretation of the data provided by Martin et al., 2016. Furthermore, Martin et al., 2016 show that the normalized branch lengths sub-tending each eukaryotic clade follow log-normal distributions, and conclude that this observation demonstrates that this is natural variation for branches meant to represent a single time interval (e.g. divergence of fungi from metazoans). By adopting this assumption they are surprisingly ignoring that eukaryotic families are also subject to differential gene loss and other processes, which would result in multiple underlying patterns of the sub-tending branches (i.e. the sub-tending branch of a fungal family, which was lost in metazoans does not derive from the divergence between fungi and metazoans, but from the deeper divergence of fungi and other unikonts). This becomes apparent when controlling for the relationship of the normalized branch lengths with the phylogenetic affiliation of the sister branch -a key step in our analyses which they ignore. Indeed applying to the eukaryotic clades an EM-based clustering and measuring enrichments in phylogenetic affiliations as we did in our previous analysis (Pittis and Gabaldón, 2016) reveals major underlying distributions related with the nature of the sister group (Figure 7.2). Thus, in this case also, the variation of sl values, interpreted by the authors as "vividly documenting abundant branch length variation", is clearly shown to naturally carry the signal of different divergence times. So yes, the sl values in eukaryotic groups do imply phases of early and late divergence times due to gene loss or other biological events, as they do in the case of LECA. Of note this is a new, independent demonstration that variation in stem lengths relate with underlying variation in phylogenetic distribution, and provides additional support to our approach.

Finally, (Martin et al., 2016) focus their criticism in only one of our comparisons and on only one of the datasets used. For that dataset, they

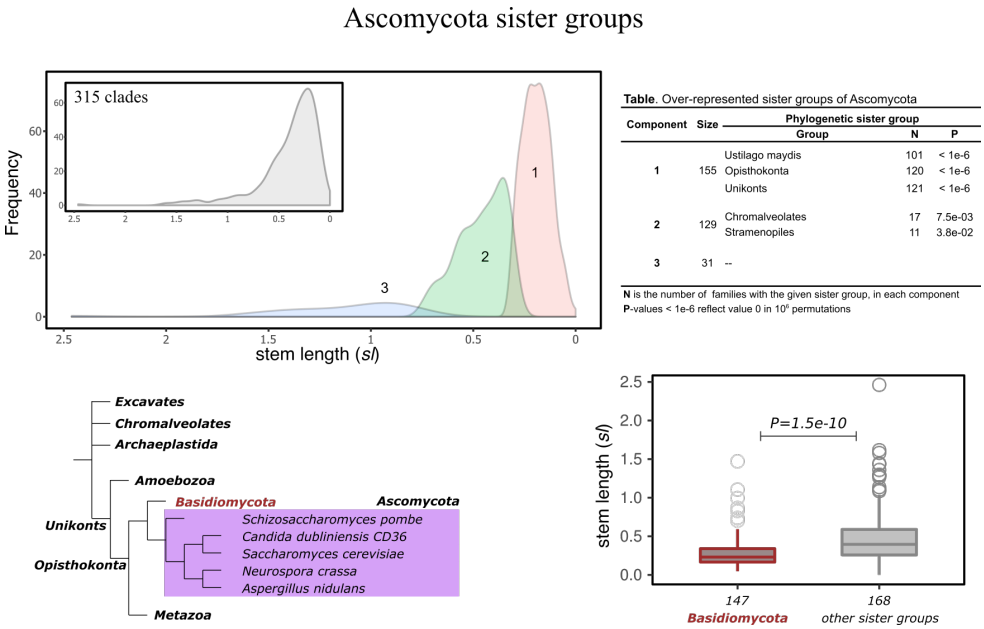


Figure 7.2: Ascomycota stem length analysis. Different phylogenetic sister groups show significant differences in stem lengths according to their divergence times from Ascomycota. Gene losses in the sister group lineage can explain the alternative tree topologies and differences in estimated stem lengths.

wrongly claim that we reused eukaryotic sequences in the different tree. This is false. Given the multidomain nature of eukaryotic protein sequences, the source of that dataset (Powell et al., 2013) may incorporate a given protein to more than one orthologous cluster. However we made sure we only used the orthologous sequence regions in a given analysis, thus never re-using a given eukaryotic sequence. Our analyses use standard filtering approaches but they claim that statistical significance for one of our comparisons (α-proteobacterial to other bacteria) is lost when applying additional ad hoc filtering on top of our previous filtering steps. We must note that even applying their filtering and using a permutation test as the one used in our paper, the α-proteobacterial sl values, remain significantly lower compared to other bacteria ($P = 10^{-2}$, accounting only for families with eukaryotic sequence lengths ≥ 100 and $P = 3.7 \times 10^{-2}$, accounting only for alignments with gaps $\leq 50\%$, 10^6 permutations). The loss of significance in some of the tests when artificially reducing the data is unsurprising. We are focusing on very ancient events and the signal we are measuring must be

necessarily weak, and the number of LECA families that can be traced back to specific ancestries is limited. Indeed the statistical significance using a Mann-Whitney U-test is often lost ($>$ 60%-70% of the times) when randomly reducing the data to sizes similar to the resulting sizes in their filtered dataset, which suggest that the mere effect of reducing the size, rather than the particular additional filtering used is having a major effect. This is why we made sure the signal was robust across different datasets, always using state of the art filtering approaches. Given the suggestion by Martin et al., 2016 that a recent phylogenetic analyses from them (which appeared after we had submitted the paper) represents a more careful dataset (Ku et al., 2015), we repeated our analyses using this dataset, which confirmed our results (650 eukaryotic clades, Archaeal vs Bacterial families, $P = 1.2 \times 10^{-41}$, two-tailed Mann-Whitney U-test and α-proteobacterial families' sl significantly smaller within Bacterial, $P = 4.7 \times 10^{-2}$, permutation test, 10^6 permutations). Again, this result lends further support to our findings.

Altogether, we show that the criticisms raised by Martin et al., 2016 do not compromise the main results and conclusions of our paper. Furthermore, we would like to stress that the new dataset and analyses brought about by this discussion lend additional support to our approach and conclusions.

7.4 Methods

The analyses were based on the same alignment and tree data used in our original publication Pittis and Gabaldón (2016) and the data provided in Ku et al., 2015. All statistical tests were performed as in Pittis and Gabaldón, 2016, using the same R and python code. For the new analyses, focusing on the various taxonomic levels within the LECA clades, the first encountered clade per level was accounted for the calculation of rsl and ebl values (analogously to the values of the LECA clades).

Part III

Discussion

8 Summarizing discussion

In this thesis I have presented analyses on different aspects of eukaryotic gene family evolution, mainly focused on the origin and evolution of the first eukaryotic cells. Using phylogenomics, I have approached problems related to co-evolution across protein families, evolutionary signal in deep phylogenetic relationships, and the long-standing question of the relevance of mitochondrial endosymbiosis in the origins of eukaryotic complexity. Every chapter of this thesis has its specific discussion section. Here, I will summarize the general implications of my work so far, and present my view on the future of comparative genomics and its role in the study of evolutionary history.

Nothing in biology makes sense except in the light of comparative genomics and vice versa

Theodosius Dobzhansky in his much-celebrated essay in 1973 "Nothing in Biology Makes Sense Except in the Light of Evolution" argued that unless through the prism of evolution, biology is meaningless. In the genomics era, where sequencing technology has truly revolutionized our means to address biological questions, the comparative study of genomes within a phylogenetic context has proven itself invaluable in giving context to genomic complextity.

Comparative genomics is the direct comparison of the (complete) genomic component of one organism to another. It is a powerful methodology not only for gaining insights into the evolution of organisms but also to understand genetic, metabolic and physiological pathways, through the analysis of the patterns of genes and non-coding regions across different species. Cur-

rently, there is almost no biological problem in which comparative genomic analysis is not directly or indirectly implicated. Co-evolution patterns in various genomic scales (nucleotide, amino acid, gene) are being explored to predict functional interactions (de Juan et al., 2013). Distributions of genetic alterations are widely used in the population or species levels to associate genome features to phenotypes, with applications from disease (Alföldi and Lindblad-Toh, 2013) to the study of any possible adaptive trait. Conservation across genomes in different species is crucial for gene finding and identification of regulatory regions (Kellis et al., 2004). The profiles of protein families in genomes derived from metagenomic samplings, can describe the biology of otherwise unseen organisms expanding significantly our view of biodiversity and address long-standing evolutionary questions (Spang et al., 2015). In the same sense that evolutionary theory gives a meaning to biology, comparative genomics give a meaning to genome sequences. The genome can only be understood within its phylogenetic context.

As of July 2016, the numbers of fully sequenced genomes present only in UniProt are 51,255 Bacteria, 1,137 Eukaryotes, 478 Archaea and 3,446 Viruses. Ironically, such extraordinary - already - availability of public data, would suffice to keep biological scientists busy for the next many years, and it will keep growing exponentially. Inevitably, a big part of this information remains largely unexplored. Definitely the computational power needed to deal with so large and complex data, the storage, the curation and cleaning of the data, all pose serious challenges, but above all the analysis and extraction of meaningful information has been proven a highly laborious task. The stockpile seems to be growing faster than our capacity to make use of it. Over the last years expertise from different disciplines has been combined, leading to significant advancements in comparative genomic methodologies, but biological input is always needed to test their validity, improve them, give a direction. If not, it can easily happen that statistically significant results reflect good fit of biased data, without any biological relevance. I believe that bridging the gap between computational analysis and biological knowledge, is going to be the greatest challenge in the field in the following years, and it will certainly need integration of knowledge coming from the widest range of scientific disciplines, from ecology and genetics, to mathematics and

computer science. Biology will be certainly dominated by genomics in the near future, to address the most difficult questions there has to be more and more biology in genomics as well.

Phylogenomics and the origin of eukaryotic cells

Since Carl Woese and George Fox presented the first phylogenetic Tree of Life (ToL) based on ribosomal RNA sequences in 1977, molecular evolution techniques have been fundamental for testing hypothesis for eukaryotic origins. Molecular phylogenies provided important insights in difficult problems on the evolutionary relationships between species, but also many times revealed an unexpected degree of biological and/or methodological complexity.

Regarding the topic of the Origin of Eukaryotes, research has been driven to a large extent by phylogenetic analysis. While a bacterial connection had been suggested previously by the work of Michael Gray and his co-workers (Spencer et al., 1984), the mitochondria - α-proteobacteria link was first demonstrated in 1985 by Carl Woese and colleagues through the comparison of 16S (18S in mitochondria) rRNA sequences from seven prokaryotic species, and another four eukaryotic mitochondria (Yang et al., 1985). Hints from the study of the evolution of key eukaryotic gene families suggested a root of the traditional three-domain ToL in the bacterial branch, pointing to the sister relationship between Archaea and Eukaryotes (Gogarten et al., 1989; Iwabe et al., 1989). As sequencing techniques advanced, in the dawn of the genomics era, and complete species' genomes became available, phylogenomics were born. The first large scale phylogenomic analyses of complete gene sets revealed that the eukaryotic nuclear genome is a mosaic of genes of bacterial, archaeal, or lineage specific eukaryotic origin (Ribeiro and Golding, 1998), hence Eukaryotes are genetic chimeras. Of note this chimerism is not organized in a random, disordered manner, but instead show clear patterns, with an archaeal derived component involved in information processing, and a bacterial one dominating cellular metabolism (Rivera et al., 1998).

It is interesting to note that the wealth of data that followed these early

breakthroughs, did not necessarily clear the way to the big questions in the field. In several cases the results from the analyses that followed, with more data, more sophisticated methods, more powerful machines, were inconclusive and were rather unveiling a complex picture, difficult to analyze and interpret. Half a century since the formulation of the modern endosymbiotic theory for the origin of the eukaryotic cell, many of the crucial details of eukaryogenesis are still strongly debated, which also shows why this is one of the most fascinating topics in evolutionary biology. The quest for the specific α-proteobacterial sister group of mitochondria, often gave conflicting results, groups among those that have been proposed being the Rickettsiales, the oceanic SAR11 and even Rhodospirillales (Wang and Wu, 2015; Thrash et al., 2011; Carvalho et al., 2015). Following the initial proposal of the Eocyte hypothesis by James Lake and co-workers in 1984 (Lake et al., 1984), and after significant efforts over the years to resolve the 3D vs 2D controversy (see also section 1.3) kept giving conflicting results, Gribaldo et al., 2010 asked "Are we at a phylogenomic impasse?". Rooting the ToL has not been less of a challenge, the proposed alternative rootings would make a long list, each time altering its interpretation (Lake et al., 2009; Gouy et al., 2015). Finally, explaining the intricate genomic mosaicism of Eukaryotes (see also section 1.5), has been at the core of all modern eukaryogenesis hypotheses, and at the core of the controversy (Ku et al., 2015; Rochette et al., 2014; Lester et al., 2006; Pittis and Gabaldón, 2016).

There is no doubt that without the availability of complete genomes most of the deep questions in the field could have never been even asked. The richness and accessibility of sequences in public databases makes possible testing the most complex hypotheses, and still it has only been possible to examine a tiny fraction of the information available. The exploration of microbial diversity is revolutionizing our view on the origin of Eukaryotes, and new data keep providing exciting new opportunities to ask new questions and test older ideas. However, the last years it is becoming clear that more abundant data will not provide answers per se, sampling efforts have start to target the exploration of the unknown diversity, the one residing in unexplored or extreme environments, and the efforts appear worthwhile already. Put in other words, in the research on eukaryotic origins some times

more proved less and more sequence abundance may simply not be enough (Philippe et al., 2011).

When I first got interested in the problem of the origin of Eukaryotes, in a discussion with my PhD supervisor, Toni Gabaldón, he mentioned that years ago a distinguished evolutionary biologist had told him "Given the lack of data, research on the origin of Eukaryotes is mostly about faith". During my PhD, I understood that many times strong beliefs (or "faith") were called forth in the absence of conclusive data. The microbiologist Roger Stanier, on his view about the origin of Eukaryotes in 1970 (Stanier, 1970), warned us rather vividly:

> *It might have happened thus; but we shall surely never know with certainty. Evolutionary speculation constitutes a kind of metascience, which has the same intellectual fascination for some biologists that metaphysical speculation possessed for some medieval scholastics. It can be considered a relatively harmless habit, like eating peanuts, unless it assumes the form of an obsession; then it becomes a vice.*

Scientific speculation is important when the data are simply not there, yet. It provides a working scheme for hypothesis testing when sufficient data become available or when new ideas arise. The study of the evolutionary history of organisms often requires making assumptions that we cannot be certain about, but at least there are good reasons to believe they are true. Especially, in the study of events that took place maybe more than 2 billion years ago, we can rely almost exclusively on indirect evidence and interpret the data based on assumptions that may be or not accurate. Speculation of course is practiced in the study of eukaryogenesis eminently. Hypotheses should be based on robust assumptions, inferred from scientific facts, and tested constantly on real data. Inevitably in practice, personal criteria and beliefs often have a big influence on most speculative hypotheses.

The work presented in this thesis on the phylogenetic signal inferred to LECA points to a complex host that acquired mitochondria. An implication of that is that the α-proteobacterial endosymbiont was incorporated late during the stem phase leading to LECA, after eukaryotic complexity was already along the way. This is not a new idea. On contrary, this was the

prevailing view during the first years after the revival of the endosymbiotic theory for the origin of mitochondria, based on basic principles of cell biology and microbiological observations. Until the discovery that true, primarily amitochondrial Eukaryotes have never been found (see also section 1.4). The loss of mitochondria appears to have been a secondary adaptation to specific environments and lifestyles. As Anthony Poole and David Penny mention in their article "Engulfed by speculation" (Poole and Penny, 2007), the fact that amitochondrial Eukaryotes have not been found among modern eukaryotic species does not imply that those never existed. I anticipate in the future the focus to get back to testable hypotheses as new data arrive, free as possible of preconceptions that withhold further advancements in the field.

Our work uncovers a new angle to look at complex evolutionary problems using phylogenomics. The distribution of branch lengths as discussed in **Chapters 6** and **7** can prove an important source of information, largely unexplored until now, in several problems where the phylogenetic signal has been conflicting and too complicated to make sense of. However, I believe it is important to remember that the outcomes of all probabilistic methods, as most of those used in phylogenomics and genomics generally, are statistical inferences on the data they set out to explain. In modern evolutionary analysis sometimes it seems like "molecular and genomic data, which were originally harnessed to answer questions about cell evolution, now so dominate our thinking that they largely define the question" (Keeling, 2014). If we are aware of that, this can also be interesting and prolific from a different perspective. But if the objective is the study of the evolution of the cell, the biological interpretation of the data should always come first.

Conclusions

- Retargeting of proteins among the various subcellular compartments has been extensive during the evolution of eukaryotes. The retargeting patterns reveal an evolutionary link between peroxisomes and mitochondria.

- Mitochondrial calcium signaling is an ancestral feature in Eukaryotes. The current calcium signalling machinery has no alpha-proteobacterial origin and constitutes a eukaryotic innovation in the lineage leading to the common ancestor of all extant eukaryotic species. The molecular machinery involved in the uptake of calcium in mitochondria is tightly linked and co-evolving.

- We proposed a new method for the analysis of the phylogenetic origin of genes based on fast similarity searches.

- Using similarity-based and phylogenomic approaches we confirmed previous observations that the reconstructed ancestral Eukaryotic proteome is a composite of genes originating from different prokaryotic sources and genes that are specific to Eukaryotes. Their ancestry correlates with different biological functions.

- The distribution of branch lengths from phylogenetic trees can provide interpretable information for the evolution of genes and lineages. Particularly those cases where certain lineages have acquired different genes in different times, the analysis of branch lengths can help to disentangle the order of ancient events.

- Mitochondrial endosymbiosis occurred relatively late during the process of eukaryogenesis. The phylogenetic signal carried by modern eukaryotic genomes, indicates that mitochondria were acquired by a host

cell of chimeric archaeal and bacterial origins, which already possessed a certain degree of genomic complexity.

- The more ancient origin of archaeal and bacterial LECA proteins other than those from α-protobacteria, suggests that the formation of other key features of eukaryotes was already underway prior to the acquisition of mitochondria. Part of the pre-existing proteome was later retargeted to the newly formed organelle.

Appendix: List of publications

1. **Pittis A. A.** & Gabaldón, T. (2016). Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature*, 531(7592), 101-104.

2. **Pittis A. A.** & Gabaldón, T. Response to "Late mitochondrial acquisition is pure artefact". *submitted*.

3. **Pittis A. A.** & Gabaldón, T. Assessing the origins of the genetic repertoire of the Last Eukaryotic Common Ancestor through the analysis of similarity distributions. *in preparation*.

4. **Pittis A. A.**, Perocchi F. & Gabaldón, T. Phylogenomics of mitochondrial calcium homeostasis. *in preparation*.

5. Gabaldón, T. & **Pittis, A. A.** (2015). Origin and evolution of metabolic sub-cellular compartmentalization in eukaryotes. *Biochimie*, 119, 262-268.

6. Krypotou, E.,Evangelidis, T., Bobonis, J., **Pittis, A. A.**, Gabaldón, T., Scazzocchio, C., Mikros, E. & Diallinas, G. Origin, diversification and substrate specificity in the family of NCS1/FUR transporters. *Mol. Microbiol.* 96, 927–50 (2015).

Appendix: Tables

Table A1: The 243 eukaryotic species used in chapter 4 and the sources of their proteomes.

Taxid	Organism name	Source
10020	*Dipodomys ordii*	ensembl
10090	*Mus musculus*	ensembl
10116	*Rattus norvegicus*	ensembl
10141	*Cavia porcellus*	ensembl
104341	*Postia placenta*	JGI
104355	*Gloeophyllum trabeum*	ensembl
109760	*Spizellomyces punctatus*	broad institute
109871	*Batrachochytrium dendrobatidis*	broad institute
114155	*Dichomitus squalens*	ensembl
114524	*Saccharomyces kudriavzevii*	Duke
114525	*Saccharomyces mikatae*	Duke
117187	*Gibberella moniliformis*	broad institute
121225	*Pediculus humanus*	VECTORBASE-VECTORBASE bis
132908	*Pteropus vampyrus*	ensembl
13563	*Heterobasidion annosum*	JGI
13616	*Monodelphis domestica*	ensembl
140110	*Nectria haematococca*	JGI
148960	*Wallemia sebi*	ensembl
153609	*Moniliophthora perniciosa*	ensembl
15368	*Brachypodium distachyon*	Phytozome-phylomeDB
164328	*Phytophthora ramorum*	JGI
180454	*Anopheles gambiae str. PEST*	integr8
184922	*Giardia lamblia ATCC 50803*	integr8
186039	*Fragilariopsis cylindrus*	JGI
192875	*Capsaspora owczarzaki*	Broad-broad ensembl
202698	*Punctularia strigosozonata*	ensembl
203908	*Melampsora larici-populina*	JGI
208348	*Puccinia triticina*	ensembl
208960	*Fomitiporia mediterranea*	ensembl
214684	*Cryptococcus neoformans var. neoformans JEC21*	integr8
227321	*Aspergillus nidulans FGSC A4*	integr8
237561	*Candida albicans SC5314*	CGD-CGD
237631	*Ustilago maydis 521*	integr8
242159	*Ostreococcus 'lucimarinus'*	integr8-phylomeDB
242507	*Magnaporthe oryzae 70-15*	integr8
246410	*Coccidioides immitis RS*	broad ensembl
27288	*Naumovozyma castellii*	YGOB-YGOB
27291	*Saccharomyces paradoxus*	Duke
27335	*Verticillium albo-atrum*	broad ensembl
278021	*Antonospora locustae*	unknown-VECTORBASE
28377	*Anolis carolinensis*	ensembl
284590	*Kluyveromyces lactis NRRL Y-1140*	integr8
284591	*Yarrowia lipolytica CLIB122*	integr8
284592	*Debaryomyces hansenii CBS767*	integr8
284593	*Candida glabrata CBS 138*	Genolevures-NIH
284811	*Ashbya gossypii ATCC 10895*	YGOB
284812	*Schizosaccharomyces pombe 972h-*	broad ensembl
284813	*Encephalitozoon cuniculi GB-M1*	integr8

28583	*Allomyces macrogynus*	ensembl
29001	*Alternaria brassicicola*	JGI-broad ensembl nogene
294381	*"Entamoeba histolytica HM-1*	IMSS:KEGG-KEGG
294746	*Meyerozyma guilliermondii ATCC 6260*	integr8
296543	*Thalassiosira pseudonana CCMP1335*	JGI
29760	*Vitis vinifera*	Genoscope-phylomeDB
29875	*Hypocrea virens*	JGI
29879	*Neurospora discreta*	JGI
29883	*Laccaria bicolor*	JGI
29898	*Rhodotorula graminis*	ensembl
30538	*Vicugna pacos*	ensembl
3055	*Chlamydomonas reinhardtii*	Phytozome-phylomeDB
30608	*Microcebus murinus*	ensembl
30611	*Otolemur garnettii*	ensembl
306901	*Chaetomium globosum CBS 148.51*	integr8
31033	*Takifugu rubripes*	ensembl
31271	*Plasmodium chabaudi chabaudi*	integr8
31870	*Glomerella graminicola*	ensembl
321614	*Phaeosphaeria nodorum SN15*	integr8
3218	*Physcomitrella patens*	Phytozome-phylomeDB
322104	*Scheffersomyces stipitis CBS 6054*	integr8
325569	*NOT FOUND*	integr8
331117	*Neosartorya fischeri NRRL 181*	integr8
33188	*Uncinocarpus reesii*	broad ensembl
332648	*Botryotinia fuckeliana B05.10*	integr8
333668	*Theileria parva strain Muguga*	integr8
339724	*Ajellomyces capsulatus NAm1*	broad ensembl
341663	*Aspergillus terreus NIH2624*	integr8
34387	*Trichophyton tonsurans*	ensembl
344612	*Aspergillus clavatus NRRL 1*	integr8
347515	*Leishmania major strain Friedlin*	integr8
352472	*Dictyostelium discoideum AX4*	integr8
352914	*Plasmodium yoelii yoelii 17XNL*	integr8
353151	*Cryptosporidium hominis TU502*	integr8
353152	*Cryptosporidium parvum Iowa II*	integr8
35720	*Thielavia terrestris*	ensembl
36080	*Mucor circinelloides*	JGI
36329	*Plasmodium falciparum 3D7*	integr8
367110	*Neurospora crassa OR74A*	integr8
36911	*Clavispora lusitaniae*	broad ensembl
3702	*Arabidopsis thaliana*	integr8
370354	*Entamoeba dispar SAW760*	KEGG
37347	*Tupaia belangeri*	ensembl
379508	*Lodderomyces elongisporus NRRL YB-4239*	integr8
381046	*Lachancea thermotolerans*	Genolevures-NIH
3847	*Glycine max*	Phytozome-phylomeDB
39416	*Tuber melanosporum*	ensembl
39947	*Oryza sativa Japonica Group*	integr8
40127	*Neurospora tetrasperma*	JGI
402676	*Schizosaccharomyces japonicus yFS275*	broad ensembl
40483	*Fomitopsis pinicola*	ensembl
40492	*Stereum hirsutum*	ensembl
40563	*Sporobolomyces roseus*	JGI
40993	*Aspergillus carbonarius*	ensembl
412133	*Trichomonas vaginalis G3*	KEGG
42254	*Sorex araneus*	ensembl
425011	*Aspergillus niger CBS 513.88*	integr8
42742	*Ceriporiopsis subvermispora*	ensembl
43179	*Ictidomys tridecemlineatus*	ensembl
436017	*Ostreococcus lucimarinus CCE9901*	integr8
436907	*Vanderwaltozyma polyspora DSM 70294*	YGOB
44056	*Aureococcus anophagefferens*	JGI
45151	*Pyrenophora tritici-repentis*	broad ensembl
451804	*Aspergillus fumigatus A1163*	integr8
45351	*Nematostella vectensis*	integr8
454284	*Coccidioides posadasii RMSCC 3488*	broad ensembl
4558	*Sorghum bicolor*	JGI-phylomeDB

4577	*Zea mays*	www.maizesequence.org-phylomeDB
46245	*Drosophila pseudoobscura pseudoobscura*	flybase
4784	*Phytophthora capsici*	JGI
4787	*Phytophthora infestans*	Broad-broad ensembl
481877	*Enterocytozoon bieneusi H348*	integr8
483514	*Schizosaccharomyces octosporus yFS286*	broad ensembl
4837	*Phycomyces blakesleeanus*	JGI
489714	*Microsporum gypseum*	broad ensembl
4914	*Lachancea waltii*	Duke
4922	*Komagataella pastoris*	NCBI-NCBI
4931	*Saccharomyces bayanus*	YGOB
4934	*Lachancea kluyveri*	Genolevures-NIH
4956	*Zygosaccharomyces rouxii*	Genolevures-NIH
498257	*Verticillium dahliae VdLs.17*	broad ensembl
500485	*Penicillium chrysogenum Wisconsin 54-1255*	integr8
5016	*Cochliobolus heterostrophus*	JGI
502780	*Paracoccidioides brasiliensis Pb18*	broad ensembl
5059	*Aspergillus flavus*	broad ensembl
5062	*Aspergillus oryzae*	broad ensembl-broad ensembl nogene
507601	*Toxoplasma gondii GT1*	NCBI-phylomeDB
5116	*Cryphonectria parasitica*	JGI
51453	*Hypocrea jecorina*	JGI
5145	*Podospora anserina*	Podospora anserina genome database-VECTORBASE
51511	*Ciona savignyi*	ensembl
5217	*Tremella mesenterica*	JGI
5297	*Puccinia graminis*	broad ensembl
5306	*Phanerochaete chrysosporium*	JGI
5322	*Pleurotus ostreatus*	JGI
5325	*Trametes versicolor*	ensembl
5334	*Schizophyllum commune*	JGI
5341	*Agaricus bisporus*	ensembl
5346	*Coprinopsis cinerea*	broad ensembl
54734	*NOT FOUND*	JGI
5480	*Candida parapsilosis*	broad ensembl
5482	*Candida tropicalis*	broad ensembl
5507	*Fusarium oxysporum*	unknown-broad ensembl
5518	*Gibberella zeae*	broad ensembl
554155	*Arthroderma otae CBS 113480*	broad ensembl
5551	*Trichophyton rubrum*	ensembl
556484	*Phaeodactylum tricornutum CCAP 1055/1*	JGI
559292	*Saccharomyces cerevisiae S288c*	SGD-CGD
559297	*Ajellomyces dermatitidis ER-3*	broad ensembl
5660	*Leishmania braziliensis*	integr8
5671	*Leishmania infantum*	integr8
5691	*Trypanosoma brucei*	KEGG
5693	*Trypanosoma cruzi*	integr8
573728	*NOT FOUND*	JGI
573826	*Candida dubliniensis CD36*	integr8
5762	*Naegleria gruberi*	JGI
578460	*Nosema ceranae BRL01*	unknown-VECTORBASE
5786	*Dictyostelium purpureum*	JGI
5823	*Plasmodium berghei ANKA*	integr8
5851	*Plasmodium knowlesi strain H*	integr8
5855	*Plasmodium vivax*	integr8
5865	*Babesia bovis*	integr8
5874	*Theileria annulata*	integr8
58839	*Encephalitozoon intestinalis*	ensembl
5888	*Paramecium tetraurelia*	integr8-phylomeDB
5911	*Tetrahymena thermophila*	KEGG
59463	*Myotis lucifugus*	ensembl
59689	*Arabidopsis lyrata*	Phytozome-phylomeDB
59729	*Taeniopygia guttata*	ensembl
6238	*Caenorhabditis briggsae*	integr8
6239	*Caenorhabditis elegans*	ensembl
63418	*Trichophyton equinum*	broad ensembl
63577	*Trichoderma atroviride*	JGI
64363	*Mycosphaerella pini*	ensembl

64495	*Rhizopus oryzae*	broad ensembl
6669	*Daphnia pulex*	JGI
67593	*Phytophthora sojae*	JGI
69293	*Gasterosteus aculeatus*	ensembl
7029	*Acyrthosiphon pisum*	NIH-Aphid
70448	*Ostreococcus tauri*	integr8
7070	*Tribolium castaneum*	NCBI-phylomeDB
7091	*Bombyx mori*	SILKDB-SILKDB
7159	*Aedes aegypti*	integr8
7175	*Culex pipiens*	VECTORBASE-VECTORBASE
7217	*Drosophila ananassae*	flybase
7220	*Drosophila erecta*	flybase
7222	*Drosophila grimshawi*	flybase
7227	*Drosophila melanogaster*	flybase
7230	*Drosophila mojavensis*	flybase
7234	*Drosophila persimilis*	flybase
7238	*Drosophila sechellia*	flybase
7240	*Drosophila simulans*	flybase
7244	*Drosophila virilis*	flybase
7245	*Drosophila yakuba*	flybase
7260	*Drosophila willistoni*	flybase
73824	*Populus balsamifera*	JGI-phylomeDB
7425	*Nasonia vitripennis*	NCBI-nasonia
76773	*Malassezia globosa*	Hyphal tip-hiphal
7719	*Ciona intestinalis*	ensembl
7955	*Danio rerio*	ensembl
80637	*Coniophora puteana*	ensembl
80884	*Colletotrichum higginsianum*	ensembl
8090	*Oryzias latipes*	ensembl
81056	*Wolfiporia cocos*	ensembl
81824	*Monosiga brevicollis*	integr8
83344	*Mycosphaerella fijiensis*	JGI
8364	*Xenopus (Silurana) tropicalis*	ensembl
85929	*Mycosphaerella populorum*	ensembl
866546	*Schizosaccharomyces cryophilus*	ensembl
88036	*Selaginella moellendorffii*	Phytozome-phylomeDB
9031	*Gallus gallus*	ensembl
9258	*Ornithorhynchus anatinus*	ensembl
9315	*Macropus eugenii*	ensembl
9358	*Choloepus hoffmanni*	ensembl
9361	*Dasypus novemcinctus*	ensembl
9365	*Erinaceus europaeus*	ensembl
9371	*Echinops telfairi*	ensembl
9478	*Tarsius syrichta*	ensembl
9483	*Callithrix jacchus*	ensembl
9544	*Macaca mulatta*	ensembl
9593	*Gorilla gorilla*	ensembl
9598	*Pan troglodytes*	ensembl
9600	*Pongo pygmaeus*	ensembl
9606	*Homo sapiens*	ensembl
9615	*Canis lupus familiaris*	ensembl
9685	*Felis catus*	ensembl
9739	*Tursiops truncatus*	ensembl
9785	*Loxodonta africana*	ensembl
9796	*Equus caballus*	ensembl
9813	*Procavia capensis*	ensembl
9823	*Sus scrofa*	ensembl
9913	*Bos taurus*	ensembl
9978	*Ochotona princeps*	ensembl
9986	*Oryctolagus cuniculus*	ensembl
99883	*Tetraodon nigroviridis*	ensembl

References

Alföldi, J. and Lindblad-Toh, K. (2013). Comparative genomics as a tool to understand evolution and disease. *Genome research*, 23(7):1063–1068.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Andersen, J. S. and Mann, M. (2006). Organellar proteomics: turning inventories into insights. *EMBO reports*, 7(9):874–879.

Archibald, J. M. (2009). The puzzle of plastid evolution. *Current biology : CB*, 19(2):R81–8.

Archibald, J. M. (2015). Endosymbiosis and eukaryotic cell evolution. *Current Biology*, 25(19):R911–R921.

Baldauf, S. L., Palmer, J. D., and Doolittle, W. F. (1996). The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proceedings of the National Academy of Sciences*, 93(15):7749–7754.

Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V., and Robinson-Rechavi, M. (2008). Bgee: integrating and comparing heterogeneous transcriptome data among species. In *Data Integration in the Life Sciences*, pages 124–131. Springer.

Baughman, J. M., Perocchi, F., Girgis, H. S., Plovanich, M., Belcher-Timme, C. A., Sancak, Y., Bao, X. R., Strittmatter, L., Goldberger, O., Bogorad, R. L., et al. (2011). Integrative genomics identifies mcu as an essential component of the mitochondrial calcium uniporter. *Nature*, 476(7360):341–345.

Bazylinski, D. A. and Frankel, R. B. (2004). Magnetosome formation in prokaryotes. *Nature Reviews Microbiology*, 2(3):217–230.

Bell, E. A., Boehnke, P., Harrison, T. M., and Mao, W. L. (2015). Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proceedings of the National Academy of Sciences*, 112(47):201517557.

Bengtson, S., Belivanova, V., Rasmussen, B., and Whitehouse, M. (2009). The controversial "cambrian" fossils of the vindhyan are real but more than a billion years older. *Proceedings of the National Academy of Sciences*, 106(19):7729–7734.

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, 21(2):163–193.

Berridge, M. J., Bootman, M. D., and Roderick, H. L. (2003). Calcium signalling: dynamics, homeostasis and remodelling. *Nature reviews Molecular cell biology*, 4(7):517–529.

Bick, A. G., Calvo, S. E., and Mootha, V. K. (2012). Evolutionary diversity of the mitochondrial calcium uniporter. *Science*, 336(6083):886–886.

Birdsey, G. M., Lewin, J., Cunningham, A. A., Bruford, M. W., and Danpure, C. J. (2004). Differential enzyme targeting as an evolutionary adaptation to herbivory in carnivora. *Molecular biology and evolution*, 21(4):632–646.

Bodył, A., Mackiewicz, P., and Stiller, J. W. (2009). Early steps in plastid evolution: current ideas and controversies. *Bioessays*, 31(11):1219–1232.

Bolte, K., Gruenheit, N., Felsner, G., Sommer, M. S., Maier, U.-G., and Hempel, F. (2011). Making new out of old: Recycling and modification of an ancient protein translocation system during eukaryotic evolution. *Bioessays*, 33(5):368–376.

Booth, A. and Doolittle, W. F. (2015). Eukaryogenesis, how special really? *Proceedings of the National Academy of Sciences*, 112(33):10278–10285.

Brocks, J. J., Logan, G. A., Buick, R., and Summons, R. E. (1999). Archean molecular fossils and the early rise of eukaryotes. *Science*, 285(5430):1033–1036.

Brown, J. R. and Doolittle, W. F. (1995). Root of the universal tree of life based on ancient aminoacyl-trna synthetase gene duplications. *Proceedings of the National Academy of Sciences*, 92(7):2441–2445.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973.

Carafoli, E. and Lehninger, A. L. (1971). A survey of the interaction of calcium ions with mitochondria from different tissues and species. *Biochemical Journal*, 122(5):681–690.

Carvalho, D. S., Andrade, R. F., Pinho, S. T., Góes-Neto, A., Lobão, T. C., Bomfim, G. C., and El-Hani, C. N. (2015). What are the evolutionary origins of mitochondria? a complex network approach. *PloS one*, 10(9):e0134988.

Cavalier-Smith, T. (1983). A 6-kingdom classification and a unified phylogeny. *Endocytobiology II*, pages 1027–1034.

Cavalier-Smith, T. (1988). Origin of the cell nucleus. *BioEssays*, 9(2-3):72–78.

Cavalier-Smith, T. (1989). Molecular phylogeny. archaebacteria and archezoa. *Nature*, 339(6220):l00.

Cavalier-Smith, T. (2000). Membrane heredity and early chloroplast evolution. *Trends in plant science*, 5(4):174–182.

Cavalier-Smith, T. (2010). Origin of the cell nucleus, mitosis and sex: roles of intracellular coevolution. *Biology direct*, 5:7.

Cavalier-Smith, T. and Lee, J. J. (1985). Protozoa as hosts for endosymbioses and the conversion of symbionts into organelles. *The Journal of protozoology*, 32(3):376–379.

Chyba, C. and Sagan, C. (1992). Endogenous production, exogenous delivery and impact-shock synthesis of organic molecules: an inventory for the origins of life. *Nature*.

Clapham, D. E. (2007). Calcium signaling. *Cell*, 131(6):1047–1058.

Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R., and Embley, T. M. (2008). The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences*, 105(51):20356–20361.

Csordás, G., Thomas, A. P., and Hajnóczky, G. (1999). Quasi-synaptic calcium signal transmission between endoplasmic reticulum and mitochondria. *The EMBO journal*, 18(1):96–108.

Dalrymple, G. B. (2001). The age of the earth in the twentieth century: a problem (mostly) solved. *Geological Society, London, Special Publications*, 190(1):205–221.

Danpure, C. J. (1997a). The molecular basis of alanine: glyoxylate aminotransferase mistargeting: the most common single cause of primary hyperoxaluria type 1. *Journal of nephrology*, 11:8–12.

Danpure, C. J. (1997b). Variable peroxisomal and mitochondrial targeting of alanine: glyoxylate aminotransferase in mammalian evolution and disease. *Bioessays*, 19(4):317–326.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). Prottest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–1165.

de Alda, J. A. O., Esteban, R., Diago, M. L., and Houmard, J. (2014). The plastid ancestor originated among one of the major cyanobacterial lineages. *Nature communications*, 5.

de Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261.

Deluca, H. and Engstrom, G. (1961). Calcium uptake by rat kidney mitochondria. *Proceedings of the National Academy of Sciences*, 47(11):1744–1750.

Derelle, R. and Lang, B. F. (2012). Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Molecular biology and evolution*, 29(4):1277–89.

Derelle, R., Torruella, G., Klimeš, V., Brinkmann, H., Kim, E., Vlček, Č., Lang, B. F., and Eliáš, M. (2015). Bacterial proteins pinpoint a single eukaryotic root. *Proceedings of the National Academy of Sciences*, 112(7):E693–E699.

Do, C. B. and Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897–9.

Dominguez, D. C. (2004). Calcium signalling in bacteria. *Molecular Microbiology*, 54(2):291–297.

Doolittle, W. F. (1998). You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics*, 14(8):307–311.

Duhita, N., Satoshi, S., Kazuo, H., Daisuke, M., Takao, S., et al. (2010). The origin of peroxisomes: The possibility of an actinobacterial symbiosis. *Gene*, 450(1):18–24.

Duve, C. (1969). Evolution of the peroxisome. *Annals of the New York Academy of Sciences*, 168(2):369–381.

Eddy, S. R. (2011). Accelerated profile hmm searches. *PLoS Comput Biol*, 7(10):e1002195.

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797.

Emanuelsson, O., Nielsen, H., Brunak, S., and Von Heijne, G. (2000). Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *Journal of molecular biology*, 300(4):1005–1016.

Embley, M., Van Der Giezen, M., Horner, D. S., Dyal, P. L., and Foster, P. (2003). Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 358(1429):191–203.

Embley, T. M. and Martin, W. F. (2006). Eukaryotic evolution, changes and challenges. *Nature*, 440(7084):623–30.

Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., Henze, K., Kretschmann, E., Richly, E., Leister, D., et al. (2004). A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular Biology and Evolution*, 21(9):1643–1660.

Esser, C., Martin, W., and Dagan, T. (2007). The origin of mitochondria in light of a fluid prokaryotic chromosome model. *Biology Letters*, 3(2):180–184.

Forner, F., Foster, L. J., Campanaro, S., Valle, G., and Mann, M. (2006). Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Molecular & Cellular Proteomics*, 5(4):608–619.

Fraley, C., Raftery, A. E., and Scrucca, L. (2012). Normal mixture modeling for model-based clustering, classification, and density estimation. *Department of Statistics, University of Washington, Available online at http:// cran. r-project. org/ web/ packages/ mclust/ index. html. Accessed September*, 23:2012.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science*, 296(5568):750–752.

Gabaldón, T. (2010). Peroxisome diversity and evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1541):765–73.

Gabaldón, T. (2014a). Evolutionary considerations on the origin of peroxisomes from the endoplasmic reticulum, and their relationships with mitochondria. *Cellular and Molecular Life Sciences*, 71(13):2379.

Gabaldón, T. (2014b). A metabolic scenario for the evolutionary origin of peroxisomes from the endomembranous system. *Cellular and Molecular Life Sciences*, 71(13):2373.

Gabaldón, T. and Capella-Gutiérrez, S. (2010). Lack of phylogenetic support for a supposed actinobacterial origin of peroxisomes. *Gene*, 465(1):61–65.

Gabaldón, T. and Huynen, M. A. (2004). Shaping the mitochondrial proteome. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1659(2):212–220.

Gabaldón, T. and Huynen, M. A. (2007). From endosymbiont to host-controlled organelle: the hijacking of mitochondrial protein synthesis and metabolism. *PLoS Comput Biol*, 3(11):e219.

Gabaldón, T. and Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14(5):360–366.

Gabaldón, T. and Pittis, A. A. (2015). Origin and evolution of metabolic sub-cellular compartmentalization in eukaryotes. *Biochimie*, 119:262–268.

Gabaldón, T., Snel, B., Van Zimmeren, F., Hemrika, W., Tabak, H., and Huynen, M. A. (2006). Origin and evolution of the peroxisomal proteome. *Biology Direct*, 1(1):8.

Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, R. J., Date, T., and Oshima, T. (1989). Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 86(17):6661–5.

Gouy, R., Baurain, D., and Philippe, H. (2015). Rooting the tree of life: the phylogenetic jury is still out. *Phil. Trans. R. Soc. B*, 370(1678):20140329.

Gray, M. W. (1992). The endosymbiont hypothesis revisited. *International review of cytology*, 141:233–357.

Gray, M. W. (2012). Mitochondrial evolution. *Cold Spring Harbor perspectives in biology*, 4(9):a011403.

Gray, M. W., Burger, G., and Lang, B. F. (1999). Mitochondrial evolution. *Science*, 283(5407):1476–1481.

Gray, M. W. and Doolittle, W. F. (1982). Has the endosymbiont hypothesis been proven? *Microbiological Reviews*, 46(1):1.

Greenacre, M. (2007). *Correspondence analysis in practice*. CRC press.

Gribaldo, S., Poole, A. M., Daubin, V., Forterre, P., and Brochier-Armanet, C. (2010). The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nature reviews. Microbiology*, 8(10):743–52.

Gunter, T. E. and Pfeiffer, D. R. (1990). Mechanisms by which mitochondria transport calcium. *American Journal of Physiology-Cell Physiology*, 258(5):C755–C786.

Gupta, R. S. and Golding, G. B. (1996). The origin of the eukaryotic cell. *Trends in biochemical sciences*, 21(5):166–171.

Han, T.-M. and Runnegar, B. (1992). Megascopic eukaryotic algae from the 2.1-billion-year-old negaunee iron-formation, michigan. *Science*, 257(5067):232–235.

Hashimoto, T. and Hasegawa, M. (1996). Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1α/tu and 2/g. *Advances in biophysics*, 32:73–120.

Hirt, R. P., Logsdon, J. M., Healy, B., Dorey, M. W., Doolittle, W. F., and Embley, T. M. (1999). Microsporidia are related to fungi: evidence from the largest subunit of rna polymerase ii and other proteins. *Proceedings of the National Academy of Sciences*, 96(2):580–585.

Hoepfner, D., Schildknegt, D., Braakman, I., Philippsen, P., and Tabak, H. F. (2005). Contribution of the endoplasmic reticulum to peroxisome formation. *Cell*, 122(1):85–95.

Horiike, T., Hamada, K., Kanaya, S., and Shinozawa, T. (2001). Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria is revealed by homology-hit analysis. *Nature cell biology*, 3(2):210–4.

Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L. P., Marcet-Houben, M., and Gabaldón, T. (2013). Phylomedb v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic acids research*, page gkt1177.

Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010). Ete: a python environment for tree exploration. *BMC bioinformatics*, 11(1):1.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). Ete 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, 33(6):1635–1638.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95.

Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J., and O'Donovan, C. (2014). The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Research*.

Iwabe, N., Kuma, K.-i., Hasegawa, M., Osawa, S., and Miyata, T. (1989). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proceedings of the National Academy of Sciences*, 86(23):9355–9359.

Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *TRENDS in Genetics*, 22(4):225–231.

Kamer, K. J. and Mootha, V. K. (2015). The molecular era of the mitochondrial calcium uniporter. *Nature Reviews Molecular Cell Biology*, 16(9):545–553.

Karnkowska, A., Vacek, V., Zubáčová, Z., Treitli, S. C., Petrželková, R., Eme, L., Novák, L., Žárskỳ, V., Barlow, L. D., Herman, E. K., et al. (2016). A eukaryote without a mitochondrial organelle. *Current Biology*, 26(10):1274–1284.

Katoh, K. and Toh, H. (2008). Recent developments in the mafft multiple sequence alignment program. *Briefings in bioinformatics*, 9(4):286–298.

Katz, L. A. (2002). Lateral gene transfers and the evolution of eukaryotes: theories and data. *International Journal of Systematic and Evolutionary Microbiology*, 52(5):1893–1900.

Kauffman, S. A. (2011). Approaches to the origin of life on earth. *Life*, 1(1):34–48.

Keeling, P. J. (2010). The endosymbiotic origin, diversification and fate of plastids. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1541):729–748.

Keeling, P. J. (2013). The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annual review of plant biology*, 64:583–607.

Keeling, P. J. (2014). The impact of history on our perception of evolutionary events: endosymbiosis and the origin of eukaryotic complexity. *Cold Spring Harbor perspectives in biology*, 6(2):1–14.

Kellis, M., Patterson, N., Birren, B., Berger, B., and Lander, E. S. (2004). Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *Journal of Computational Biology*, 11(2-3):319–355.

Knoll, A. H., Javaux, E. J., Hewitt, D., and Cohen, P. (2006). Eukaryotic organisms in proterozoic oceans. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361(1470):1023–1038.

Koonin, E. V. (2009). Darwinian evolution in the light of genomics. *Nucleic acids research*, page gkp089.

Koonin, E. V. (2010). The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome biology*, 11(5):209.

Koonin, E. V., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Krylov, D. M., Makarova, K. S., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., et al. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome biology*, 5(2):R7.

Koonin, E. V. and Yutin, N. (2014). The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. *Cold Spring Harbor perspectives in biology*, 6(4):a016188.

Koumandou, V. L., Wickstead, B., Ginger, M. L., van der Giezen, M., Dacks, J. B., and Field, M. C. (2013). Molecular paleontology and complexity in the last eukaryotic common ancestor. *Critical reviews in biochemistry and molecular biology*, 48(4):373–96.

Ku, C., Nelson-Sathi, S., Roettger, M., Garg, S., Hazkani-Covo, E., and Martin, W. F. (2015). Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. *Proceedings of the National Academy of Sciences*, 112(33):10139–10146.

Lake, J. A., Henderson, E., Oakes, M., and Clark, M. W. (1984). Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proceedings of the National Academy of Sciences*, 81(12):3786–3790.

Lake, J. A. and Rivera, M. C. (1994). Was the nucleus the first endosymbiont? *Proceedings of the National Academy of Sciences*, 91(8):2880–2881.

Lake, J. A., Skophammer, R. G., Herbold, C. W., and Servin, J. A. (2009). Genome beginnings: rooting the tree of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1527):2177–2185.

Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC evolutionary biology*, 7 Suppl 1:S4.

Latorre, A. and Moya, A. (2013). *Role of Symbiosis in Evolution*, pages 63–70. Springer New York, New York, NY.

Lester, L., Meade, A., and Pagel, M. (2006). The slow road to the eukaryotic genome. *BioEssays*, 28(1):57–64.

López-García, P. and Moreira, D. (2015). Open questions on the origin of eukaryotes. *Trends in ecology & evolution*, 30(11):697–708.

Lynch, M., Koskella, B., and Schaack, S. (2006). Mutation pressure and the evolution of organelle genomic architecture. *Science*, 311(5768):1727–1730.

Makarova, K. S., Wolf, Y. I., Mekhedov, S. L., Mirkin, B. G., and Koonin, E. V. (2005). Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic acids research*, 33(14):4626–38.

Margulis, L. (1993). *Symbiosis in Cell Evolution: Microbial Communities in the Archean and Proterozoic Eons*. Freeman.

Margulis, L. (1996). Archaeal-eubacterial mergers in the origin of eukarya: phylogenetic classification of life. *Proceedings of the national academy of sciences*, 93(3):1071–1076.

Margulis, L., Dolan, M. F., and Guerrero, R. (2000). The chimeric eukaryote: origin of the nucleus from the karyomastigont in amitochondriate protists. *Proceedings of the National Academy of Sciences*, 97(13):6954–6959.

Martijn, J. and Ettema, T. J. G. (2013). From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. *Biochemical Society transactions*, 41(1):451–7.

Martin, W. (1999). A briefly argued case that mitochondria and plastids are descendants of endosymbionts, but that the nuclear compartment is not. *Proceedings of the Royal Society of London B: Biological Sciences*, 266(1426):1387–1395.

Martin, W. (2010). Evolutionary origins of metabolic compartmentalization in eukaryotes. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1541):847–855.

Martin, W. et al. (1999). Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays*, 21(2):99–104.

Martin, W. and Koonin, E. V. (2006). A positive definition of prokaryotes. *Nature*, 442(7105):868–868.

Martin, W. and Kowallik, K. (1999). Annotated english translation of mereschkowsky's 1905 paper 'über natur und ursprung der chromatophoren impflanzenreiche'. *European Journal of Phycology*, 34(3):287–295.

Martin, W. and Müller, M. (1998). The hydrogen hypothesis for the first eukaryote. *Nature*, 392(6671):37–41.

Martin, W. F. (2005). Archaebacteria (Archaea) and the origin of the eukaryotic nucleus. *Current opinion in microbiology*, 8(6):630–7.

Martin, W. F., Roettger, M., Ku, C., Garg, S. G., Nelson-Sathi, S., and Landan, G. (2016). Late mitochondrial origin is pure artefact. *bioRxiv*, page 055368.

Matrin, W. and Kowallik, K. V. (1999). Annotated english translation of mereschkowsky's 1905 paper 'Über natur und ursprung der chromatophoren im pflanzenreiche'. *European Journal of Phycology*, 34:287–295.

Moreira, D. and López-García, P. (1998). Symbiosis between methanogenic archaea and δ-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *Journal of molecular evolution*, 47(5):517–530.

Mossel, E. (2003). On the impossibility of reconstructing ancestral data and phylogenies. *Journal of computational biology*, 10(5):669–676.

Müller, M., Mentel, M., van Hellemond, J. J., Henze, K., Woehle, C., Gould, S. B., Yu, R.-Y., van der Giezen, M., Tielens, A. G., and Martin, W. F. (2012). Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiology and Molecular Biology Reviews*, 76(2):444–495.

Nass, M. M. and Nass, S. (1963). Intramitochondrial fibers with dna characteristics i. fixation and electron staining reactions. *The Journal of cell biology*, 19(3):593–611.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2014). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Nickelsen, J., Rengstl, B., Stengel, A., Schottkowski, M., Soll, J., and Ankele, E. (2011). Biogenesis of the cyanobacterial thylakoid membrane system–an update. *FEMS microbiology letters*, 315(1):1–5.

Nowack, E. C. and Melkonian, M. (2010). Endosymbiotic associations within protists. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1541):699–712.

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304.

Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20.

Olsen, G. J. and Woese, C. R. (1993). Ribosomal rna: a key to phylogeny. *The FASEB journal*, 7(1):113–123.

O'Malley, M. A., Wideman, J. G., and Ruiz-Trillo, I. (2016). Losing complexity: The role of simplification in macroevolution. *Trends in ecology & evolution*.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, 276(5313):734–740.

Pace, N. R. (2006). Time for a change. *Nature*, 441(7091):289–289.

Pace, N. R. (2009). Mapping the tree of life: progress and prospects. *Microbiology and Molecular Biology Reviews*, 73(4):565–576.

Penel, S., Arigon, A.-M., Dufayard, J.-F., Sertier, A.-S., Daubin, V., Duret, L., Gouy, M., and Perrière, G. (2009). Databases of homologous gene families for comparative genomics. *BMC bioinformatics*, 10 Suppl 6:S3.

Perocchi, F., Gohil, V. M., Girgis, H. S., Bao, X. R., McCombs, J. E., Palmer, A. E., and Mootha, V. K. (2010). Micu1 encodes a mitochondrial ef hand protein required for ca2+ uptake. *Nature*, 467(7313):291–296.

Philippe, H. (2000). Early–branching or fast–evolving eukaryotes? an answer based on slowly evolving positions. *Proceedings of the Royal Society of London B: Biological Sciences*, 267(1449):1213–1221.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*, 9(3):e1000602.

Pittis, A. A. and Gabaldón, T. (2016). Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature*, 531(7592):101–104.

Plovanich, M., Bogorad, R. L., Sancak, Y., Kamer, K. J., Strittmatter, L., Li, A. A., Girgis, H. S., Kuchimanchi, S., De Groot, J., Speciner, L., et al. (2013). Micu2, a paralog of micu1, resides within the mitochondrial uniporter complex to regulate calcium handling. *PloS one*, 8(2):e55785.

Poole, A. and Penny, D. (2007). Eukaryote evolution: engulfed by speculation. *Nature*, 447(7147):913–913.

Poole, A. M. and Gribaldo, S. (2014). Eukaryotic Origins: How and When Was the Mitochondrion Acquired? *Cold Spring Harbor perspectives in biology*.

Porter, S. M. (2004). The fossil record of early eukaryotic diversification. *Paleontological Society Papers*, 10:35.

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., Gabaldón, T., Rattei, T., Creevey, C., Kuhn, M., et al. (2013). eggnog v4. 0: nested orthology inference across 3686 organisms. *Nucleic acids research*, page gkt1253.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490.

Quang, L. S., Gascuel, O., and Lartillot, N. (2008). Empirical Profile mixture models for phylogenetic reconstruction Supplementary Material Appendix : EM algorithm. *Bioinformatics*, 24:2317–2323.

Raff, R. A. and Mahler, H. R. (1972). The non symbiotic origin of mitochondria. *Science*, 177(4049):575–582.

Raffaello, A., De Stefani, D., Sabbadin, D., Teardo, E., Merli, G., Picard, A., Checchetto, V., Moro, S., Szabò, I., and Rizzuto, R. (2013). The mitochondrial calcium uniporter is a multimer that can include a dominant-negative pore-forming subunit. *The EMBO journal*, 32(17):2362–2376.

Rasmussen, M. D. and Kellis, M. (2007). Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome research*, 17(12):1932–1942.

Ribeiro, S. and Golding, G. B. (1998). The mosaic nature of the eukaryotic nucleus. *Molecular Biology and Evolution*, 15(7):779–788.

Ris, H. and Plaut, W. (1962). Ultrastructure of dna-containing areas in the chloroplast of chlamydomonas. *The Journal of cell biology*, 13(3):383–391.

Rivera, M. C., Jain, R., Moore, J. E., and Lake, J. a. (1998). Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11):6239–6244.

Rivera, M. C. and Lake, J. A. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, 431(7005):152–155.

Rizzuto, R., Pinton, P., Carrington, W., Fay, F. S., Fogarty, K. E., Lifshitz, L. M., Tuft, R. A., and Pozzan, T. (1998). Close contacts with the endoplasmic reticulum as determinants of mitochondrial ca2+ responses. *Science*, 280(5370):1763–1766.

Rochette, N. C., Brochier-Armanet, C., and Gouy, M. (2014). Phylogenomic Test of the Hypotheses for the Evolutionary Origin of Eukaryotes. *Molecular biology and evolution*, 31(4):1–14.

Rodríguez-Ezpeleta, N. and Embley, T. M. (2012). The SAR11 group of alpha-proteobacteria is not related to the origin of mitochondria. *PloS one*, 7(1):e30520.

Rogozin, I. B., Basu, M. K., Csürös, M., and Koonin, E. V. (2009). Analysis of rare genomic changes does not support the unikont-bikont phylogeny and suggests cyanobacterial symbiosis as the point of primary radiation of eukaryotes. *Genome biology and evolution*, 1:99–113.

Sagan, C., Tyson, N., and Druyan, A. (2013). *Cosmos*. Ballantine Books. Ballantine.

Sagan, L. (1967). On the origin of mitosing cells. *Journal of theoretical biology*, 14(3):255–74.

Sancak, Y., Markhard, A. L., Kitami, T., Kovács-Bogdán, E., Kamer, K. J., Udeshi, N. D., Carr, S. A., Chaudhuri, D., Clapham, D. E., Li, A. A., et al. (2013). Emre is an essential component of the mitochondrial calcium uniporter complex. *Science*, 342(6164):1379–1382.

Santarella-Mellwig, R., Pruggnaller, S., Roos, N., Mattaj, I. W., and Devos, D. P. (2013). Three-dimensional reconstruction of bacteria with a complex endomembrane system. *PLoS Biol*, 11(5):e1001565.

Schlüter, A., Fourcade, S., Ripp, R., Mandel, J. L., Poch, O., and Pujol, A. (2006). The evolutionary origin of peroxisomes: an er-peroxisome connection. *Molecular biology and evolution*, 23(4):838–845.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., and Ettema, T. J. G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*.

Speijer, D. (2011). Oxygen radicals shaping evolution: why fatty acid catabolism leads to peroxisomes while neurons do without it. *BioEssays*, 33(2):88–94.

Spencer, D. F., Schnare, M. N., and Gray, M. W. (1984). Pronounced structural similarities between the small subunit ribosomal rna genes of wheat mitochondria and escherichia coli. *Proceedings of the National Academy of Sciences*, 81(2):493–497.

Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, page btu033.

Stanier, R. (1963). *The microbial world, 2nd edition*. Prentice-Hall.

Stanier, R. Y. (1970). Some aspects of the biology of cells and their possible evolutionary significance. In *Symp Soc Gen Microbiol*, volume 20, pages 1–38.

Stanier, R. Y. and Niel, C. v. (1962). The concept of a bacterium. *Archives of Microbiology*, 42(1):17–35.

Subramanian, A. R., Kaufmann, M., and Morgenstern, B. (2008). Dialign-tx: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms for Molecular Biology*, 3(1):1.

Szalai, G., Csordás, G., Hantash, B. M., Thomas, A. P., and Hajnóczky, G. (2000). Calcium signal transmission between ryanodine receptors and mitochondria. *Journal of Biological Chemistry*, 275(20):15305–15313.

Szathmary, E. and Smith, J. M. (2000). The major evolutionary transitions. *Shaking the Tree: Readings from*, pages 32–47.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., et al. (2014). String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, page gku1003.

Takayama, T., Fujita, K., Suzuki, K., Sakaguchi, M., Fujie, M., Nagai, E., Watanabe, S., Ichiyama, A., and Ogawa, Y. (2003). Control of oxalate formation from l-hydroxyproline in liver mitochondria. *Journal of the American Society of Nephrology*, 14(4):939–946.

Taylor, S. W., Fahy, E., and Ghosh, S. S. (2003). Global organellar proteomics. *Trends in biotechnology*, 21(2):82–88.

Thiergart, T., Landan, G., Schenk, M., Dagan, T., and Martin, W. F. (2012). An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. *Genome biology and evolution*, 4(4):466–85.

Thrash, J. C., Boyd, A., Huggett, M. J., Grote, J., Carini, P., Yoder, R. J., Robbertse, B., Spatafora, J. W., Rappé, M. S., and Giovannoni, S. J. (2011). Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Scientific reports*, 1:13.

Timmis, J. N., Ayliffe, M. A., Huang, C. Y., and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics*, 5(2):123–135.

Torruella, G., Derelle, R., Paps, J., Lang, B. F., Roger, A. J., Shalchian-Tabrizi, K., and Ruiz-Trillo, I. (2012). Phylogenetic relationships within the opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Molecular Biology and Evolution*, 29(2):531–544.

Tovar, J. (2007). Mitosomes of parasitic protozoa: biology and evolutionary significance. In *Origin of Mitochondria and Hydrogenosomes*, pages 277–300. Springer.

Turner, B. M. (2007). Defining an epigenetic code. *Nature cell biology*, 9(1):2–6.

Vasington, F. D. and Murphy, J. V. (1962). Ca++ uptake by rat kidney mitochondria and its dependence on respiration and phosphorylation. *Journal of Biological Chemistry*, 237(8):2670–2677.

Wallace, I. M., O'Sullivan, O., Higgins, D. G., and Notredame, C. (2006). M-coffee: combining multiple sequence alignment methods with t-coffee. *Nucleic acids research*, 34(6):1692–1699.

Wallin, I. E. et al. (1927). *Symbionticism and the origin of species*. Williams & Wilkins company.

Wang, Z. and Wu, M. (2015). An integrated phylogenomic approach toward pinpointing the origin of mitochondria. *Scientific reports*, 5:7949.

Wickstead, B. and Gull, K. (2011). The evolution of the cytoskeleton. *The Journal of cell biology*, 194(4):513–25.

Wier, A. M., Sacchi, L., Dolan, M. F., Bandi, C., MacAllister, J., and Margulis, L. (2010). Spirochete Attachment Ultrastructure: Implications for the Origin and Evolution of Cilia. *Biological Bulletin*, 218(November 2009):25–35.

Williams, T. a., Foster, P. G., Cox, C. J., and Embley, T. M. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, 504(7479):231–6.

Williams, T. a., Foster, P. G., Nye, T. M. W., Cox, C. J., and Embley, T. M. (2012). A congruent phylogenomic signal places eukaryotes within the Archaea. *Proceedings. Biological sciences / The Royal Society*, 279(1749):4870–9.

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579.

Wolf, Y. I. and Koonin, E. V. (2013). Genome reduction as the dominant mode of evolution. *Bioessays*, 35(9):829–837.

Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J., and Woese, C. R. (1985). Mitochondrial origins. *Proceedings of the National Academy of Sciences*, 82(13):4443–4447.

Yutin, N., Makarova, K. S., Mekhedov, S. L., Wolf, Y. I., and Koonin, E. V. (2008). The deep archaeal roots of eukaryotes. *Molecular biology and evolution*, 25(8):1619–1630.

Zhang, J. and Yang, J.-R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, 16(7):409–420.

Zhaxybayeva, O. and Gogarten, J. P. (2007). Horizontal gene transfer, gene histories, and the root of the tree of life. *Planetary systems and the origin of life*.

Zhu, A., Guo, W., Jain, K., and Mower, J. P. (2014). Unprecedented heterogeneity in the synonymous substitution rate within a plant genome. *Molecular biology and evolution*, 31(5):1228–36.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. *Evolving genes and proteins*, 97:97–166.