UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1861

Genomic and evolutionary exploration of Asgard archaea

EVA F. CACERES

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019

ISSN 1651-6214
ISBN 978-91-513-0761-9
urn:nbn:se:uu:diva-393710

Dissertation presented at Uppsala University to be publicly examined in B22, Biomedical Center (BMC), Husargatan 3, Uppsala, Thursday, 14 November 2019 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Simonetta Gribaldo (Institut Pasteur, Department of Microbiology).

Abstract

Caceres, E. F. 2019. Genomic and evolutionary exploration of Asgard archaea. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1861. 88 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0761-9.

Current evolutionary theories postulate that eukaryotes emerged from the symbiosis of an archaeal host with, at least, one bacterial symbiont. However, our limited grasp of microbial diversity hampers insights into the features of the prokaryotic ancestors of eukaryotes. This thesis focuses on the study of a group of uncultured archaea to better understand both existing archaeal diversity and the origin of eukaryotes.

In a first study, we used short-read metagenomic approaches to obtain eight genomes of *Lokiarchaeum* relatives. Using these data we described the Asgard superphylum, comprised of at least four different phyla: Lokiarchaeota, Odinarchaeota, Thorarchaeota and Heimdallarchaeota. Phylogenetic analyses suggested that eukaryotes affiliate with the Asgard group, albeit the exact position of eukaryotes with respect to Asgard archaea members remained inconclusive. Comparative genomics showed that Asgard archaea genomes encoded homologs of numerous eukaryotic signature proteins (ESPs), which had never been observed in Archaea before. Among these, there were several components of proteins involved in vesicle formation and membrane remodelling.

In a second study, we used similar approaches to uncover additional members of the Asgard superphylum. Based on genome-centric metagenomics we recovered 69 new genomes from which we identified five additional candidate phyla: Freyarchaeota, Baldrarchaeota, Gefionarchaeota, Friggarchaeota and Idunnarchaeota. In this expanded dataset we could detect additional homologs for unreported ESPs. Updated phylogenies showed support for a scenario in which eukaryotes emerged from within Asgard archaea.

We further took advantage of the increased Asgard diversity to delimit the gene content of the last common archaeal ancestor of eukaryotes using ancestral reconstruction analyses. The results suggest that the archaeal host cell who gave rise to eukaryotes already contained many of the genes associated with eukaryotic cellular complexity. Based on these analyses, we discussed the metabolic capabilities of the archaeal ancestor of eukaryotes.

Finally, we reconstructed several nearly complete Lokiarchaeota genomes, one of them in only three contigs, using both short- and long-read metagenomics. These analyses indicate that long-read metagenomics is a promising approach to obtain highly complete and contiguous genomes directly from environmental samples, even from complex populations in the presence of microdiversity and low abundant members. This study further supports that the presence of ESPs in Asgard genomes is not the result of assembly and binning artefacts.

In conclusion, this thesis highlights the value of using culture-independent approaches together with phylogenomics and comparative genomics to improve our understanding of microbial diversity and to shed light into relevant evolutionary questions.

Keywords: archaea, Asgard, eukaryogenesis, metagenomics, genome binning, phylogenetics, phylogenomics, comparative genomics, gene tree-species tree reconciliation, ancestral reconstruction, long-read metagenomics

Eva F. Caceres, Department of Cell and Molecular Biology, Box 596, Uppsala University, SE-75124 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-0761-9
urn:nbn:se:uu:diva-393710 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-393710)

To my high school teachers
Rufino and Charo who sparked
my interest in science

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Zaremba-Niedzwiedzka, K.*, **Caceres, EF**.*, Saw, JH.*, Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, KW., Anantharaman, K., Starnawski, P., Kjeldsen , KU., Stott, MB., Nunoura, T., Banfield, JF., Schramm, A., Baker, BJ., Spang, A., Ettema, TJG. (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, 541:353–358

II Eme, LE.*, **Caceres, EF.*,** Tamarit, D., Seitz, KW., Dombrowski, N., Homa, F., Saw, JH., Lombard, J., Li, W., Hua, Z., Chen, L., Banfield, JF., Reysenbach, A., Nunoura, T., Stott, MB., Schramm, A., Kjeldsen, KU., Baker, BJ., Ettema, TJG. (2019) Expanded diversity of Asgard archaea points to Idunnarchaeota as closest relatives of eukaryotes. *Manuscript*

III **Caceres, EF.*,** Eme, LE.*, De Anda, V., Baker, BJ., Ettema, TJG. (2019) Ancestral reconstruction of Asgard archaea provides insight into the gene content of the archaeal ancestor of eukaryotes. *Manuscript*

IV **Caceres, EF**., Lewis, WH., Homa, F., Martin, T., Schramm, A., Kjeldsen, KU., Ettema, TJG. (2019) Reconstruction of a near-complete Lokiarchaeota genome using long- and short-read metagenomics of complex sediment samples. *Manuscript*

(*) Equal contribution

Papers by the author not included in this thesis

1. Spang, A., Stairs, CW., Dombrowski, N., Eme, L., Lombard, J., **Caceres, EF.**, Greening, C., Baker, BJ., Ettema, TJG. (2019) Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nature Microbiology 10(1):1822.*

2. Narrowe, AB., Spang, A., Stairs, CW., **Caceres, EF.**, Baker, BJ., Miller, CS., Ettema, TJG. (2018) Complex Evolutionary History of Translation Elongation Factor 2 and Diphthamide Biosynthesis in Archaea and Parabasalids. *Genome Biology and Evolution, 10(9):2380-2393*

3. Spang, A., Eme, L., Saw, JH., **Caceres, EF.**, Zaremba-Niedzwiedzka, K., Lombard, J., Guy, L., Ettema, TJG. (2018) Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLoS Genetics, 14(3):e1007080*

4. Hennell James, R., **Caceres, EF.**, Escasinas, A., Alhasan, H., Howard, JA., Deery, MJ., Ettema, TJG., Robinson, NP. (2017) Functional reconstruction of a eukaryotic-like E1/E2(RING) E3 ubiquitylation cascade from an uncultured archaeon. *Nature Communications, 8:1120*

5. Gomez-Velazquez, M., Badia-Careaga, C., Lechuga-Vieco, AV., Nieto-Arellano, R., Tena, JJ., Rollan, I., Alvarez, A., Torroja, C., **Caceres, EF.**, Roy, AR., Galjart, N., Delgado-Olguin, P., Sanchez-Cabo, F., Enriquez, JA., Gomez-Skarmeta, JL., Manzanares, M. (2017) CTCF counter-regulates cardiomyocyte development and maturation programs in the embryonic heart. *PLoS Genetics, 13(8):e1006985*

6. Spang, A., **Caceres, EF.**, and Ettema, TJG. (2017) Genomic exploration of the diversity, ecology and evolution of the archaeal domain of life. *Science, 357:6351*

7. Marshall, IPG., Starnawski, P., Cupit, C., **Caceres, EF.,** Ettema, TJG., Schramm, A., Kjeldsen, KU. (2017) The novel bacterial phylum Calditrichaeota is diverse, widespread and abundant in marine sediments

and has the capacity to degrade detrital proteins. *Environmental Microbiology Reports, 9(4):397-403*

8. **Caceres, EF.**, Hurst, LD. (2013) The evolution, impact and properties of exonic splice enhancers. *Genome Biology, 14(12):R143*

9. Wu, X., Tronholm, A., **Caceres, EF.**, Tovar-Corona, JM., Chen, L., Urrutia, AO., Hurst, LD. (2013) Evidence for deep phylogenetic conservation of exonic splice-related constraints: splice-related skews at exonic ends in the brown alga Ectocarpus are common and resemble those seen in humans. *Genome Biology and Evolution, 5(9):1731-1745*

Contents

Abbreviations

AAG	Ancient archaeal group
ANME	ANaerobic MEthane-oxidizing archaea
ARP	Actin-related protein
DBG	De Bruijn graph
DNA	Deoxyribonucleic acid
DPANN	Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota
DSAG	Deep sea archaeal group
ESCRT	Endosomal sorting complex required for transport
ESP	Eukaryotic signature protein
GTR	General time reversible
HGT	Horizontal gene transfer
HMW	High molecular weight
LACAE	Last archaeal common ancestor of eukaryotes
LBA	Long-branch attraction
LECA	Last eukaryotic common ancestor
LG	Le and Gascuel
MAG	Metagenome-assembled genome
MCMC	Markov chain Monte Carlo
MHVG	Marine hydrothermal vent group
MRO	Mitochondria-related organelle
MSA	Multiple sequence alignment
OLC	Overlap, layout, consensus
PCR	Polymerase chain reaction
PVC	Planctomycetes, Verrucomicrobia and Chlamydiae
RNA	Ribonucleic acid
SR	Short reads
SSU rRNA	Small subunit ribosomal RNA
TACK	Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota
TRAPP	Transport protein particle
WAG	Whelan and Goldman

Introduction

Over the last decades, thanks to the development of sequencing technologies and culture-independent approaches, we have started to unravel the genomic diversity of the microorganisms that inhabit our planet. With the current methods, we now have the possibility to study numerous microbial groups that, for so long, have remained out of our reach.

In this thesis work, I will describe our efforts to understand one of these understudied groups of microorganisms, now known as the Asgard archaea superphylum. We used metagenomic approaches to obtain genome sequences of Asgard lineages for which no genomic information was available before. By studying their genomes within a comparative genomics and evolutionary framework, we have learnt not only about the cellular capabilities of this group but also about their role in the early evolution of eukaryotes.

In the forthcoming sections, I will introduce the topic and the methods around which this thesis is centred, and summarize the main results of the analyses carried out as part of my doctoral work. This introduction is followed by the four articles that comprise my main research projects. Given the format constraints, the supplementary material is only attached when the size of the figures and tables allowed. Alternatively, electronic links are provided, with the exception of Paper I, for which the supplementary material can be found on the publisher's website.

Finally, I would like to mention that the research presented here is the result of collaborative efforts and that it would not have been possible without the concerted work of many people. The contribution of each person involved is fundamental, from taking samples and preparing sequencing libraries to providing guidance and supervision. I firmly believe in the strength of collaborative science, in which researchers with different skill sets can all work together to make more compelling and comprehensive studies. To recognize the efforts of all the people involved, I will refrain from using "I" and "my" for the most part of the text.

Archaea

The discovery of the Third Domain

Archaea were recognized as a group of prokaryotes fundamentally different from bacteria in 1977 by Woese and Fox (Woese and Fox, 1977). At that time, all organisms were divided into two categories, *eukaryotes* and *prokaryotes,* with the latter group composed solely of bacteria. While eukaryotes had cells with a nucleus and internal organelles, prokaryotes lacked such structures (McLaughlin and Dayhoff, 1970). This eukaryote-prokaryote dichotomy was considered the most basic evolutionary division of life. Woese and Fox showed that, in spite of their apparent morphological similarities, Archaea formed a domain of life different from Bacteria and, based on these results, they proposed a tripartite view of life, with Eukarya, Bacteria and Archaea being the most basal divisions (Woese and Fox, 1977).

At that time, taxonomical systems were primarily reliant on phenotypical and morphological traits. Prokaryotes were classified based on the absence of eukaryotic traits such as the nucleus and certain intracellular organelles (Stanier and Van Niel, 1962). One of the problems of classifying organisms based on the absence of a certain feature is that there are no degrees of variation of such trait between different organisms – the feature is not present – and, therefore, it cannot be used to generate phylogenies. As a consequence, the taxonomical system at the time largely excluded microbes.

On the other hand, the construction of phylogenies was still at its infancy and mostly based on protein sequences (Fitch and Margoliash, 1967; Zuckerkandl and Pauling, 1965). Woese realized that the ribosomal RNA could be a good molecular marker to generate continuous classifications between all organisms as it was conserved and present in all life forms (Woese, 1987). Woese and Fox created oligonucleotides catalogues of the small subunit of the ribosomal RNA (SSU rRNA) of several prokaryotes and eukaryotes and compared them to produce an evolutionarily-coherent taxonomy that was solely based on molecular data allowing the identification of the so-called Third Domain of life (Woese and Fox, 1977).

The only Archaea included in this study were methanogens, a group of microorganisms that produce methane in anaerobic conditions. That strange metabolism was believed to reflect the primitive atmospheric conditions of the planet and, thus, it was considered an ancient phenotype (Woese, 1977). The original term “Archaebacteria” (from the Greek “ancient” “rod”) made

reference to that idea; although this assumption is today is disregarded, as we know that other metabolisms exist in Archaea. By 1990, Woese and others recommended abandoning the original term "Archaebacteria" in favour of the shorter version "Archaea", since it incorrectly suggested that Archaea and Bacteria were related to one another (Woese et al., 1990a). Notwithstanding, the word Archaebacteria is still in use in the scientific literature, propagating misleading connections to Bacteria.

As many dogma-challenging theories, Woese and Fox' work was criticized by many scientists who strongly rejected their methodology and did not accept Archaea as an independent domain of life. The paradigm shift required some time and the work of many other scientists. During the following years, data supporting the distinctiveness of Archaea started to pile-up. Even though in terms of size and morphology Archaea resembled Bacteria, there were important differences between them. For example, the cell walls in Archaea lacked peptidoglycan (Kandler and Hippe, 1977) and their lipids were crosslinked via ether bonds instead of the ester bonds found in Bacteria (Langworthy et al., 1972). Furthermore, it was soon realized that in many other aspects archaea were more similar to eukaryotes than to bacteria. Certain proteins were more closely related to their eukaryotic homologs – such as the DNA-dependent RNA polymerase (Zillig et al., 1979) – and some were only found in Archaea and Eukarya to the exclusion of Bacteria. The publication of the first archaeal genome, almost 20 years later, marked the end of a period of the denial of the Archaea as a separate domain of life (Bult et al., 1996).

During the first years after their discovery, archaea were mainly found in environments with extreme conditions (e.g., high temperatures or high salinity) where they can be abundant players of the microbial communities. By that time, the study of microbes was carried out by isolating and culturing strains, an approach with important limitations (see "Traditional methods"). Initially, archaea that successfully grew in laboratory conditions showed similar lifestyles (e.g., methanogenesis, halophilism and thermophilic sulfur metabolism) giving the false impression that most archaeal phenotypes/diversity were already discovered by 1987 (Woese, 1987). The lack of adequate technologies and approaches needed for their study together with their relatively low interest in human and human-associated research made Archaea go unnoticed and remain understudied for many years.

In the mid-1980s, Norman Pace and co-workers established a method that allowed the exploration of the microbial diversity bypassing the culturing step (Pace et al., 1986). Their approach consisted of recovering rRNA gene sequences from all organisms present in a sample to estimate the relative abundances and identities of the community members living in an environment. These rRNA gene sequence surveys revealed that, contrary to what it was thought, archaea were ubiquitous and diverse, ultimately

falsifying the assumption that all archaea are extremophiles. Over time, this approach became a standard procedure and phylogenies of SSU rRNA gene sequences showed an increasing number of existing archaeal lineages. However, in-depth analyses and available complete genomes were still restricted to a small number of cultivated representatives (Pace, 2009).

During the past decade, the rise of independent-culture approaches such as metagenomics and single-cell genomics has made possible genomic reconstructions of uncultivated archaea, advancing our understanding of the archaeal biology and evolution (see "Genomic exploration of archaea"). The more recent use of long-read sequencing technologies in metagenomics will prove invaluable for generating high-quality genomes of uncultivated microorganisms (see "Paper IV") (Nicholls et al., 2018). Indeed, with innovative technology and software, it will soon become common practice to recover complete genomes from an environment, allowing for continued studies of these fascinating organisms.

Archaeal diversity

Molecular investigations of diverse environments have revealed that archaea can live in a wide range of environments, including sediments and soils, aquatic habitats, hot springs, hydrothermal vents, the rumen and gut of certain animals, etc. (Chaban et al., 2006). The estimated average abundance of archaea is around 20% in oceanic waters (Karner et al., 2001), 2% in surface soil layers (Bates et al., 2011) and 37% in subseafloor sediments (Hoshino and Inagaki, 2019), although these percentages can show important deviations depending on the specific location. In humans, archaea have been found living in the gastrointestinal tract, the oral cavity, the skin, and the vagina (Bang and Schmitz, 2015), where some species can amount to 14% of the microbiome according to some estimates (Tyakht et al., 2013).

The ubiquity of archaea in diverse environments is mirrored in the disparate lifestyles that different lineages display. A wide variety of metabolisms have been reported in Archaea including methanogenesis, methane oxidation, ammonia oxidation, denitrification and sulfate reduction among others (Kletzin, 2007). Through these biochemical reactions, archaea can significantly change the chemical composition in these environments, impacting availability and form of the elements and molecules present. This makes some archaea major contributors to the nutrient cycles (Offre et al., 2013).

In addition, archaea can be free-living or depend on one or several organisms to survive. Archaea can establish close associations with other archaea, bacteria or eukaryotes (Moissl-Eichinger and Huber, 2011). Examples of this are the archaeal symbiont *Nanoarchaeum equitans (Huber et al., 2002)*, the archaeal-bacterial consortium formed by anaerobic

methane-oxidizing archaea (ANME) and sulfate-reducing bacteria (Boetius et al., 2000); and the eukaryotic endosymbiont *Methanobrevibacter* (Gijzen et al., 1991; Lind et al., 2018), respectively. Strikingly, no archaeal parasite of animals has been found until now (Abedon, 2013). Even though there are several studies indicating potential correlations between some archaea and human diseases, no evidence for direct pathogenic effects of any archaeal species has been reported up to date (Mahnert et al., 2018).

The archaeal tree has undergone a dramatic transformation since 1977 (Adam et al., 2017; Spang et al., 2017). Originally, the archaeal taxonomy consisted uniquely in two phyla (originally considered kingdoms): Euryarchaeota and Crenarchaeota. To date, there are four high-level archaeal ranks recognized: Euryarchaeota, TACK or Proteoarchaeota (the group that includes the original Crenarchaeota), DPANN and Asgard archaea (see "Paper I and II"). However, the position of various clades and members is still unresolved. Understudied clades for which only few representatives are sequenced or fast-evolving taxa are especially difficult to place (see "Inferring evolution"), such as Korarchaeota and DPANN. Additionally, inferring the archaeal root has also turned out to be challenging, with studies suggesting conflicting placements (Petitjean et al., 2014; Raymann et al., 2015; Williams et al., 2017a).

Unfortunately, the current archaeal classification is inconsistent and paradoxical. During years, clades of uncultured lineages have been assigned to different taxonomic levels without following any systematic criteria. Therefore, some taxonomical decisions might seem arbitrary, as illustrated by the case of the Euryarchaeota and the Proteoarchaeota. While the first is considered a phylum the latter has received a superphylum rank. The need of a congruent archaeal classification with updated taxonomical criteria and nomenclature has already been stressed (Gribaldo and Brochier-Armanet, 2012; Hugenholtz et al., 2016; Konstantinidis et al., 2017; Yarza et al., 2014). Reaching a consensus on the archaeal classification that is congruent with the evolutionary relationships between archaea will inevitably require the use of reliable phylogenetic reconstructions and the study of diverse lineages that can fill the gaps existing in the archaeal tree.

Archaea and the origin of the eukaryotes

The eukaryotic cell

Independently of their evolutionary histories, cells can be divided into eukaryotic and prokaryotic according to their cellular organization. A typical eukaryotic cell has a higher grade of intracellular compartmentalization than the average prokaryotic cell. This is typified by the presence of membrane-bound organelles – such as mitochondria – and a developed endomembrane system that includes the nuclear membrane and the continuous endoplasmic reticulum, the Golgi apparatus, lysosomes, endosomes and vesicles among others. Such intricate internal compartmentalization is absent in prokaryotes. Nevertheless, intracellular structures have been observed in both Bacteria and Archaea. Some examples are the magnetosomes used by some bacteria to align themselves to geomagnetic field lines; the anammoxosomes in which anaerobic ammonia oxidation occurs; or other intracellular membrane structures observed in members of the *Planctomycetes*, *Verrucomicrobiae*, and *Chlamydiae* (PVC) bacterial superphylum and the thermophilic archaeon *Ignococcus hospitalis* (Grant et al., 2018; Shively, 2006).

Generally speaking, eukaryotes have larger cells than prokaryotes. A typical bacterium such as *Escherichia coli* or *Bacillus subtilis* has average cell volumes between ~1-2 μm^3 (Heim et al., 2017; Lynch and Marinov, 2017) while human cells can range between ~30-4000000 μm^3 (Gilmore et al., 1995; Goyanes et al., 1990). However, this is by no means a delimiting trait and cases of very large prokaryotes and tiny eukaryotes do exist. For example, the bacteria *Thiomargarita namibiensis* is visible by the human eye, reaching cell volumes of 2.2×10^8 μm^3 (Levin and Angert, 2015; Schulz et al., 1999). On the opposite side of the spectrum, the green algae *Ostreoccocus tauri* is considered the smallest eukaryote identified until now with a cellular volume of 0.91μm^3 (Courties et al., 1994; Henderson et al., 2007).

Similarly, the eukaryotic genomes are usually bigger than the prokaryotic ones, albeit overlap in sizes exists between them. The haploid nuclear genome size of eukaryotes ranges between 2.3 Megabase pairs (Mbp) and 150 000 Mbp; whereas the prokaryotic genome sizes are between 140 kilobase pairs (kbp) and 15 Mbp (Elliott and Ryan Gregory, 2015). Commonly, these large eukaryotic genomes display low gene densities that contrast with prokaryotes, in which non-coding regions represent a small

fraction of their genome. Exceptions are seen in non-free-living eukaryotes whose chromosomes have been independently reduced and/or compacted (Keeling and Slamovits, 2005). Another feature characteristic of eukaryotic genomes is the presence of telomeres, centromeres and complex regulatory elements that are absent in prokaryotes.

Furthermore, eukaryotic genes consist of coding sequences (exons) disrupted by non-coding fragments (introns) that need to be removed before translation to generate functional proteins. By keeping or removing introns, eukaryotes can generate slightly different versions of the same gene, also referred to as isoforms, increasing the complexity of their proteomes. The machinery responsible for the removal of the introns is the spliceosome, an intricate eukaryotic complex absent in Bacteria and Archaea. Nevertheless, introns that are independent of this complex are found in prokaryotes (Lambowitz and Zimmerly, 2004; Nawrocki et al., 2018). In eukaryotes, splicing takes place inside the nucleus and is coupled with the export of mature transcripts to the cytoplasm, where translation takes place. This is in contrast to Bacteria and Archaea, where transcription and translation are coupled and occur simultaneously.

In general, the eukaryotic cell is associated with a high degree of complexity that can be observed at many levels. Eukaryotes have molecular machineries that are generally more elaborate than the archaeal and bacterial versions, with some protein complexes being completely absent in prokaryotes. Numerous gene duplications, functionalization and *de novo* originations observed in their genomes have probably allowed such high level of specialization and sophistication (Conant and Wolfe, 2008; Makarova et al., 2005; McLysaght and Guerzoni, 2015) and the support of eukaryotic specific functions such as the ability to perform meiotic sex and phagocytosis. Nonetheless, although previously many features have been considered eukaryotic hallmarks, we know now that prokaryotic versions exist for many of them (Koonin, 2010) and their presence in eukaryotes is less unique than previously thought (Booth and Doolittle, 2015).

The origin of eukaryotes

The origin of the eukaryotic cell represents one of the major evolutionary transitions in the history of life. How did the cellular complexity observed in eukaryotes arise from simpler prokaryotic cells? Through the years, numerous hypotheses have attempted to provide an explanation to this question (Embley and Martin, 2006; Martin et al., 2001). These theories differ in the timing, the underlying mechanisms and the identity and nature of the ancestors involved. Yet, some key aspects are largely accepted.

First, it is widely recognized that mitochondria and mitochondria-related organelles (MROs) – such as hydrogenosomes and mitosomes – are the

descendants of a bacterial lineage whose closest living relatives belong to the Alphaproteobacteria (Roger et al., 2017; Sagan, 1967; Yang et al., 1985), albeit the exact lineage is still unclear (Martijn et al., 2018). The ancestor of mitochondria established an endosymbiotic relationship with a host cell and ultimately became an organelle. It is broadly accepted that mitochondria were already present in the last eukaryotic common ancestor (LECA) (Adl et al., 2012; Heiss et al., 2018; Pittis and Gabaldón, 2016) and that any loss of mitochondria occurred later in evolution (Karnkowska et al., 2016; Martijn et al., 2018; McInerney et al., 2014).

Second, eukaryotes genomes are chimeric and, in addition to eukaryotic specific genes, they harbour genes derived both from Archaea and Bacteria (Rivera et al., 1998). Many eukaryotic genes of archaeal origin are part of the systems that process and store genetic information in the cell (referred to as informational genes) (Yutin et al., 2008). In contrast, numerous metabolic genes are thought to be of bacterial origin (referred to as operational genes). Yet, just a fraction of the bacterial genes trace back to the Alphaproteobacteria and the origin of these other bacterial genes is still unclear with several possible explanations being proposed, including horizontal gene transfers (HGT), additional symbiotic events and phylogenetic noise (Ku et al., 2015; Pittis and Gabaldón, 2016; Thiergart et al., 2012). If the transfer of these bacterial genes happened before or after the acquisition of mitochondria is likewise debated (Eme et al., 2018).

Finally, eukaryotes harbour genes absent in both Archaea and Bacteria. Proteins present in all main eukaryotic groups that lack homologs in prokaryotes have been initially referred to as eukaryotic signature proteins (ESPs) and are often involved in key functions of the eukaryotic cell (Hartman and Fedorov, 2002). However, a fraction of ESPs might not be *bona fide* eukaryotic innovations and are likely to be present in prokaryotes or viruses, but remain unidentified. Since the definition of ESPs is based on homology criteria (or the absence thereof), with the development of more sensitive methods for homology detection and access to more comprehensive genomic databases, the number of ESPs is expected to change. In fact, many of the proteins originally defined as ESPs have now been identified in prokaryotes. However, referring to them as ESPs is still useful in such cases as the term highlights the prevalence of these proteins in eukaryotes and the fact that they are rarely found in prokaryotes.

Regarding the evolutionary relationship between Archaea and Eukarya, two opposing scenarios have coexisted in the literature for many years (Figure 1) (reviewed in Gribaldo et al. (2010)). The first one, known as the three domains (3D), suggests that Archaea and Eukarya are sisters lineages derived from a common ancestor that was neither an archaeon nor a eukaryote (Cavalier-Smith, 1987; Woese et al., 1990a). Interestingly, this theory implies that the homologs genes shared between Archaea and

Eukarya were transmitted from their common ancestor and are, therefore, ancestral to the diversification of any of these domains.

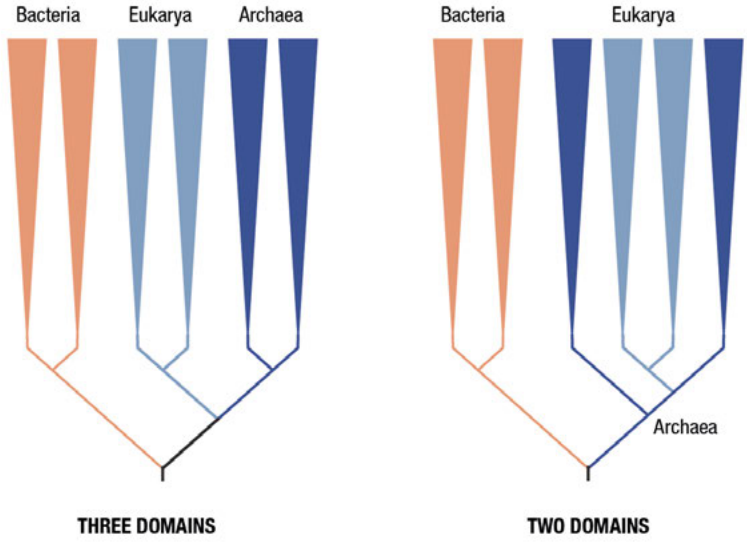

Figure 1. Schematic representation of the relationship between Archaea and Eukarya according to the "three domains" and "two domains" scenarios. In the three domains hypothesis, Bacteria, Eukarya and Archaea are seen as primary domains of life. In the two domains scenario, Eukarya is considered a secondary domain that originated from within Archaea.

The rival scenario, the two domains (2D) view, suggests that eukaryotes emerged from within the Archaea (Lake et al., 1984; Williams et al., 2013). According to this view, there were only two primary domains of life – Bacteria and Archaea – and Eukarya is seen as a secondary domain that evolved later from the Archaea. In this scenario, the term Archaea only refers to the cellular domain and lacks any phylogenetic connotation since it is viewed as a paraphyletic group. In contrast to the 3D view, it implies that the features shared between Archaea and eukaryotes arose after the diversification of Archaea.

Although these competing scenarios have been the subject of intense debates, the most recent data strongly favour the 2D topology (reviewed in Williams et al. (2013)). Phylogenetic analyses of concatenated protein-alignments using more complex evolutionary models and including a broader archaeal representation show convincingly that eukaryotes evolved from within Archaea and, thus, the host-cell was of archaeal nature (Cox et al., 2008; Foster et al., 2009b; Guy and Ettema, 2011; Guy et al., 2014; Lasek-Nesselquist and Gogarten, 2013; Spang et al., 2015; Williams and Embley, 2014; Williams et al., 2013; Williams et al., 2012). These results are further supported by the discovery of many ESPs in specific archaeal

groups, first within TACK (Guy and Ettema, 2011) and later within *Lokiarchaeum* (Spang et al., 2015) and Asgard (see "Paper I, II and IV")

Coupled with the 3D/2D debate is the controversy about the timing and mechanisms of the mitochondrial endosymbiosis (Eme et al., 2018; Lopez-Garcia and Moreira, 2015; Poole and Gribaldo, 2014). There are two main scenarios with regard of the relative timing and contribution of mitochondria acquisition: the mito-late and the mito-early. Different mechanistic models that explain the origin of eukaryotes have been proposed that are compatible with both scenarios. Mito-late favouring models suggest that most eukaryotic features associated with cellular complexity – such as developed endomembrane system, nucleus and cytoskeleton – arose before the symbiosis event. Having such features made possible the engulfment of the mitochondrial ancestor, with phagocytosis been proposed as a possible mechanism (Cavalier-Smith, 1983). On the other hand, mito-early models postulate that the mitochondrial endosymbiosis was the major event that led to the cellular complexity observed in eukaryotes. In this context, it is often argued that mitochondria provided an energy surplus that allowed the increase in complexity (Martin and Müller, 1998). Through the years, numerous variations of these and other models have been suggested (reviewed in Zachar and Szathmáry (2017)), including mito-intermediate models that assume a certain degree of cellular complexity in the host before the mitochondrial acquisition (Baum and Baum, 2014; Martijn and Ettema, 2013).

Nevertheless, none of the proposed models is exempt from criticism (Booth and Doolittle, 2015; Lynch and Marinov, 2017; Zachar and Szathmáry, 2017). The mito-late models are theoretically compatible with the existence of amitochondriate eukaryotes and, the fact that up to date no truly amitochondriate eukaryote has been found (Clark and Roger, 1995; Tovar et al., 1999; Tovar et al., 2003; Williams et al., 2002) is used as an argument against these models. Similarly, mito-early models were originally criticized because they provided no explanation about how phagocytosis – that was thought to be required to engulf the alphaproteobacterium – could have occurred without cellular complexity. Finding bacterial endosymbionts living within non-phagocytic bacteria weakened this argument (von Dohlen et al., 2001). Likewise, the reasoning behind theories claiming that the energy boost provided by the establishment of mitochondria was the trigger of cellular complexity has been challenged (Hampl et al., 2019; Lynch and Marinov, 2017; Zachar and Szathmáry, 2017).

Independently of the timing of the mitochondrial acquisition, current models provide different explanations for the lifestyle of the partners involved, the nature of their relationship, the selective advantage of their association and the mechanism of inclusion of the alphaproteobacterium. Various models suggest syntrophic interactions in which one species live off the products of another – with several types of metabolism being proposed

for the partners – (Martin and Müller, 1998; Moreira and Lopez-Garcia, 1998) or predation as the nature of the relationship (Cavalier-Smith, 2007). Apart from phagocytosis (Martijn and Ettema, 2013), other mechanisms to explain the acquisition of the mitochondrial ancestor have been hypothesized, such as an increasing contact surface followed by eventual membrane fusion (Baum and Baum, 2014).

The limited amount of information that can be obtained about a process that happened at least 1.9 billion years ago (Betts et al., 2018; Chernikova et al., 2011; Eme et al., 2014; Parfrey et al., 2011) has made difficult to judge which model is more accurate. Since evolution is a continuous process that never ceases, there is no living lineage reflecting the intermediate state of "prokaryote evolving into an eukaryote" as they went extinct or changed since then (Eme et al., 2018). The only way that we, nowadays, could find some "direct" evidence of these intermediate stages would be through microfossils or ancient DNA of such lineage. Nevertheless, the probabilities of finding such microfossils or DNA are extremely low and, even if we could detect them, they would add little information confidently. Other microbial fossil records are scarce and, by itself, not very helpful to answer questions about the features of the prokaryotic ancestors, and the mechanisms and order of the evolutionary events that happened during the eukaryogenesis. Hence, our knowledge about the origin of the eukaryotes mostly comes from comparative and phylogenetic analyses based on information of extant organisms. By studying their features and molecular sequences we can have a glimpse to their evolutionary past. Thus, the more we know about living microorganisms, the more accurate the evolutionary reconstructions are and the more realistic the proposed hypotheses become.

Phylogenetic methods based on molecular data can provide information about the pattern of diversification of species (see "Evolutionary history of species"). This information, together with molecular clocks and geological age estimates, can additionally be used to date such events (dos Reis et al., 2016; Ho and Duchêne, 2014). However, the information that geological records can provide for microbial evolution is minimal and not existent for the majority of the known clades. This has motivated the development of methods that make use of the information provided by horizontal gene transfer events between microorganisms to time speciation events. Albeit these approaches are promising, they still require further development and testing (Chauve et al., 2017a; Davin et al., 2018). A recent study based on genomic and fossil data has inferred a timescale of the early evolution of life on Earth. Their results show a long branch preceding the last eukaryotic common ancestor and suggest a late acquisition – in absolute times – of the mitochondria followed by a rapid diversification of eukaryotes. However, their analyses cannot discriminate between mito-early or -late hypotheses which are relative to the origination of other eukaryotic features (e.g., endomembrane system) (Betts et al., 2018).

In addition, comparative genomics can provide insights about which genes were present in the archaeal ancestor of eukaryotes and the LECA. However, these approaches often lack an evolutionary framework, which could result in parsimonious but inaccurate inferences. Ancestral reconstructions methods, which take into account the pattern of diversification of species and the evolutionary dynamics of genomes (or genes inside them), have the potential of generating accurate results if the evolutionary models used are realistic (see "Gene content of ancestral lineages" and "Paper III"). Yet, the information that existing methods can provide about the intermediate states between the archaeal ancestor of eukaryotes and LECA is very limited and therefore, little is known about what happened during that period. In this respect, a recent study has attempted to shed some light into the relative timing of the mitochondria and the nature of the host cell. Their results suggest that the acquisition of mitochondria occurred relatively late during eukaryogenesis by a host that already contained many genes of bacterial and archaeal descent (Pittis and Gabaldón, 2016). However, the methodology used by the authors is currently debated (Martin et al., 2017; Pittis and Gabaldon, 2016) and new analyses are needed to confirm or deny such results.

The identity and nature of the archaeal ancestor

Defining the identity and capabilities of the prokaryotic ancestors of eukaryotes can help to refine the hypotheses on eukaryogenesis by setting realistic assumptions. Our understanding of the identity of the archaeal host has been changing as we uncover more archaeal groups. Initially, it was suggested that members of the TACK superphylum were the closest living descendants of the archaeal host (Cox et al., 2008; Foster et al., 2009b; Guy and Ettema, 2011; Guy et al., 2014; Kelly et al., 2011; Lasek-Nesselquist and Gogarten, 2013; Raymann et al., 2015; Williams and Embley, 2014; Williams et al., 2012). Nevertheless, the exact placement within this superphylum was unclear. While most analyses could not confidentially pinpoint an exact placement within this superphylum, various pointed to an archaeal ancestor affiliated with Korarchaeota (Guy and Ettema, 2011; Guy et al., 2014; Kelly et al., 2011; Williams and Embley, 2014; Williams et al., 2012). However, these analyses could not exclude the possibility that the observed Eukaryota-Korarchaeota affiliation was an artefact arising from the presence of a single and deeply branching Korarchaeota representative (Guy et al., 2014). Another explanation for such placement was that eukaryotes were affiliated to other groups distantly related to Korarchaeota that lacked sequenced relatives, such as the Deep Sea Archaeal Group (DSAG), Marine Hydrothermal Vent Group (MHVG), and Ancient Archaeal Group (AAG) (Guy and Ettema, 2011; Guy et al., 2014).

The discovery of the first genome belonging to the DSAG group (renamed as Lokiarchaeota after the sampling location from which this lineage was retrieved, Loki's Castle) has provided additional clues about the identity of the archaeal ancestor (Spang et al., 2015). Phylogenetic analyses including Lokiarchaeota – originally considered a deeply branching clade of the TACK superphylum – show a monophyletic relationship between eukaryotes and Lokiarchaeota. This affiliation is further supported by the presence of a large number ESPs in its genome, some of which have been previously identified in various archaea albeit with patchy taxonomical distributions. Interestingly, the *Lokiarchaeum* genome also encodes for homologous of ESPs that had never been observed in prokaryotes before. Although a recent study has questioned the quality of this genome due to its metagenomic origin and argued against the Eukaryota-Lokiarchaeota affiliation (Cunha et al., 2017), such re-analyses and interpretations have been themselves criticized and rebutted (Spang et al., 2018).

The genomic capabilities of *Lokiarchaeum*, whose genome encodes for several homologs of genes that are required for key cellular processes in eukaryotes, support a scenario in which the archaeal ancestor of eukaryotes was relatively complex. The archaeal host is thought to harbour homologs of eukaryotic components involved in replication, transcription and translation machineries, as well as the proteasome, exosome, and ubiquitin modifier systems (Gribaldo and Brochier-Armanet, 2006; Koonin, 2015; Koonin and Yutin, 2014). Furthermore, the additional ESPs identified in Lokiarchaeota suggest that the ancestor also contained homologs of genes comprising the eukaryotic cytoskeleton (e.g., actin and actin regulators, such as gelsolin and profilin), as well as, and various genes involved in eukaryotic membrane remodeling and trafficking (e.g., components of the endosomal sorting complexes required for transport (ESCRT) and numerous small GTPases) (Klinger et al., 2016; Spang et al., 2015). Although the biological function of such proteins in *Lokiarchaeum* remains unknown, it is likely that at least some of them perform functions equivalent or related to their eukaryotic counterparts (Akil and Robinson, 2018). Yet, culturing and experimental efforts are required to be able to understand the role of these proteins *in vivo* and the general cell biology and metabolism of uncultured microorganisms such as *Lokiarchaeum*. This information will be crucial for refining our understanding of the eukaryotic evolution.

Genomic exploration of archaea

Traditional methods

Traditionally, the study of archaea and other microbes required their isolation and cultivation in a laboratory. Once in culture, these microbes were often characterized through growth studies, biochemical profiling and microscopy. With current technologies, it is now possible to also study their genomes, transcriptomes, proteomes and metabolites. Altogether, we can obtain detailed information about both the genotypes and phenotypes of organisms growing in culture. Nevertheless, it is important to keep in mind that functional characterizations performed under artificial laboratory conditions do not necessarily reflect the behaviours of microbes in their natural environment. Our current understanding of cultured microorganisms is thus somewhat biased, and interesting physiologies and characteristics have probably been overlooked.

Unfortunately, most microbial groups lack cultured representatives that we can investigate using these culture-dependent techniques. A recent study estimates that 81-98% of microbial cells on Earth belong to genera or higher taxonomic ranks without cultured representatives (Lloyd et al., 2018). These high numbers reflect the intrinsic difficulty of isolating and growing microorganisms in culture. Since growth conditions and nutritional requirements are unknown at first, culturing new isolates becomes an iterative and time-consuming process, which is usually carried out manually. Complicating culturing efforts further, some microbes are obligate syntrophs, extreme oligotrophs, slow growers or require conditions that are difficult to maintain in the laboratory, preventing them from being grown in pure culture (Lloyd et al., 2018). Hence, to understand the diversity and physiologies of most microbial life, culture-independent approaches are required.

Culture-independent approaches

Since the development of environmental SSU rRNA gene sequencing approaches, SSU rRNA surveys have been widely used for taxonomic identification and abundance estimation of microbes (Doolittle, 1999; Hou et al., 2013; Jorgensen et al., 2012; Pace et al., 1986; Sogin et al., 2006;

Turnbaugh et al., 2007). The mainstream version of this approach takes advantage of the architecture of the SSU rRNA gene – which contains alternating conserved and variable regions – to generate PCR amplified products that are sequenced in a high-throughput manner. Although the reads recovered are usually short, representing just a small part of the gene, this is generally sufficient to get an overall idea of the identity and abundance of the microorganisms living in an environment.

However, SSU rRNA gene surveys have several limitations (Bonk et al., 2018; von Wintzingerode et al., 1997). Most importantly, the PCR step introduces amplification bias towards studied microorganisms (von Wintzingerode et al., 1997). Since primers are designed based on sequences of known genes, they can fail to hybridize and amplify atypical sequences and, thus, organisms encoding such divergent genes can go undetected (Eloe-Fadrosh et al., 2016). Secondly, chimeric molecules can be generated during PCR amplification, resulting in sequences that do not belong to any existing organism (von Wintzingerode et al., 1997). Furthermore, given that SSU rRNA genes can be present in a variable copy number, abundance estimates of community members are often biased (Farrelly et al., 1995). Lastly, the phylogenetic signal contained in the short sequenced fragments is insufficient to resolve the phylogenetic placement for many of these organisms.

To overcome some of these disadvantages, variants of this technique have been developed. They include the use of different phylogenetic markers (such as the long subunit ribosomal RNA), sequencing full SSU rRNA genes or several genes simultaneously (Karst et al., 2018; Martijn et al., 2019) and versions without primer biases (Karst et al., 2018) among others.

Although convenient for getting an idea of the microbial community in an environmental sample, SSU rRNA gene approaches are not suitable for understanding the genomic potential of uncultured microorganisms. Instead, single-cell genomics (Lasken, 2013; Stepanauskas, 2012) and metagenomics (Tyson et al., 2004; Venter et al., 2004) can be used to study the genomes of organisms without the need for culturing. Both techniques are based on the same idea: sequencing DNA extracted directly from an environmental sample. However, while single-cell approaches rely on capturing and isolating individual cells before sequencing, metagenomic techniques sequence the DNA of all microorganisms at once. When the aim of a metagenomic study is reconstructing the genomes of microorganisms present in a sample, the term genome-centric metagenomics is used. Alternatively, we refer to gene-centric metagenomics if the objective is to analyse the genes and functions of a community as a whole. Both single-cell genomic and genome-centric metagenomic techniques can produce genomes of comparable accuracy (Alneberg et al., 2018). In addition, other meta-omics approaches can be used to study gene expression (metatranscriptomics), protein content (metaproteomics) and, to a lesser extent, metabolites (meta-

metabolomics) of microbial communities (Simon and Daniel, 2011; Tang, 2011).

Genome-centric metagenomics

The first steps in every metagenomic workflow are: 1) obtaining a sample from an environment of interest, 2) extracting DNA from it, and 3) sequencing (Figure 2). Depending on the sequencing platform, short and accurate or long and error-prone reads will be obtained. Former metagenomics approaches required the construction of plasmid or fosmid libraries, followed by Sanger or another type of shotgun sequencing (Daniel, 2005; Kunin et al., 2008). However, such approaches are rarely in use today, and will not be covered here.

In genome-centric metagenomics, reads are subsequently assembled into longer contiguous sequences (contigs) that represent genomic fragments of the microbes present in the sample. These contigs are then classified according to the organism they were originated from in a process referred to as '*binning*', which is commonly followed by a refinement step to ensure the accuracy of the classification. The end of this process will result in complete or, more commonly, partial genomes: the so-called (genome) bins or metagenome-assembled genomes (MAGs).

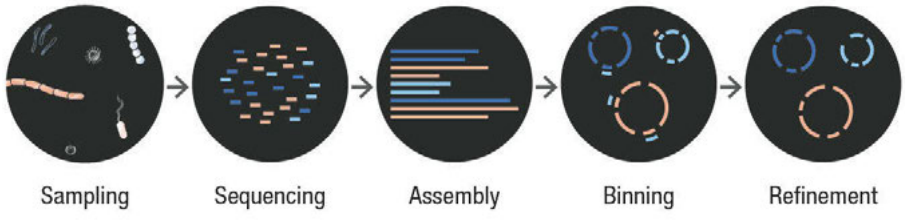

Figure 2. Overview of the standard workflow used in genome-centric metagenomics.

In the last years, the field of genome-centric metagenomics has changed substantially. Numerous tools have been developed and improved within a short period of time, and standards are now established for short-read based metagenomics. Furthermore, third generation sequencing technologies have recently erupted in this field and quick progress is expected to happen in the coming few years.

In the next sections, I will give explain in more details the different steps of the metagenomic workflow with considerations for both short- and long-read metagenomics. In particular, I will highlight the relevant steps needed to reconstruct the genome of specific target organisms from environmental samples comprised of complex communities, such as sediments.

Sample selection

Ideally, an assessment of the complexity of the microbial community in a sample should be performed prior to metagenome sequencing. The complexity of the sample depends on the number of species in it and their relative abundances. Samples with more species that are present in similar proportions are more complex than those with fewer species in uneven abundances (Kunin et al., 2008). Some trends in sample types can be observed in which sediments and soils are usually among the most complex communities (Torsvik et al., 2002). In general, downstream bioinformatics analyses of low complexity samples will be more straightforward and result in more contiguous and complete genomes.

Due to their relatively low price, SSU rRNA gene surveys are commonly used to assess the community composition of samples, and to identify those most suitable for further metagenome sequencing. When the aim is studying certain species rather than the whole population, the ideal sample would be a simple community in which the microorganism of interest is present in high abundance but in which closely related organisms are absent. Such characteristics give the best prognosis for the recovery of high-quality genomes in subsequent assembly and binning steps (see sections below).

The community composition of samples can be modified through additional experimental procedures. For example, size filtering (Castelle et al., 2015) or culture-based enrichment (Park et al., 2014) can reduce the sample complexity and increase the relative abundance of the target microorganisms. Restricting the sample collection to a homogenous and precise location might limit the presence of related strains within the population (Kunin et al., 2008). However, if the species of interest are rare, the biomass of the sample is insufficient, or if enrichment and filtering procedures are not successful, suboptimal samples become the best available option. To ensure the recovery of low abundant microorganisms in such cases, high sequencing depth is often required.

Furthermore, sequencing several related samples in which organisms co-occur at different abundances might be advantageous in genome-centric metagenomics projects, as they aid the classification of contigs into genome bins (see “Genome binning”). Such samples can be obtained by, for example, using different DNA extraction methods or by sampling either at different time-points or neighbouring locations (Albertsen et al., 2013; Alneberg et al., 2014).

DNA extraction

Once the presence of the target organism in a sample has been verified, it is equally important to ensure that the cells are lysed and the DNA accessible. Not all microbial cells are equally easy to lyse. Lysis susceptibility can vary among microorganisms depending on the composition of their cell wall and the extracellular matrix of biofilms. Failure to lyse certain cells will result in variations in DNA extraction efficiencies between microorganisms, introducing a bias in the relative DNA abundances of community members (Frostegård et al., 1999; Jiang et al., 2011).

There is no DNA extraction method that is suitable for all organisms and all environments. Protocols, therefore, need to be optimized to the sample or microbe of interest by selecting appropriate lysis methods, which could include mechanical force, temperature, sonication, chemicals or enzymatic digestion. Subtle variations in protocols can lead to important differences when it comes to the observed microbial composition (Albertsen et al., 2015). For example, methods that use physical force such as bead beating can help to extract DNA from hard-to-lyse microbes, and have been shown to increase the extraction efficiency of archaea and some bacteria (Albertsen et al., 2015; Salonen et al., 2010).

The issue with aggressive DNA extraction methods (such as bead beating) is that they also cause DNA shearing and can be problematic in recovering high-molecular weight (HMW) genomic DNA necessary for long-reads sequencing. For instance, the distribution of read lengths obtained with long-read Nanopore sequencing seems to be dependant on the quality of DNA after library preparation rather than on the sequencing chemistry itself (Branton and Deamer, 2019). Since long reads can span repetitive regions aiding in the assembly of sequences that would otherwise be problematic reconstruct, being able to extract high quality HMW DNA can be crucial to obtain complete genomes (Branton and Deamer, 2019). Therefore, it becomes essential to optimize protocols for long-read metagenomics that allow the lysis of most microorganisms present in a sample while, at the same time, maximize the quality of the HMW DNA. However, given that the field of long-read metagenomics is still in its infancy, the conditions required for ensuring good results for different types of environmental samples are still under evaluation.

Metagenome sequencing

High throughput DNA sequencing can be done using different technologies, with the Illumina sequencing platform currently being the most used for genome-centric metagenomics. This technology allows for the generation of hundreds of millions of short DNA sequencing reads that have a very low

error rate (lower than 0.1%) (Liu et al., 2012). The high quality of the generated reads together with the low cost per sequenced base is what has made this sequencing technology a very attractive choice for metagenomic studies. In this respect, the reasonable price makes deep sequencing affordable, and thus allows for the identification of low abundant members of microbial communities. On the other side, the short length associated with Illumina reads – ranging from 50 to 300bp long – is considered the main disadvantage of this sequencing platform. This is particularly problematic for genomic and metagenomic studies in which the short read length complicates the assembly process hampering the reconstruction of complete genomes (see sections below).

Alternatively, third-generation sequencing platforms, such as Pacific Biosciences (PacBio) and Oxford Nanopore, can produce long DNA sequencing reads. These platforms have been widely used in sequencing projects, allowing for the completion of numerous genomes (Loman et al., 2015; Rhoads and Au, 2015). However, the relatively low throughput and high cost of these technologies have limited their use in the metagenomic field. The development of the Oxford Nanopore PromethION sequencer, which can produce up to several hundreds gigabases of long reads in real-time, has supposed an inflexion point for the use of long reads in other applications. Albeit still limited, the field of long-read metagenomics is rapidly growing and early results already show the benefit of having long reads to obtain complete genomes directly from metagenomic samples (Bertrand et al., 2019; Nicholls et al., 2019; Somerville et al., 2019; Warwick-Dugdale et al., 2019). Nevertheless, third-generation sequencing technologies still have important disadvantages, particularly concerning their high error rates. Despite being continuously improving, long-read error rates are still around 14% for PacBio and 15-20% for Nanopore (Jain et al., 2015; Weirather et al., 2017). To increase their accuracy, both PacBio and Nanopore technologies have protocols that can sequence the same read multiple times to generate consensus reads with decreased error rates, although at the expense of read length and throughput (Ip et al., 2015; Travers et al., 2010). Nevertheless, additional Illumina sequencing is often required to correct sequencing errors in order to produce high-quality genomes, thus increasing the costs per sample.

Other promising options are the reads produced by companies such as 10X Genomics, which allow for the reconstruction of artificially generated long reads with an error rate comparable to that of Illumina sequencing. Such reads can be extremely valuable for the reconstruction of complex eukaryotic genomes that contain many repeats and structural variants. Although their use is in metagenomics still limited (Bishara et al., 2018), such reads could also be promising for assembling genomes from complex metagenomic samples, especially for samples with high strain diversity.

Sequence assembly

Assembly is the process of creating contiguous stretches of sequences (contigs) by combining multiple sequencing reads (Kunin et al., 2008). From a theoretical perspective, having long reads lacking errors would allow for a relatively straightforward reconstruction of a genome. In practice, we rarely have access to such reads, at least not in a high-throughput manner. Currently available short reads contain few – but still some – errors, whereas long reads have high error rates that create additional challenges in the assembly process. From such a starting point, assembling one single genome can be an arduous problem to solve, which is compounded when assembling multiple genomes simultaneously, as is the case for metagenomes.

Overlap layout consensus (OLC) and de Bruijn graph (DBG) are two of the main strategies used by assemblers, for which numerous variants and implementations exist (Figure 3). Both methods are based on translating the problem of genome-sequence reconstruction into mathematical graph theory and implementing solutions for graph theory problems. In a graph, nodes represent the basic elements and connections between them are the edges. Usually, the basic elements (nodes) of an assembly graph represent reads or read fragments and the edges indicate overlaps between them. From such representation, contigs can be generated by traversing (walking) the assembly graph.

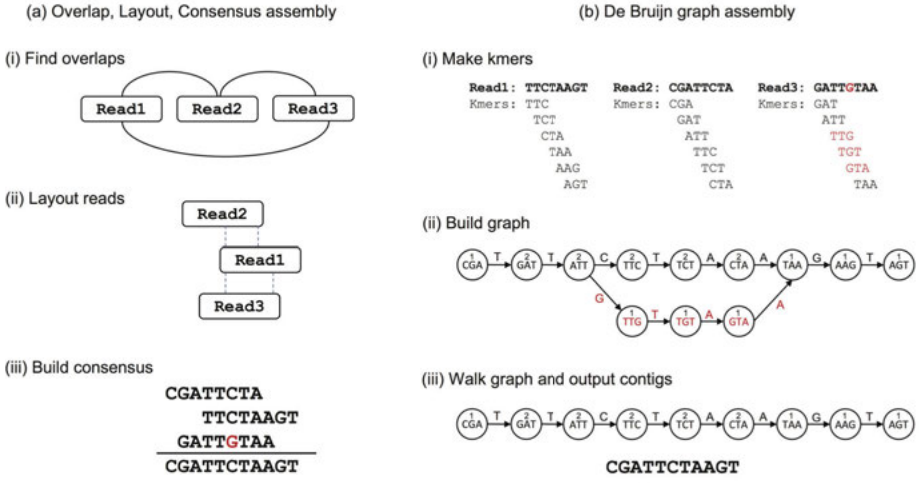

Figure 3. Schematic representation of two different assembly strategies: Overlap, Layout, Consensus (a) and de Bruijn graph (b). Polymorphisms or sequencing errors (red) form branching structures in DBG-based assembly graphs. Original figure by Ayling et al. (2019).

Overlap, Layout, Consensus

As its name indicates, OLC methods are based on three steps: overlap, layout and consensus (Miller et al., 2010). In the overlap stage, which is by far the most computationally demanding, every read is compared to every other read to identify overlaps between them. The assembly graph is then built using read sequences as nodes and overlaps between them as edges. The second phase, the layout, groups the overlaps previously generated to form contigs. Finally, a consensus sequence is determined by choosing the most represented nucleotide at each position in the layout.

Repeats contained within reads can be resolved by OLC approaches if the ends of the reads can be unambiguously overlapped and positioned during the layout step. Any repeat longer than the read will be unresolved. Thus, ultra-long reads are the most useful to solve repeats and have a huge impact on the contiguity and quality of the assembly.

OLC approaches were popular with Sanger reads and their use has re-emerged with long reads from PacBio and Nanopore sequencing technologies. In their new implementations, many of which combine elements from other assembly strategies (e.g., string graphs), overlap-based methods use heuristics to address the higher throughput of the current technologies. Furthermore, most overlap-based assemblers designed for third generation sequencing tackle the high error rates of the reads by including an initial pre-correction stage, in which reads are aligned to each other to generate more accurate consensus reads (Chin et al., 2013; Koren et al., 2017). However, the inclusion of an additional alignment step is computationally costly. This has motivated the development of alternative assembly tools that can use uncorrected reads directly to produce unrefined contigs that retain numerous errors (Li, 2016).

De Bruijn Graph

In DBG approaches, reads are split into overlapping subsequences of length k, called k-mers. Each different k-mer becomes a node in the assembly graph that can be connected to other k-mers if they overlap without mismatches in all but one of their bases (Miller et al., 2010). In other words, edges in the graph are formed by perfect overlaps of length k-1. Note that other variations of the assembly graph definition exist, although they are not mentioned here for simplicity. Once created, the graph is traversed guided by heuristics to generate contigs.

Although the graph construction is done very efficiently, navigating the graph in the correct order to reconstruct sequences corresponding to actual genomes can be daunting. This can be particularly challenging when sequencing errors, repetitive regions, heterozygosity, strain variation and structural variants are present in the sample, as they create complicated

branching structures that increase the complexity of the graph (Ayling et al., 2019; Olson et al., 2017). Discerning which of the many possible graph traversals is correct can be an impossible task without any further information. Therefore, assemblers usually incorporate additional data to create constraints that aid in the reconstruction of contigs, such as the alignment of reads back onto the graph, coverage information or graph connectivity. If such information is not enough to resolve ambiguities, the graph traversal breaks at such points, generating fragmented assemblies (Olson et al., 2017; Vollmers et al., 2017).

Unlike OLC approaches, sequencing errors heavily affect the graph construction by creating false k-mers and overlaps that increase memory requirements (i.e., more k-mers need to be stored) and add branches to the graph. Each sequencing error can affect up to *k* different k-mers and thus, their impact increases with the length of *k*. Therefore, most DBG assemblers include a step previous to the graph construction to detect and correct such errors. In single-genome assembly, with relatively even coverage, errors can be identified by detecting rare k-mers that have low multiplicity values (i.e., the number of times that a given k-mer appears). Erroneous k-mers are subsequently corrected by applying the minimum number of changes that can lead to a correct k-mer sequence. Nonetheless, such approaches are suboptimal for metagenomics datasets – which contain organisms present at various abundances – since they remove k-mers from low-abundant species. Hence, revised methods have been developed for metagenomic datasets that avoid the assumption of uniform coverage, for example, by removing rare k-mers only from reads with high coverage (Olson et al., 2017; Vollmers et al., 2017).

In addition to sequencing errors, repetitive regions also impact the structure of the graph by adding additional edges between nodes that increase the number of possible traversals (Olson et al., 2017). In this regard, the k-mer length plays an important role. The longer the k-mer, the lower the probability of finding overlaps of length k-1. Hence, longer k-mers increase the specificity and create fewer edges, leading to a better resolution of repeats. However, they also require higher sequencing depth to allow sufficient overlaps between nodes, and thus avoid unconnected graphs. On the contrary, short k-mers result in the creation of more edges and are therefore more suitable for shallow sequencing depths, albeit while having a limited power to resolve complicated repetitive structures (Vollmers et al., 2017). In order to obtain better assembly results, most currently used assemblers are able to incorporate the information from several k-mer lengths (Bankevich et al., 2012; Li et al., 2015; Nurk et al., 2017; Peng et al., 2012).

In single-genome assembly with even coverage, k-mers originating from repetitive regions can be identified as having higher multiplicity values. Those values can, at the same time, be used to navigate the graph and

resolve repetitive structures. Yet, in metagenomic datasets – with species present at different abundances – such coverage assumptions are violated, and additional information or specific algorithms are required to resolve repeats (Breitwieser et al., 2017; Ghurye et al., 2016; Vollmers et al., 2017).

DBG approaches became popular with the introduction of high throughput and low error sequencing technologies, as their efficiency and computational requirements do not scale with the depth of coverage and number of sequencing reads. Currently, new tools that combine overlap-based with approximate DBG approaches have been implemented that can make use of third generation sequencing reads (Kamath et al., 2017; Lin et al., 2016).

Assembling metagenomes

The uneven abundances of different members of a microbial community create additional challenges for metagenomic assembly. First, as mentioned above, simplistic graph constraints based on coverage cannot be used to identify errors and to resolve repeats. Second, without deep sequencing, low abundant members may be barely sequenced, and thus give rise to highly fragmented and incomplete genomes. On the contrary, high abundant species may exhibit an excessive depth of coverage dramatically increasing the computational costs in overlap-based assembly methods. Furthermore, such excessive coverage amplifies the effect of sequencing errors, leading to complex assembly graphs that ultimately produce fragmented assemblies in DBG approaches. An optimal depth of coverage of 50x has been shown to produce longer contigs in DBG assemblies based on short reads (Desai et al., 2013). Yet, in metagenomic samples, no single tradeoff value is achievable for all members of the community simultaneously.

Unlike microorganisms isolated in culture, microbial populations in nature are non-clonal, which increases the genomic diversity of metagenomics samples. Multiple closely related lineages and strains are often found co-inhabiting the same niche. The presence of closely related lineages (i.e., microdiversity) or other organisms that share similar genomic regions (e.g., horizontally transferred sequences, highly conserved regions) can hamper the assembly process (Awad et al., 2017). Similar regions shared between different genomes behave like repeats in the assembly process, generating complex branching structures in the graph that often cannot be resolved, and thus result in fragmented assemblies (Olson et al., 2017). More importantly, if such graph-structures are incorrectly traversed, co-assembly of different genomes can occur producing chimeric contigs. Co-assembly is more likely to happen in the presence of multiple highly similar strains. However, depending on the application, obtaining longer consensus sequences for a species, possibly mixing related strains, might be desired

over reconstructing microvariation within highly similar strains in fragmented assemblies.

Finally, the assembly of complex communities is a computationally costly process. The large number of lineages and strains in such samples requires the storage and processing of massive amounts of data. Although new assembly implementations have considerably reduced the computational resources needed, the extensive amount of sequencing data required to assemble genomes from complex environmental samples is still a limiting factor for both overlap- and DBG-based approaches. Therefore, to avoid practical limitations, planning for bioinformatics resources and costs becomes as important as having the experimental data.

Given these additional difficulties, the quality of metagenome assembly is generally inferior to the assembly of clonal isolates. Typically, this is translated in the generation of shorter contigs and the failure to obtain complete genomes. Furthermore, as any other sequence assembly process, errors (i.e., misassemblies) could be present in genome sequence reconstructions and should be kept into consideration. Yet, current tools for short-read metagenomics use algorithms optimized for the metagenomic nature of the data to limit such errors and improve the assembly quality (Li et al., 2015; Nurk et al., 2017). While several assemblers have shown to be accurate for single-genome assembly of long reads, it is currently unclear how such methods perform with metagenomic data. Tools specialized for long-read metagenomics are currently in active development, and, while promising, their accuracy still needs to be evaluated.

Scaffolding

Contigs generated during the assembly process can be subsequently ordered and oriented to form longer genomic fragments called scaffolds. In contrast to contigs, scaffolds are non-contiguous and contain unknown regions between contigs formed by gaps (represented by stretches of Ns). Commonly, the scaffolding process makes use of the information given from pairs of reads generated from the two ends of the same DNA fragment (i.e., paired-end reads) to infer which contigs are consecutive (Ghurye and Pop, 2019). Paired reads that align to two different contig-ends suggest a link between those contigs. If the number of paired-reads connections between two contigs is high enough, contigs can be combined into scaffolds, in which the inter-contig gap size can be estimated from the insert size of the sequencing library (i.e., the size of the DNA fragment excluding the size of the sequencing library adapters). Libraries with long insert sizes can thus improve the scaffolding step considerably (Kunin et al., 2008). Other methods for scaffolding based on optical maps (generated by using restriction enzymes), long sequencing reads or Hi-C data (based on

chromosome conformation capture) also exist, although their application to metagenomics is still limited (Ghurye and Pop, 2019). However, scaffolding is a difficult process that is subject to errors (i.e., combining contigs erroneously), especially when assembling large and complex metagenomes. Consequently, even though many assembly tools can also perform the scaffolding step based on paired-end reads, when short reads are the only available data, the use of contigs over scaffolds in metagenomic analyses might be preferred to minimize potential artefacts at the expense of having shorter sequences.

Assembly validation

Assessing the accuracy of genomic reconstructions can be relatively straightforward when the expected output is known. Metagenomic assemblers are generally evaluated by using mock communities or synthetic datasets in which the number of lineages and their abundances is pre-defined (Nicholls et al., 2018; Sczyrba et al., 2017). However, such data can fail to reflect the complexity of real environmental samples and other biases related to the sample preparation or sequencing process. Since the performance of assemblers heavily depends on the complexity of the underlying data, the predicted accuracy for a given tool can show large variations from sample to sample and among genomes within a given sample.

In the absence of reference genomes, assembly validation is less simple. Yet, the simultaneous use of several statistics can give an idea of the accuracy of the resulting genome reconstruction (Vollmers et al., 2017). Some of these measurements aim to assess the contiguity (i.e., the length and fragmentation) of the assembly. The most popular is the N50 value, that represents the minimum contig length of the set of contigs that comprises over half of the total assembly size. To be a useful measure, the compared assemblies should have the same minimum contig length cut-off (note that different assemblers use different cut-off values). Additionally, the N50 value should be inspected together with the total assembly size to avoid misinterpretations. Checking the number of contigs and the contig length distribution (maximum, mean, median and quartiles) can also provide some insights about fragmentation in the assembly. However, such statistics do not inform about the accuracy of the (meta)genome reconstruction itself and can be misleading when used in isolation (Ayling et al., 2019). For example, assemblies could show low fragmentation and high contiguity but still contain numerous misassemblies (e.g., chimeras).

Measuring accuracy is generally more complicated. In single-genome assembly, potential misassemblies can be detected by aligning the sequencing reads to the reconstructed contigs and inspecting both the depth of coverage and the alignment pattern of reads (or read-pairs) within contigs.

Sudden drops in coverage are typically found at breakpoints flanking inversions, deletions and insertions, and regions with unexpected high coverage could indicate collapsed repeats. Other indications of potential misassemblies could be given by single reads mapping to two non-contiguous regions in the middle of a contig or by read-pairs that align with exceedingly large distances between them or in the wrong orientation (Olson et al., 2017). However, in metagenomic assemblies, such arrangements are not always easy to interpret, especially when closely related lineages are present in a sample. Reads originating from different lineages can behave as repeats and align to multiple regions producing variations in the coverage pattern. Discerning whether such variations are the result of assembly artefacts or biological differences between organisms can be difficult. Still, some errors can be identified by detecting bases that are not supported by any reads (i.e., regions of zero coverage). Nonetheless, even such cases can be complicated to assess when assembling low abundant community members in which the read depth is often minimal. The evaluation of such regions can be further complicated by, again, highly similar lineages that are prone to co-assemble, as well as, the presence of sequencing errors.

Genome binning

Once contigs (or scaffolds) have been reconstructed, the next step is to identify which of them originated from the same organisms and to group them together to produce MAGs. The simplest binning methods consist of similarity or k-mer based searches of contig sequences against databases of reference genomes (supervised binning). However, since representative genomes are generally lacking for most microorganisms, supervised approaches are often ineffective. Alternatively, two properties of contig sequences are mainly used to classify them into MAGs: nucleotide composition and read coverage.

Composition-based binning is centred on the observation that oligonucleotide frequencies are species-specific and conserved across genomes within species (Dick et al., 2009). This compositional signature of genomes is observed for oligonucleotides of length two or higher (Abe et al., 2003; Nakashima et al., 1998; Noble et al., 1998; Pride et al., 2003; Sandberg et al., 2003) and its specificity increases with the oligonucleotide length (Bohlin et al., 2008). While longer oligomers could potentially allow for a higher binning resolution, the fact that the number of possible oligonucleotide sequences grows exponentially with their length creates computational constraints that limit their use. A good trade-off between specificity and computational feasibility is the use of oligomers of length four (i.e., tetranucleotides), for which 256 (4^4) different oligomers exist. Although tetranucleotides frequencies work generally well for classifying

contig sequences from different species, they have little power to separate closely related organisms that exhibit similar compositional signatures.

Furthermore, there are regions within genomes that deviate from the overall oligonucleotide composition of a given organism. Ribosomal RNAs, sequences of different origin (i.e., viral or horizontally acquired), and plasmids often display biases in compositional signatures (Noble et al., 1998; Pride et al., 2003). If such regions are present in long contigs, the oligonucleotide pattern is overpowered by the unbiased part of the sequence, and they thus have a minor impact on the classification. Conversely, short contigs containing biased regions are often misclassified by composition-based binning methods. In fact, current binning tools often group contigs encoding rRNA gene sequences from several organisms together into a single artefactual bin. As a rule of thumb, the longer the contig, the more robust the compositional signatures, and the more reliable the classification.

Coverage-based binning is based on the premise that contigs generated from the same genome should exhibit similar depth of coverage values on average. However, since organisms can display similar abundances in a population, coverage information alone is usually not enough to classify contig sequences. Instead, coverage information is used in conjunction with nucleotide composition data. Binning accuracy can be further improved by the use of read coverage information across several related samples. The idea behind this approach – referred to as differential coverage binning – is that organisms present in similar abundances in a sample can show uneven abundances in another, thus allowing for a higher accuracy in the classification (Albertsen et al., 2013; Alneberg et al., 2014). Therefore, the larger the number of related samples, the higher the probabilities of distinguishing contigs originated from different organisms. Ideally, the samples included should originate from microbial communities that largely overlap in terms of the identity of the microorganisms present but that exhibit considerable variation in their abundances. Samples collected at different time points, neighbouring locations, or that have been generated using different DNA extraction methods that lyse some species preferentially over others, can be used for differential coverage binning.

Over the last few years, numerous binning tools have been developed (Alneberg et al., 2014; Kang et al., 2019; Lin and Liao, 2016; Wu et al., 2016). The most recent ones use both composition and coverage information to generate MAGs. Some of them also include additional information to improve their classification; such as linkage information of contigs given by pair-end reads (Lu et al., 2017). Current binning tools mostly differ in the distance metrics and clustering methods they use, which lead to variations in their results. The performance of these tools depends on the structure and complexity of the sample with no single binning method outperforming the others in all scenarios (Sieber et al., 2018). To overcome such situation and improve the accuracy of binning, methods to combine the results from

various binning tools have been developed (Sieber et al., 2018; Song and Thomas, 2017; Uritskiy et al., 2018). Yet, such approaches rely on the accuracy of existing binning methods and cannot guarantee the generation of error-free MAGs.

MAG validation

Metagenomic binning requires careful validation and verification to ensure high quality. MAGs generated from binning tools should be inspected to identify misclassified contig sequences and other binning artefacts (e.g., co-binning of several lineages, partial MAGs). During the last years, the use of tools to estimate the completeness and contamination percentages of MAGs has become popular for evaluating their quality (Bowers et al., 2017). Such estimations are based on the percentage of universal single-copy genes found in each MAG. For example, if 80 of 100 marker genes are found in a genome, completeness is estimated to be roughly 80% (note that a correction for the co-occurrence of neighbouring markers is commonly included). Single copy markers found in several copies may indicate that contigs originating from different genomes were clustered together (i.e., contamination or redundancy). As mentioned above, closely related strains can be difficult to separate and often co-occur together in the same genome bin. In these cases, the estimated contamination values will be high, in spite of such values not indicating contamination from unrelated species. To distinguish between actual contamination and the presence of related strains, some tools, such as CheckM, use an additional measurement, known as strain heterogeneity. Strain heterogeneity is estimated by comparing the sequence similarity between markers found in more than one copy. If their amino acid identity is over a certain threshold (90% by default), CheckM considers that those markers belong to related strains. The heterogeneity value is based on the percentage of duplicated markers that pass this identity threshold.

However, these estimates can easily deviate from the actual completeness and contamination values of MAGs. First, a given set of marker genes might not be suitable for certain lineages that could lack some of the marker genes considered universal, or could, conversely, have duplicated copies of others. If this were the case, even a complete genome would have suboptimal contamination and completeness values. Second, these estimations are restricted to the presence of contigs encoding a limited number of genes. This is particularly important to consider in fragmented assemblies where contigs without any marker gene can be the majority.

Hence, it is important to not blindly rely on such tools and to use them in combination with other approaches that are based on the characteristics of the contig sequences. The distribution of GC content, tetranucleotide

frequencies or coverage values of contig sequences can be used to detect outliers that might represent misclassified contigs (Karst et al., 2018; Parks et al., 2015; Parks et al., 2017). Another source of information can be provided by the taxonomical classification of genes within contigs. A limitation of this approach comes from the microbial representation in sequence databases, where novel organisms lack relevant representatives. The information provided by read pairs connecting contigs and the presence of rRNAs and tRNAs is also useful in quality validation of MAGs, together with other statistics typically used for assembly assessment (Bowers et al., 2017; Karst et al., 2016). It is important to notice that for individual contigs, the amount of supporting evidence for their presence in a given MAG will vary.

Inferring evolution

Evolutionary history of species

One aspect of studying microorganisms is the determination of the evolutionary relationships between them to understand how they came to be. In this context, however, it is important to distinguish between organismal and genome evolution. On the one side, we can consider prokaryotes as populations of cellular entities that propagate through cell division and are related by a series of bifurcations in a tree-like manner. On the other side, we can think in terms of genomes evolving within species. Before cell division, DNA replicates to ensure that each daughter cell receives a copy of the genome. Throughout the cell cycle, different types of mutational events can occur generating variation in the genome. Parts of the genome can be modified (i.e., nucleotide substitutions), rearranged, duplicated or lost. More importantly, foreign DNA can also be incorporated into the genome by a variety of mechanisms such as transformation, transduction and conjugation (Lerat et al., 2005; Soucy et al., 2015; Wagner et al., 2017). In Archaea and Bacteria, HGT is considered one of the main evolutionary forces to generate genetic variation (Lerat et al., 2005; Wagner et al., 2017). The integration of new genetic material, which can originate from closely or distantly related species, allows the acquisition of new functions that might be instrumental to adapt to new or changing niches. Therefore, the evolution of genomes cannot only be considered in terms of vertical inheritance, but it also involves horizontal transfers that are better described in a network-like fashion (Soucy et al., 2015). Genome evolution is often regarded as the evolution of the entirety of genes encoded into them. Although genomes are not only composed by genes, such simplification bypasses many computational challenges (Boussau and Daubin, 2010) and it is more suitable for the study of distantly related sequences by using protein sequences instead of DNA.

And yet, although they are not one and the same, solving the evolutionary history of prokaryotes is not possible without studying the evolutionary history of (part of) their genomes. Molecular phylogenetic approaches have been developed to model the evolutionary processes and infer patterns of gene and species diversification. During the early days of molecular phylogenetics, it was common to interpret the evolutionary history of certain universal gene families as if they fully represented the phylogeny of their respective species (e.g., SSU rRNA) (Woese et al., 1990b). Nowadays we

know that gene histories do not always mirror species evolution. In prokaryotes, discord between gene and species phylogenies can arise not only from methodological errors and artefacts but also, as mentioned above, due to gene-specific events (e.g., gene duplications, losses, transfers) (Boussau and Daubin, 2010; Som, 2015). Because of that, the evolutionary history of a single gene cannot be directly considered as the organismal phylogeny.

To overcome this problem, current methods – such as supermatrices or supertrees – combine the information of several orthologous gene families (i.e. genes that evolved by speciation) to infer the vertical component of evolution. Similar to the early approaches, these methods rely on the identification of a limited amount of genes that evolved through speciation and are unlikely to have experienced horizontal gene transfer or duplication events. In contrast to single-gene phylogenies, such approaches average the phylogenetic – and often conflicting – signal of a set of genes, thus buffering the effect that the inclusion of undetected non-orthologous sequences might have in the phylogenetic reconstructions. Albeit numerous studies suggest that such approaches are able to capture true species diversification patterns (Abby et al., 2012; Galtier, 2007; Galtier and Daubin, 2008; Szöllosi et al., 2012), the fact that variations in the datasets or phylogenetic reconstruction methods used can result in different topologies has made them somewhat controversial. Hence, a great effort is often put into understanding the source of conflicting topologies and to minimize errors that could affect the reconstructions (Philippe et al., 2011; Rodríguez-Ezpeleta et al., 2007a).

Other approaches that model the evolution of gene trees along a species trees have the potential of providing a direct explanation of the phylogenetic discord between genes and species histories. Such methods attempt to fit the gene phylogeny inside the species tree by invoking a series of events (duplications, horizontal transfers, etc.) that explain the conflict between the topologies. Approaches that reconcile gene trees and species trees are promising not only because of their intuitive interpretation but also because they can make use of additional sources of information that have been previously overlooked. First, all homologous gene families, independently of whether they have evolved solely through speciation or not, can be used. By including gene families containing paralogs (i.e. genes that evolved by duplication) and xenologs (i.e. homologs acquired from horizontal gene transfer events), the repertoire of genes and, hence, the amount of phylogenetic information, increases considerably compared to supermatrices and other methods that require a small set of orthologous genes with clear vertical inheritance patterns. Such broader set of homologous families, even if they contain non-orthologous sequences, harbours a strong signal for vertical inheritance that can be exploited if modelled properly (Boussau et al., 2013). Moreover, if we assume that HGT only occurs between

contemporary organisms, transfer events can aid to establish the relative order of speciation events (Davín et al., 2018).

Unfortunately, no realistic method for co-estimating gene and species trees that also considers the modelling of Duplication, Transfer and Loss (DTL) events is currently available, although simplified (e.g., only considering DT) and related methods (e.g., coalescent approaches developed for eukaryotic evolution) do exist (Akerborg et al., 2009; Boussau et al., 2013; Wen and Nakhleh, 2018). Instead, most of these approaches are used to improve the topology of gene trees (Szöllősi et al., 2015b), evaluate alternative hypothesis for the evolution of species (Abby et al., 2012), infer gene histories and the gene-content in ancestral lineages (Williams et al., 2017b) or date a given species tree (Chauve et al., 2017b; Davín et al., 2018). In addition to limitations specific to these methods, most reconciliation approaches require pre-computed gene and species phylogenies to distinguish between vertical and horizontal evolution. Therefore, reconciliation methods are themselves subject to the same reconstruction errors and artefacts as other phylogenetic approaches, becoming a circular problem.

Currently, there is no phylogenetic method that can accommodate all processes driving the evolution of organisms at different levels (sequence, genomes and populations). This is partially due to computational and model limitations but also due to the fact that such processes are complex and not fully understood yet. Further development of phylogenetic approaches may result in more realistic methods and models that are able to reconstruct accurate and robust phylogenetic histories. For the time being, additional measures have to be taken to ensure that the phylogenetic inferences obtained are reliable and not the result of systematic biases and other types of errors.

In the next sections, I will assume that the reader has a basic understanding of phylogenetic reconstruction methods and evolutionary models and only explain some of the key aspects that are relevant to this thesis work. For an in-depth explanation of phylogenetic methods, see for example (Felsenstein, 2004; Yang, 2014).

Supermatrix-based approaches

Molecular phylogenetic approaches make use of mathematical methods to infer the evolutionary past from the information stored in the DNA of extant species. The reconstruction of molecular phylogenies, independently of the type of data and analysis, requires the identification and alignment of homologous characters from different species and the estimation of a phylogenetic tree using specific models and methods. The accuracy of the reconstructions depends on the quality of the dataset (e.g., the accuracy of the alignment) and the realism of the model used to infer the past.

In supermatrix-based analyses, the general approach remains the same, with two particularities. First, the characters used in the dataset (nucleotides or amino acids) originate from the concatenation of various genes. Second, gene sequences included should only be orthologous. The general workflow to infer species diversification patterns based on supermatrices of concatenated genes goes as follows (Roger et al., 2013):

1. A set of representative lineages, relevant to the question posed, is selected.
2. A set of single-copy orthologous genes (or proteins) present in all or most selected taxa is identified to represent the vertical inheritance of organisms. These genes are commonly referred to as marker genes.
3. Each orthologous gene family is aligned separately to identify homologous sites between the sequences, and filtered to minimize errors originated from automatic aligning tools.
4. Alignments for each marker gene are concatenated to create a supermatrix.
5. A model of sequence evolution that captures the dynamics of the substitution process of the given dataset is selected.
6. A phylogeny is inferred based on the data provided in the supermatrix under the selected model of sequence evolution. The most reliable phylogenetic methods currently available are based on maximum likelihood (ML) or Bayesian inference frameworks (not covered here, but for a general introduction see, for example, Holder and Lewis (2003); Roger et al. (2013)).
7. Finally, the validity of the results is assessed (statistical significance, model fit, etc.)

The use of phylogenetic reconstruction tools will always result in at least one phylogenetic tree. However, the obtained phylogeny does not necessarily represent the true evolutionary history of organisms, as errors can be introduced at various points of the workflow, leading to incorrect reconstructions (Philippe et al., 2017).

Errors and artefacts in phylogenetic reconstructions

There are three main types of errors that can affect the accuracy of phylogenetic reconstructions: 1) sampling or stochastic errors caused by insufficient number of phylogenetically informative positions; 2) errors arising from the violation of the orthology assumption and 3) systematic errors stemming from the inability of the substitution model to capture the underlying evolutionary process of the data (Philippe et al., 2011; Rodríguez-Ezpeleta et al., 2007b).

The first type of error is associated with having insufficient phylogenetic signal and is most problematic for evolutionary inferences of individual gene families in which the length of the gene is the limiting factor. The addition of more positions reduces the magnitude of sampling errors and, therefore, they generally do not represent an issue for supermatrix-based approaches that combine the information of several genes.

Failure to identify orthologous sequences and sites causes the second type of error, which can lead to unpredictable effects in tree inferences. In concatenations, this effect is averaged across sites from different genes reducing the magnitude of the error. Nevertheless, systematic failures in orthology detection can produce incorrect placements that are highly supported (Beiko et al., 2008).

Lastly, model misspecifications can lead to the consistent and systematic recovery of incorrect topologies that become statistically supported. The most common misspecifications stem from variations in the composition or rate of evolution of sequences that are generally not well captured by current substitution models. Such systematic errors may result in the artifactual grouping of sequences with similar characteristics regardless of their true evolutionary histories. Moreover, this effect is exacerbated when sites that have experienced multiple substitutions (mutational saturation) are present. In saturated sites, phylogenetic signal is overwritten by more recent substitutions. Similarities between saturated homologous sites are the result of convergent (homoplastic) mutations and do not carry any phylogenetic information. If substitution models were accurate, saturation would lead to random noise that would simply decrease the statistical support of the tree. However, in the presence of model violations, sequences that accumulate a higher number of substitutions – as the result of accelerated evolution or ancient divergence – are the most affected by systematic errors and can be incorrectly grouped together. Since these sequences are represented by long branches in the tree, this artefact is known as long-branch attraction (LBA) (Felsenstein, 1978).

Violations of the orthology assumption

Identification of orthologous genes

One of the most important assumptions made in supermatrix-based approaches is that the genes included in the analyses are universal (or nearly universal) for the taxa included, and have only evolved through speciation and therefore reflect the vertical evolution of the species. However, the identification of orthologous sequences is not always evident (Tekaia, 2016). Orthology is frequently inferred from sequence similarity searches that aim to identify single-copy genes. The hypothetical orthologous sequences are often subsequently scrutinised by inspecting the topology of the individual

gene phylogenies prior to concatenation of the final set of sequences. By doing so, it might be possible to detect and exclude sequences that are clear cases of horizontal gene transfer, paralogs or potential contamination in the dataset.

However, manual inspection of gene phylogenies cannot always guarantee that the remaining genes have only evolved through speciation (Boussau and Daubin, 2010). There are several reasons that can prevent the correct identification of non-orthologous genes. First, gene trees have limited phylogenetic signal that is often insufficient to accurately estimate the history of the gene family, especially at the deepest nodes. Second, the lack of duplicated sequences cannot be considered as absence of paralogy. Traces of duplications can be masked by incomplete datasets that exclude one of the copies or reciprocal losses in which organisms have retained different copies of an ancestrally duplicated gene (Doolittle, 1999). Although the frequency of this phenomenon – known as "hidden paralogy" – is unknown in prokaryotes, it has been shown that it can strongly affect phylogenies causing discord between them. Third, genes that have been horizontally acquired might not have a detectable pattern in gene phylogenies, especially in cases of ancient transfers mostly impacting the topology at the deepest nodes – the most difficult ones to reconstruct. Given the extent of HGT throughout the evolution of organisms, it is unrealistic to assume that certain gene families have never experienced it, independently of their function (Boussau and Daubin, 2010). In fact, a recent study suggests that all gene families present in the last archaeal common ancestor have experienced at least one horizontal transfer event during the evolution of this domain (Williams et al., 2017b).

Although simulation studies suggest that supermatrix-based approaches can reflect true species histories even in cases of substantial horizontal gene transfer events occurring randomly between organisms, they fail when HGT occurs preferentially between certain organisms (Beiko et al., 2008). This seems to take place in natural communities, where co-occurrence in the same habitat is an important facilitating factor (Smillie et al., 2011). In such cases, supermatrix-based approaches might produce well-supported topologies that do not reflect the true vertical history but instead show scenarios influenced by both vertical and horizontal signals that are difficult to interpret (Beiko et al., 2008). Failure to exclude such sequences in the dataset can thus result in inaccurate reconstructions.

Although obtaining a dataset that fully depicts the evolution of species might not be possible, reducing the number of non-orthologous sequences included in the analyses is critical to obtain meaningful results. Inclusion of a wider taxon sampling and genes from recently sequenced genomes increases the chances of detecting cases of paralogy and HGT by manual inspection of gene tree topologies. Additionally, the use of more than one set of orthologous genes (e.g. genes encoding ribosomal proteins vs. other

single-copy marker genes) to reconstruct phylogenies can help to either confirm that the resulting topologies are robust and indeed reflect the species history or to identify potential problems in the reconstructions.

Identification of homologous sites

Once orthologous genes are detected, homologous sites are inferred via multiple sequence alignment (MSA). An alignment represents a hypothesis of character homology, in which aligned characters are assumed to be derived from a common ancestor. Hence, non-homologous characters that are incorrectly aligned are treated as genuine historic signal in downstream analyses affecting phylogenetic reconstructions. The effect of errors in phylogenies can be minor when the phylogenetic signal in the data surpasses the noise created by misalignments and other sources of non-phylogenetic signal (explained below). However, alignment errors at high proportions and, especially, in cases in which there is no strong phylogenetic signal, can have bigger effects. In fact, it has been shown that different alignments of the same dataset can lead to tree estimates with different support values, branch lengths or even topologies, and have been associated to LBA artefacts (Blackburne and Whelan, 2013; Hossain et al., 2015; Ogden et al., 2006; Wong et al., 2008).

In spite of the efforts put into improving MSAs, there is no alignment software exempt of issues that can guarantee that the resulting alignment is the true one (Chatzou et al., 2016; Redeling and Suchard, 2009). Through heuristic implementations, alignment tools aim to obtain an alignment that is satisfactory according to a predetermined score that penalizes substitutions and gaps. Such a scoring system might be suboptimal for the data in question, resulting in the erroneous alignment of non-homologous characters. Not only can different tools result in alternative alignments: the order of the sequences in the input file (both left-right and up-down) may also generate variations in the MSA even when analysed with the same software (Boyce et al., 2015).

Given the importance of the accuracy of the alignment in the subsequent phylogenetic analyses, a common practice is to modify the MSA produced by tools to correct or filter unreliable sites. This step used to be performed manually. However, the increasing amount of data and the need for reproducible analyses have led to the development of specialized tools (Ali et al., 2019; Capella-Gutiérrez et al., 2009; Criscuolo and Gribaldo, 2010). Yet, detecting such erroneous sites is not trivial. Different tools make different assumptions about what an "erroneous alignment" should look like based on measures such as the conservation of characters across sites, the number of gaps or the stability of the alignment. Consequently, the sites identified as erroneous differ between tools. Filtering methods may fail to detect some non-homologous characters that are left in the data, creating noise. Additionally, they can remove true homologous sites, reducing the

phylogenetic signal. To which extent the reduction of noise is beneficial over the data removal is still debated, with some studies suggesting that filtering tools improve phylogenetic inferences (Jordan and Goldman, 2012; Karin et al., 2014; Privman et al., 2012; Talavera and Castresana, 2007) and others pointing toward detrimental effects (Spielman et al., 2014; Tan et al., 2015). To evaluate the impact of inaccurate alignments in phylogenetic reconstructions, it is advisable to compare the results from various alignments when sufficient computational resources are available.

Violations of the substitution model

Phylogenetic reconstruction methods are able to estimate evolutionary histories from sequence information found in extant organisms. By modelling the process of sequence substitution, these methods can estimate the evolutionary history that eventually resulted in the observed data. The accuracy of a substitution model resides in its ability to distinguish between similarities in homologous sites caused by homoplasy and conservation (phylogenetic signal) (Philippe et al., 2017). Homologous sites that have experienced few changes through time can appropriately reflect ancient similarities and are reliable to use in phylogenetic inferences. Sites that, on the contrary, have undergone multiple changes can erase the ancestral similarities and, hence, the phylogenetic signal. In these cases, similarities can be caused by reverse mutations (two or more substitutions in the same sequence character can revert it to an ancestral state) or convergence (two or more homologous characters can independently change to the same state). These similarities, if treated as ancestral, can bias phylogenetic estimations. Both ancestral similarities and homoplasies are found in datasets, together with incorrectly aligned non-homologous sites. As a result, the relative proportion of each type and the realism of the chosen substitution model are decisive factors to obtain accurate inferences (Gribaldo and Philippe, 2002). Simplistic models of evolution that underestimate the probability of convergence often result in artefactual reconstructions.

Substitution models

A model of sequence evolution (or substitution model) is a mathematical formalization that attempts to describe how characters change over time (Yang, 2014). They aim to capture the substitutions process and the subsequent natural selection that occurs in sequences during evolution. Substitution models are described as Markov processes in which the possible character types (i.e., the 20 amino acids or 4 nucleotides) are the states of a so-called Markov chain (Yang, 2014). The main property of Markov chains is that they have no memory, meaning that the change from one state to the next one only depends on the present state, independently of how the current state was reached (i.e., older states). Another main assumption made by

substitution models is that sites evolve independently. Different substitution models differ in which additional simplifications and constraints they set.

The core element of a substitution model is the rate matrix (Q matrix), which specifies the equilibrium frequencies of character states (i.e., nucleotides or amino acids) and the relative rates of replacement between each other (exchangeabilities) (Figure 4) (Whelan and Goldman, 2001; Yang, 2014). The number of possible character states determines the number of parameters and the dimensions of the Q matrix, being 4x4 for nucleotides and 20x20 for amino acids. Since most models assume time-reversibility – a property that defines the replacement rate between two amino-acids (or nucleotides) is the same regardless of the direction of the substitution – the number of parameters in the Q-matrix can be substantially reduced. For instance, if the rate of change from A to B is the same as the rate from B to A, only one rate is needed to explain both processes, thus decreasing the number of exchangeability rates by half. Under such assumption, and considering that the equilibrium frequencies and rates stay constant over time (stationary) and across all sites (homogeneity), the number of parameters of the amino acid rate matrix is 208: 189 exchangeability rates (20x19/2-1) and 19 equilibrium frequencies (Yang, 2014). For nucleotides this number is only 8: 5 rates (4x3/2-1) and 3 frequencies.

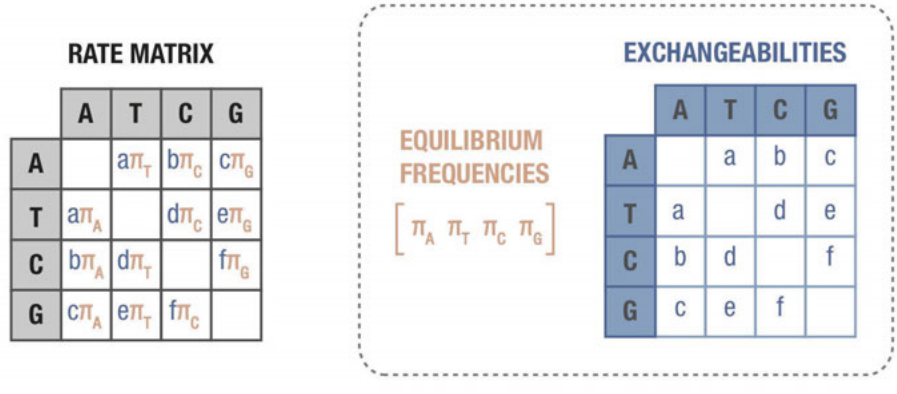

Figure 4. Representation of the rate matrix for nucleotide characters, highlighting the exchangeabilities (blue) and the equilibrium frequencies (pink).

The parameters of the substitution model can be either estimated from the data or pre-calculated from large empirical datasets (Yang, 2014). Estimating the parameters from the data has the benefit of improving the fit of the model at the expense of being more computationally demanding. A commonly used model where all 208 parameters are estimated under the assumptions of stationarity, homogeneity and time-reversibility is the so-called General Time Reversible (GTR) model (Yang, 1994a). Given the stationarity assumption, the 19 equilibrium frequencies can be set as the observed fre-

quencies and be easily obtained, but the remaining 189 are more difficult to estimate. Inferring such a large amount of parameters is computationally demanding and requires large datasets to accurately determine their values avoiding overfitting of the model, which is often not a problem for supermatrix analyses but can be inadequate for single protein phylogenies. Alternatively, one can use empirical models, in which the parameters have been pre-calculated from large protein matrices. These models receive different names depending on the protein matrix in which they are based, such as LG (Le and Gascuel, 2008) and WAG (Whelan and Goldman, 2001). A variation of empirical models is to use pre-calculated exchangeability values but replacing the equilibrium frequencies with the frequencies observed in the data (indicated by adding the suffix "+F", such as in LG+F). This alternative is generally a good balance between the fit of the model and the number of parameters to estimate (Yang, 2014).

The examples described above typify relatively simple substitution models that assume homogeneity and stationarity. For example, LG or GTR models used in isolation assume that the same Q matrix applies to all sites without accounting for compositional or rate variations across sites or taxa. Nonetheless, these assumptions, although mathematically practical, are not necessarily realistic. Sites often have different functional and structural constraints that make them evolve under different rates and selective pressures. Sites, restricted by their role and position in a protein, generally only accept certain amino acids, with the rest being selected against (Miyamoto and Fitch, 1996). Therefore, assuming that all amino acids are likely to be present in all sites with the same equilibrium frequencies is unreasonable. Similarly, organisms tend to differ in their evolutionary rates and sequence compositions (at nucleotides and amino acids level) (Collins et al., 1994; Foster and Hickey, 1999). The use of more complex models of sequence evolution can relax some of these assumptions.

Differences in rates of evolution across sites can be taken into account by assuming that substitution rates are not fixed but instead vary according to certain distribution, with the most common being the Gamma distribution (often specified by adding the suffix "+G", or "+Γ") (Yang, 1994b). Other corrections, such as the FreeRate heterogeneity distribution (suffix "+R"), do not conform to any pre-specified distribution (Soubrier et al., 2012; Yang, 1995) but instead estimate the differences between rates directly from the data. Using the gamma distribution has the advantage of only requiring one additional parameter to the model, while the number of parameters to be estimated increases in the FreeRate model. However, the latter approach has shown to better capture rate variations among sites (Soubrier et al., 2012).

Differential amino acid preferences across sites can be modelled by considering the presence of different classes of sites that can evolve under different Markov chains. In these models – known as mixture models –, sites belonging to the same class are described based on the same Q matrix. For

each class, a set of different equilibrium frequencies is estimated and combined with exchangeability rates that are globally defined. The equilibrium frequencies of the mixture model, as well as the number of classes to be used, can be either estimated from the data – as in the CAT model (Lartillot and Philippe, 2004) – or predefined (e.g. the C60 model uses 60 classes of fixed amino acid frequencies) (Quang et al., 2008).

By combining the models described above it is possible to account for rate and compositional heterogeneity across sites (e.g., CAT+GTR+G, LG+C60+R). Such complex models can detect more homoplasies due to site-specific constraints, and have been shown to fit the data better and alleviate LBA artefacts (Lartillot et al., 2007; Wang et al., 2008). However, phylogenetic reconstructions using these models are slow and often memory consuming. Other models that further relax the assumptions of stationarity and heterogeneity exist, albeit their use is limited either because their current implementations make their application computationally prohibitive for large datasets, or because they oversimplify the rest of the model (Blanquart and Lartillot, 2008; Foster et al., 2009a). When the use of more realistic models is not possible, model violations can be alleviated by optimizing the taxa representation in the dataset, filtering or transforming the data.

Alleviating model misspecifications

A diverse and even representation of organisms within the taxonomical group of interest can help to distinguish multiple substitutions, aiding the discrimination between phylogenetic signal and noise. From a theoretical point of view, the number of taxa selected should be high enough to give a good representation of the group in question, while keeping it computationally feasible (Roger et al., 2013). In practice, the taxa selection is limited to the number of organisms for which there are sequences available, often excluding or underrepresenting numerous clades. The inclusion of lineages that form long branches, either due to the absence of sequences from related lineages or to fast-evolving species, is discouraged since they are prone to suffer systematic errors that can cause LBA. Fast-evolving lineages can be substituted by close relatives that evolve slower, and other long branching taxa can be completely excluded if their presence is not relevant to answer the question posed. When the inclusion of highly divergent lineages is necessary, the use of complex models that account for compositional and rate variation across sites (e.g. CAT+GTR+G) can alleviate some of the systematic errors. Additionally, comparisons of the phylogenetic reconstructions obtained by including and excluding long-branching taxa can help to assess the robustness of the results and the identification of potential artefacts.

A related problem is the inclusion of compositionally biased taxa. As mentioned above, there is no efficient implementation of substitution models that can simultaneously deal with compositional heterogeneity across sites

and taxa for large datasets. Under models that assume that all lineages have the same composition, the probability of convergence between organisms with similar compositions may be underestimated. Therefore, excluding taxa with largely dissimilar amino-acid frequencies can alleviate the model violations. When the exclusion of taxa is not possible, the data can be filtered or transformed to balance the composition among lineages. Some approaches aim to filter out sites in the alignment with the most biased amino-acid compositions until the composition of the remaining residues is homogeneous. Such filtering can be performed by using, for example, χ^2 filtering (Viklund et al., 2012) or marginal homogeneity based stationary filters (Criscuolo and Gribaldo, 2010).

Another way of reducing compositional biases across taxa can be done by grouping residues that often substitute each other into a reduced number of possible character-states. Such data manipulation – known as data recoding – has shown to be effective in reducing compositional biases (Rodríguez-Ezpeleta et al., 2007b). The grouping scheme and the optimal number of categories can be either pre-calculated empirically according to the replacements of amino acids in large matrices (Dayhoff6, SR) or, preferably, estimated based on the given data (Hrdy et al., 2004; Susko and Roger, 2007). A drawback of data recoding is the reduction of the number of phylogenetically informative sites, which can be detrimental for the reconstruction (Hernandez and Ryan, 2019). This effect is exacerbated in lineages that diverged recently, in which recent substitutions can be completely masked by the categories hampering the resolution of such clades.

Furthermore, mutational saturation can magnify the effect that systematic errors have in evolutionary inferences. Mutational saturation is more problematic in deep phylogenies, in which the organisms included are distantly related and, therefore, their genomic sequences have been subject to numerous changes since they diverged from their common ancestor. Mutational saturation can be partially alleviated by using amino acid sequences instead of nucleotides. Amino acids substitutions are less frequent than nucleotides given the degeneracy of the genetic code. Furthermore, there are more amino acids than nucleotides states – 20 over 4 – allowing for a better detection of homoplasies. However, protein sequences can also become saturated when divergence times or evolutionary rates are very large. In such scenarios, the exclusion of sequences or sites that are prone to accumulate non-phylogenetic signals, such as fast-evolving sequences or sites (Philippe et al., 2011) and species lacking closely-related organisms can help to reduce artefacts. If the topology recovered is suspected to be influenced by the presence of saturated sites, their impact can be analysed by the progressive removal of the fastest-evolving sites in the data. Alternatively, data-recoding can also help to mitigate systematic errors in

evolutionary reconstructions that are caused by the presence of saturated sites.

Even though data transformation or filtering can help to ameliorate model misspecifications, such datasets can still be subject to model violations and systematic errors (e.g. changes on the substitution rate of a site through time). Hence, comparison of the results obtained from different strategies is crucial to assess the accuracy of the topologies and to spot additional biases that might be leading to incorrect reconstructions.

Gene content of ancestral lineages

The species phylogeny attempts to capture the vertical evolution of organisms, but it does not give complete information about how the genomes (and genes encoded within them) have evolved. Most genes within an organism will not follow the same evolutionary history as the species. Instead, each homologous gene family will evolve in a different manner, only sharing part of their evolutionary histories among them. Genes can be duplicated, lost, transferred from one organism to another or appear from *de novo* gene formation. By studying the evolutionary history of homologous gene families present in organisms we can infer how the gene content has evolved through time and infer which genes were present in ancestral lineages.

Current methods for ancestral reconstruction can model multiple types of evolutionary events acting on different levels, such as substitutions at the sequence level, duplications, losses and transfers at the gene level and speciation events at the species level. Ancestral reconstruction tools integrate this information (or part of it) in parsimony-based or probabilistic frameworks to infer gene content evolution (Arvestad et al., 2003; Csűrös and Miklós, 2006; Jacox et al., 2016; Szöllõsi et al., 2013). While parsimony-based methods are less computationally demanding, they require the user to specify a cost for each type of event, which will be a key factor in determining the ancestral reconstruction. Since such costs are unknown, it is usually a matter of evaluating different values and choosing the ones with more sensible outcomes (Boussau et al., 2004; Dagan and Martin, 2007; David and Alm, 2011). A better approach is to use probabilistic methods in which the rate at which each event occurs are parameters of the model that can be estimated from the data. The most popular approaches rely on birth-death models that consider rates of gene duplication, transfer and loss (DTL) (Csűrös and Miklós, 2006; Szöllősi and Daubin, 2012).

Ancestral reconstruction methods require two types of data: a rooted species phylogeny and information about homologous gene families (in contrast to orthologous gene families required, for instance, for supermatrix-based phylogenies) (Szöllősi et al., 2015a). The species tree is indispensable

to establish the relationships between organisms and to account for vertical inheritance. Information about homologous gene families is used in conjunction with the species phylogeny to infer different types of transmission events – vertical or horizontal – across the different branches of the species tree and to determine the gene content of ancestral nodes. Information about homologous gene families has been traditionally provided in the form of numerical profiles. A numerical profile consists of a matrix indicating how many copies of a given gene family are present in each one of the organisms under study. Less informative profiles that only indicate presence or absence can also be used but are discouraged.

More recent methods have been developed to use phylogenies of individual gene families instead of profiles (Jacox et al., 2016; Szöllõsi et al., 2013; Szöllősi et al., 2013). In addition to providing information about the number of copies, gene trees also inform about the individual evolutionary histories of gene families. Gene tree-aware methods can integrate this information by means of phylogenetic reconciliations, which attempt to fit the gene tree into the species tree. Inconsistencies between the gene tree and species tree are captured by different events such as duplications, losses and transfers. This additional layer of information makes these methods more sensitive to discern between different types of events than profile-based approaches. For example, without any prior information about the topology of a gene family, a horizontal transfer into a given lineage could look like as a loss, affecting the inferred gene content of the ancestral nodes (Figure 5). In fact, it has been suggested that profile-based approaches are inadequate to detect transfer events in gene families with a widespread taxonomic distribution resulting in overestimations of the ancestral genome sizes due to transfers being erroneously classified as different combinations of duplications and losses (Szöllősi et al., 2015a).

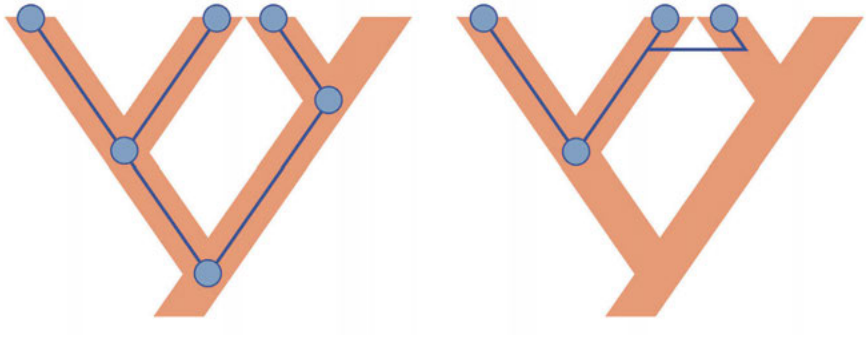

Figure 5. Two alternative ancestral reconstructions inferred for the same gene. The estimated evolutionary history of the gene (blue line) is represented inside the species phylogeny (pink). The presence of the gene in the ancestral and terminal nodes is shown by a blue dot. The use of gene phylogenies in ancestral reconstructions provides an additional source of information that facilitates the discrimination between scenarios.

Ancestral reconstruction using ALE undated

ALEml_undated (ALE hereafter) is a probabilistic gene tree-species tree reconciliation method to reconstruct gene-content evolution (Szöllősi et al., 2015a; Szöllõsi et al., 2013). ALE considers that gene families can evolve under a DTL model. In a first stage, rates for the different types of events are independently estimated for each gene family. Those rates are used in a second stage to reconstruct the evolution of the gene families by reconciling gene-tree and species trees. Two important considerations are taken into account by ALE that make it an attractive choice.

First, ALE can account for incomplete taxonomic sampling, by considering that lineages can be extinct or unsampled and contemplates HGT events involving unrepresented lineages. Notwithstanding, incomplete taxon sampling will probably lead to some events being incorrectly interpreted, especially when entire clades with distinctive characteristics are missing. However, the magnitude of such biases has not been properly evaluated.

Second, ALE incorporates the uncertainty associated with gene trees and reconciliations. As mentioned in previous sections, phylogenies of individual gene families are often not reliable either because their sequences are too short to carry enough phylogenetic information or due to methodological error mainly associated with model misspecifications. Moreover, for certain gene topologies, there can be several ways of reconciling the gene tree with the species tree, each implying a different evolutionary history. ALE takes into account both of those factors by computing the joint sequence-reconciliation likelihood of the reconciled tree in a sample – with the *sequence* term accounting for the accuracy of the gene tree and the second term for the reconciliation. For each gene family, instead of a single gene topology, it considers a distribution of gene trees. By a process known as amalgamation, the tree distribution (i.e., a set of bootstrap or MCMC trees) is combined to explore an even larger tree space than the one provided (David and Alm, 2011). Simply put, the amalgamation process combines clades coming from different trees to generate new trees that might or might not be found in the original distribution. The probability of a given (amalgamated) tree will depend on the number of times that such tree is found in the distribution. ALE approximates posterior probabilities for a large number of topologies from a reduced tree sample using conditional clade probabilities (Höhna and Drummond, 2012). Simultaneously, ALE considers all possible reconciliation scenarios for these gene trees. The output from ALE will be a distribution of a pre-specified number (a hundred by default) of reconciled trees that can give an idea of the uncertainty associated with the reconstruction. Strongly supported events will be captured in all or most reconciled trees while poorly supported scenarios will only be found in a small subset. Nevertheless, systematic errors in the

phylogenetic reconstruction of the gene family can generate biased gene tree distributions, which can be incorrectly interpreted as high certainty in the reconciled genes.

In this process, several assumptions are made. Possibly the most important one is that ALE, similarly to other ancestral reconstruction methods, does not account for uncertainty in the species phylogeny. In many cases, assuming that a certain species topology is correct is unrealistic given the current difficulties to obtain reliable phylogenies (as explained above). Incorrect species trees will necessarily affect the prediction of gene-specific evolutionary events.

Moreover, ALE considers that gene families evolve independently. This assumption might not be realistic, especially among neighbouring genes that often co-evolve (e.g., a single horizontal transfer or duplication involving several genes would be considered as different events). Additionally, DTL rates are considered constant across all branches of the species tree in a given gene family. In reality, different lineages evolve at different rates and, hence, models that account for heterogeneous rates at different branches are more suitable. The profile-based probabilistic model Count can consider rate variations across branches (Csurös, 2010), albeit its current implementation is not parallelized and the long running times limits its use to small datasets.

An additional consideration is that ALEml_undated, as indicated by its name, uses undated species trees, which differentiates it from the dated version of the same method that requires ultrametric species trees (i.e., trees in which nodes are ordered with respect to each other by absolute or relative times). This has an important implication: ALE undated generates ancestral reconstructions that are time-inconsistent. This means that transfers can occur between any lineages, even if they were not contemporary. Hence, incompatible scenarios, in which different transfers imply different speciation orders, can be present, with the only exception that transfers from a lineage to any of its ancestors are forbidden.

Finally, current ancestral reconstruction methods, including ALE, have been developed to be used with data generated from complete genomes. However, given that complete genomes are lacking for numerous lineages, partial genomes generated from metagenomic and single-cell approaches may need to be used to study the evolution of certain groups. The lower quality associated with such genomes (i.e., incompleteness and redundancy) probably has an impact on the ancestral reconstruction inferences, albeit the magnitude of this effect remains unknown.

Aims

The general aim of this thesis work has been to expand our knowledge on a specific group of archaea, now known as the Asgard superphylum, by using culture-independent approaches within an evolutionary genomic framework. Specifically, the main objectives of this work were:

1. To explore the diversity of Asgard archaea present in several environmental samples and to reconstruct their genomes.
2. To infer the evolutionary relationships between Asgard archaea and eukaryotes, and to identify eukaryotic signature proteins encoded in Asgard archaea genomes.
3. To estimate the gene content present in the last archaeal common ancestor of eukaryotes to infer its cellular and metabolic capabilities.
4. To obtain a complete genome of a Lokiarchaeota lineage to corroborate that the presence of ESPs in Asgard members is not the result of contamination and assembly artefacts present on published MAGs.

Results

Paper I. The Asgard superphylum

The origin of eukaryotes and their cellular complexity remains one of the biggest enigmas in biology. The discovery of *Lokiarchaeum* represents a major step forward in advancing our understanding of the identity and biology of the archaeal ancestor of eukaryotes. Phylogenetic inferences reveal that *Lokiarchaeum* affiliates with eukaryotes, further supporting scenarios in which Eukarya emerges from within Archaea. Moreover, the genome of *Lokiarchaeum* encodes a large number of proteins previously thought to be specific of eukaryotes.

To further explore the archaeal diversity and to better understand the role of Archaea in the evolution of eukaryotes, we used metagenomic approaches to reconstruct the genomes of seven lineages related to *Lokiarchaeum*. Collectively, these lineages form the Asgard archaea superphylum, comprised of at least four different phyla: Loki-, Odin-, Thor- and Heimdallarchaeota. Based on analyses of the SSU rRNA genes we observed that Asgard archaea are found worldwide, mostly in aquatic sediments, albeit frequently in low abundances. The presence of Odinarchaeota lineages seemed to be restricted to environments with high temperatures.

We performed in-depth phylogenetic analyses to determine the position of the Asgard superphylum in the tree of life. Our analyses showed high support for the monophyly of eukaryotes and Asgard archaea, with TACK archaea branching sister to this clade. In particular, we observed that Heimdallarchaeota often appeared as a sister group to eukaryotes, albeit this relationship was never statistically supported. In consequence, our analyses could not resolve the exact placement of eukaryotes, which could either branch as a sister group to Asgard archaea itself or emerge from within this superphylum.

Finally, we assessed the presence of eukaryotic signature proteins in Asgard genomes. Comparative genomic analyses revealed that, similarly to *Lokiarchaeum*, all Asgard members encoded for numerous homologs of ESPs. The eukaryotic-like proteins identified in Asgard archaea include homologs of components of the ubiquitin modifier system, ESCRT, proteins that in eukaryotes are required for protein translocation and glycosylation, as well as, elements of the cytoskeleton, such as actin, profilin, the subunit 4 of the actin-related protein (ARP) 2/3 complex, gelsolin domain-containing

proteins and tubulin. Furthermore, all Asgard genomes encoded numerous small GTPases.

Interestingly, our results also indicated the presence of additional homologs of eukaryotic proteins involved in vesicle formation and membrane trafficking in Thorarchaeota genomes. In particular, homologs of the eukaryotic transport protein particle (TRAPP) and the Sec23/Sec24 family were identified. Independent phylogenetic analyses determined that the Thorarchaeota homologs branched sister to their eukaryotic versions, suggesting an archaeal origin for these eukaryotic proteins that predates the mitochondrial acquisition. In addition, we observed that Thorarchaeota genomes encoded for two adjacent proteins with predicted beta-propeller and alpha-solenoid folds, respectively. Structural predictions based on the concatenation of these two proteins showed that they resembled eukaryotic coatomer proteins. The presence of these proteins opens up the possibility for an archaeal ancestor of eukaryotes with capabilities to form primordial coatomers.

Paper II. New Asgard lineages and updated evolutionary scenarios

To better understand the diversity within the Asgard superphylum we sampled sediments from twelve different geographical locations. Using metagenomics, we reconstructed 69 MAGs of novel Asgard members, from which we could identify five previously undescribed candidate phyla. We referred to these clades as Idunn-, Freya-, Baldr-, Frigg- and Gefionarchaeota. Temperature information taken from the sampling sites together with optimal growth temperature predictions based on genomic features suggested that Loki-, Thor- and most Heimdallarchaeota phyla were mesophiles, while the remaining groups were thermophiles, with Idunn-, Freya-, and Baldrarchaeota predicted to be hyperthermophiles.

We further explored the phylogenetic placement of these newly discovered lineages with respect to other Asgard archaea and eukaryotes. Initial maximum likelihood inferences based on a set of conserved ribosomal proteins placed Idunnarchaeota as a sister group to Korarchaeota. However, subsequent phylogenetic analyses ultimately indicated that this relationship was artefactual and that, in fact, Idunnarchaeota and Heimdallarchaeota were sister groups. The latter placement was further supported by the presence of numerous ESP homologs encoded in Idunnarchaeota genomes. While the results included in this thesis are preliminary and further work is still required, our phylogenetic reconstructions pointed to Idunnarchaeota as the closest archaeal relatives of eukaryotes. Yet, our analyses could not confidently exclude alternative scenarios in which eukaryotes branch sister

to the clade formed by Heimdall- and Idunnarchaeota, or, alternatively, from within Heimdallarchaeota.

Based on comparative analyses of these genomes we updated the distribution of ESPs present in Asgard archaea. Interestingly, we found additional homologs of proteins involved in the eukaryotic N-glycosylation encoded in some Asgard lineages. In particular, our results showed that components of the three subcomplexes that form the eukaryotic oligosaccharyltransferase are found in Asgards, suggesting that they were probably present in the archaeal host that ultimately gave rise to eukaryotes. Furthermore, we also reported the presence of several other eukaryotic components involved in various functions, such as membrane trafficking or informational processes.

Paper III. The nature of the Asgard ancestor of eukaryotes

The discovery of Asgard archaea has led to the proposal of updated hypotheses about the origin of eukaryotes. However, most of these studies have been based on the genomic features of one or few Asgard members. In this study, we took advantage of the expanded Asgard diversity to shed light into the nature of the last common archaeal ancestor of eukaryotes (LACAE).

We used a probabilistic approach based on gene tree-species tree reconciliations on a dataset comprised of members of the Asgard, TACK and Euryarchaeota clades, to estimate the evolutionary history of archaeal homologous gene families. We inspected the evolutionary dynamics of Asgard genomes with respect to other archaeal genomes. The gene content estimated for ancestral Asgard nodes is generally higher than for other groups. In particular, Lokiarchaeota show the largest genomes, whereas Odinarchaeota the most reduced. Additionally, we observed that Asgard genomes were more prone to gene duplication and loss events than other archaea, while no differences regarding the number of gene gains (i.e., horizontal gene transfer and *de novo* origination) were detected. Additional exploration of the inferred evolutionary events suggested that gene duplication and loss correlated with the proteome sizes and thus, the pattern observed in Asgard could be explained by their larger genomes.

We investigated the neighbouring nodes of the inferred branching point for the eukaryotes as a proxy for the gene content of LACAE (LACAE-proxy nodes). Based on previous phylogenetic analyses we considered the sister relationship between Idunnarchaeota and eukaryotes as our main hypothesis. Alternatively, we also considered two additional scenarios in which eukaryotes branch either sister to the Heimdall-Idunnarchaota clade or to Heimdallarchaeota.

Our analyses suggested that LACAE encoded many proteins currently classified as bacterial, suggesting numerous cases of ancient horizontal transfers between Archaea and Bacteria. The results of the ancestral reconstruction also indicated that a majority of the ESPs reported in extant Asgard lineages were inferred to be present in LACAE. Nevertheless, extant Asgard members showed important differences in the copy number of some of these ESPs. In particular, our analyses suggested that Lokiarchaeota experienced multiple duplications probably rendering its cytoskeleton substantially different from LACAE. Furthermore, preliminary metabolic reconstructions showed support for the absence of the Wood-Ljungdahl pathway in LACAE according to two of the three scenarios considered.

Paper IV. A near-complete Lokiarchaeota genome

Members of the Asgard superphylum have been characterized based on culture-independent approaches and, at the time of these analyses, there was no complete genome available for any Asgard lineage. Due to the metagenomic nature of the data, some researchers have questioned the quality and accuracy of previous analyses. It has been argued that the ESPs encoded in Asgard MAGs are the result of eukaryotic contamination present in their genomes and/or other assembly artefacts. Even though several lines of evidence have refuted such criticisms, having access to complete and high-quality Asgard archaeal genomes would surely alleviate doubts in the field.

To obtain complete Asgard genomes, we sequenced marine sediment samples using a combination of long and short read sequencing platforms. This generated a large amount of metagenomic data. To ease the assembly process, we reduced size of the dataset by enriching for reads originating from Asgard genomes. Hybrid assembly and subsequent binning of the subsampled data resulted in six Lokiarchaeota MAGs. We were able to assemble one near-complete Lokiarchaeota MAG (L04 hereafter) into only three contigs. Other MAGs were still high quality albeit slightly more fragmented. Our results showed that it is possible to obtain near-complete genomes from complex sediments samples using a combination of long and short sequencing reads, even for low abundant lineages in the presence of microdiversity.

Additionally, we performed assembly and binning using only short reads. We compared the quality of the hybrid L04 MAG with the corresponding MAG generated from short reads (SR-L04). The SR-L04 MAG was highly fragmented and more incomplete than the hybrid genome. However, the contamination levels were low in spite of being generated without performing any additional bin refinement step, indicating that the general criticism against the quality of MAGs is unfounded. Based on this data, we

emphasised that the accuracy of MAGs needs to be determined independently for each case taking into consideration the characteristics of the metagenomic data and the binning procedure.

We further assessed the presence of ESPs in the hybrid L04 MAG. We identified a similar number of ESPs as previously described for other Lokiarchaeota MAGs. Our results further indicated that the ESPs encoded in these genomes were not the result of contamination and/or binning artefacts as it has been previously suggested.

Perspectives

The work presented here highlights the value of culture-independent approaches to learn about organisms that, for now, cannot be grown in culture. Over the past years, the field of genome-centric metagenomics has consolidated. Matured tools and standard approaches are currently available for short-read metagenome assembly and genome binning. Recently, improvements on third-generation sequencing, especially in terms of throughput, have opened the door to a new era of long reads-based metagenomics. With long reads, it is possible to start thinking about recovering complete genomes of most microorganisms from natural populations. Insights gained during the development of short-read metagenomics and long-read genomics can be now used to develop tools and methods adapted to the new requirements of long-read metagenomics. This field is expected to grow quickly in the coming years. In consequence, we can expect to soon have complete or near-complete representative genomes for most Asgard phyla, from which much is yet to be known. These complete genomes will no doubt allow for more comprehensive analyses and, in turn, for rigorous testing of existing and forthcoming evolutionary hypotheses. Simultaneously, further exploration of undiscovered Asgard lineages might result in the identification of even closer archaeal relatives of eukaryotes and lead to new surprises.

The use of culture-independent approaches will only be strengthened by experimental efforts to culture these fascinating organisms. In the past years, culturing attempts have been mostly unsuccessful. However, these endeavours are now coming to fruition. In the weeks previous to the submission of this thesis work a manuscript was published in bioRxiv revealing that a species of Lokiarchaeota had been successfully cultured. This Lokiarchaeota co-culture shows cells displaying a bizarre morphology that underscores how little we know about uncultured microorganisms and archaea in particular. Having additional Asgard representatives in culture will help us to understand the capabilities and lifestyles of these organisms. The simultaneous combination of comparative genomics and evolutionary approaches together with experimental information obtained from the culture of Asgard lineages will be fundamental to elucidate the role of Asgard archaea in the early evolution of eukaryotes.

Svensk sammanfattning

Allt liv på jorden kan delas in i tre olika grupper, baserat på vilken typ av celler de består av: eukaryoter, bakterier och arkéer. Eukaryota celler är vanligtvis större och innehåller olika cellstrukturer, vilket gör dem till den mest komplexa typen av celler. Två exempel är cellkärnan, som innehåller cellens DNA, och mitokondrier, som är cellens energifabriker. Människor, djur, växter, svampar samt många typer av mikroorganismer tillhör den här gruppen. De andra två typerna av celler – bakterier och arkéer – är relativt små, har en enklare cellstruktur och är encelliga. Till skillnad från eukaryoterna så saknar bakterier och arkéer en cellkärna, vilket har lett till att de ibland kallas för prokaryota celler (från grekiskans pro: "före", karion: "kärna").

De morfologiska likheterna mellan de olika prokaryota celltyperna ledde till att arkéer ursprungligen antogs vara bakterier. Det var inte förrän 1977 som Carl Woese upptäckte, baserat på en studie på de olika celltypernas DNA, att dessa två grupper var fundamentalt olika. Han föreslog då namnet arkéer (ursprungligen arkebakterier) för att skilja de två grupperna åt. Efter deras upptäckt så studerades främst arkéer från extrema miljöer, vilket ledde till antagandet att arkéer var begränsade till den typen av nischer. Idag vet vi däremot att man kan hitta arkéer överallt på jorden.

Arkéerna är också särskilt intressanta från ett evolutionärt perspektiv. När det kommer till ursprunget av den eukaryota cellen så är en av de ledande hypoteserna att det var en arké som tog upp en bakterie, vilket ledde till de mer komplexa eukaryota cellerna vi kan se idag. För närvarande vet vi ganska lite om hur denna process gick till. Vilka typer av arkéer och bakterier var inblandade, och vilka egenskaper hade dessa? När det kommer till vilka dessa två organismer var så vet vi att bakterien tillhörde en grupp av bakterier vi kallar för alfaproteobakterier, samt att det var denna bakterie som gav upphov till det vi idag kallar för mitokondrien. När det kommer till den arkéella partnern så är det mer av ett mysterium. Detta beror dels på vår okunskap om denna typ av organismer. Vi känner exempelvis inte till några arkéer som orsakar sjukdomar hos människor. Detta har lett till ett medicinskt ointresse för denna typ av organismer, och studier av arkéer har främst bedrivits av ekologer. Vidare har studier av arkéer, samt många andra mikroorganismer, hindrats av det faktum att många organismer är svåra att odla i laborativa miljöer. Det har uppskattats att ungefär 99 % av alla mikroorganismer hittills inte gått att odla, då vi saknar information om vilka

levnadsförhållanden och näringsämnen som behövs för deras tillväxt. Men tack vare tekniska genombrott under de senaste två årtiondena så kan vi idag studera arvsmassan hos dessa organismer, utan att först odla dem i renkultur. Framför allt så har en teknik som kallas metagenomik lett till att många organismer, tidigare okända för vetenskapen, nu kan studeras i detalj. Denna teknik bygger på sekvenseringen av allt tillgängligt DNA i ett prov, vilket kan komma från vilken miljö som helst, för att sedan rekonstruera kromosomer och genom via bioinformatiska metoder. Genom att använda oss av informationen från dessa genom kan vi lära oss hur olika organismer uppstått, vad de behöver för att överleva samt deras evolutionära historia.

Denna avhandling fokuserar på studier av en grupp av arkéer, kallade Asgårdarkéer, via metagenomiska metoder, ur ett evolutionärt perspektiv. Våra analyser har visat att dessa Asgårdarkéer går att hitta över hela jorden, och är speciellt vanligt förekommande i sediment. Denna grupp av arkéer visar en väldig mångfald och består av minst 10 undergrupper vilka fått namnen Loki-, Tor-, Heimdall-, Oden-, Freja-, Hel-, Baldur-, Gefjon-, Frigg- och Idunarkéer. Evolutionära studier har visat att dess arkéer är de närmast besläktade organismerna till eukaryoter samt att arvsmassan från dessa innehåller gener som tidigare endast observerats i eukaryot DNA. Bland dessa finns gener som är inblandade i formationen av de cellstrukturer som ger de eukaryota cellerna deras komplexa karaktär. Detta antyder att de arkéer som gav upphov till de eukaryota cellerna var mer komplexa än vad vi tidigare trott. Dessa resultat visar att vi behöver se över de nuvarande teorierna angående de eukaryota cellernas ursprung.

Resumen en español

Los seres vivos pueden dividirse en tres dominios según el tipo de célula que poseen: eucariotas, bacterias y arqueas. Las células eucariotas son normalmente grandes y contienen múltiples compartimentos internos que las hacen más complejas. El núcleo y la mitocondria son ejemplos de compartimentos que delimitan el ADN y la fábrica energética de la célula respectivamente. Los seres humanos y el resto de los animales, las plantas, los hongos y otros muchos organismos invisibles a simple vista formamos parte de este grupo. Por el contrario, las bacterias y las arqueas son organismos unicelulares que tienen células relativamente pequeñas y simples. A diferencia de las células eucariotas, las bacterias y las arqueas carecen de núcleo, una característica por la que reciben el nombre de células procariotas (del griego *pro*: "antes de", *karion*: "núcleo").

La similitud en cuanto a la morfología de las células procariotas hizo que las arqueas fueran, en un principio, consideradas bacterias. No fue hasta 1977 cuando Carl Woese, basándose en el estudio del ADN de estos microorganismos, descubrió que estos dos grupos eran fundamentalmente diferentes y propuso el nombre de arquea (originalmente arqueobacteria) para diferenciarlos. En los años que siguieron a este descubrimiento, las arqueas que se estudiaron eran las que habitaban ambientes con condiciones extremas, lo cual llevó a que se generalizase la idea de que estos microorganismos estaban restringidos a este tipo de entornos. Sin embargo, hoy en día sabemos que las arqueas son ubicuas, pudiéndose encontrar casi cualquier parte del planeta.

Las arqueas son también muy interesantes desde un punto de vista evolutivo. Las principales hipótesis del origen de la vida defienden que las células eucariotas surgieron a partir de un proceso por el cual una arquea engulló a una bacteria. De esta combinación de, al menos, dos células relativamente simples surgió la complejidad que caracteriza a la célula eucariota. Actualmente, poco se sabe de cómo transcurrió este proceso evolutivo que tuvo lugar hace más de 1.900 millones de años. Sobre la identidad de estos dos organismos, sí que sabemos que fue un miembro de un grupo de bacterias llamado alfa proteobacterias el que fue engullido y que más tarde dio lugar a la mitocondria. Sin embargo, la identidad de la arquea sigue siendo un misterio. En parte esto se debe al gran desconocimiento que hay sobre este grupo de organismos. Hasta la fecha, no se ha encontrado ninguna arquea que produzca enfermedades en humanos. Por tanto, al

carecer de interés médico o de aplicaciones directas para nosotros, se ha relegado su estudio a una comunidad de científicos reducida y centrada en su ecología. Además, durante años el estudio de las arqueas, así como el de otros microorganismos, ha estado muy limitado por la dificultad que conlleva el cultivo de estos organismos en las condiciones artificiales proporcionadas en un laboratorio. De hecho, se estima que el 99% de los microorganismos aún no han podido ser cultivados, ya que se desconocen las condiciones de crecimiento que requieren. Sin embargo, el avance tecnológico en las últimas dos décadas ha hecho posible que se pueda estudiar el ADN de organismos sin la necesidad de cultivarlos. En particular, gracias a una técnica llamada metagenómica, muchos microorganismos desconocidos hasta ahora han podido estudiarse con gran detalle. Esta técnica se basa en la secuenciación de todo el ADN que se encuentra en una muestra, la cual puede proceder de cualquier entorno, y la posterior reconstrucción de los genomas de los organismos presentes mediante métodos bioinformáticos. Usando la información de estos genomas se puede inferir cómo están formadas las células de estos organismos, qué necesitan para vivir y su historia evolutiva.

Esta tesis se ha centrado en el estudio de un grupo de arqueas, ahora conocidas con el nombre de Asgard arqueas, a través de técnicas metagenómicas y desde un marco evolutivo. Nuestros análisis muestran que las Asgard arqueas se encuentran distribuidas a lo largo del planeta, principalmente en sedimentos acuáticos. Este grupo de arqueas es muy diverso y está compuesto por al menos diez subgrupos que hemos bautizado con el nombre de Loki-, Thor-, Heimdall-, Odin-, Freya-, Hel-, Baldur-, Gefion-, Frigg- e Idunnarquea. Los estudios evolutivos llevados a cabo muestran que estas arqueas son el pariente vivo más cercano de las células eucariotas. Nuestros resultados también revelan que los genomas de las Asgard arqueas contienen genes que previamente solo se habían encontrado en genomas de eucariotas. Entre ellos se encuentran genes involucrados en la formación de las estructuras que dan a las células eucariotas su complejidad característica. Esto sugiere que la arquea que dio lugar a los eucariotas no era tan simple como se presuponía. Estos resultados evidencian la necesidad de replantear las teorías evolutivas que explican el origen de la célula eucariota. Asimismo, sientan las bases a la hora de comprender en más detalle uno de los eventos más importantes en la historia de la vida en la Tierra.

Acknowledgements

First, I would like to thank my supervisor Thijs Ettema for giving me the opportunity to work in such a great lab. I'm grateful for your support during all these years and the trust you have always put on me. Thanks to Laura, Courtney, Kasia and Jimmy, my co-supervisors over the years, for all the great discussions, feedback and encouragement. I have learnt a lot from working with you.

I also want to thank all the amazing people who have made of this an unforgettable experience. I can very happily say that during these years I have gained many new families. Families full of weirdos who I love. Thanks to all the amazing colleagues and friends at Molevo for making these years so enjoyable. For your enthusiasm and for always helping without hesitation. For the fikas and the beers. I take many good lasting friendships with me. Also, thanks to the MIB family for being so welcoming and warm. You guys have filled this last year with countless good moments and laughs. I never expected to find such a nice group of people who are so caring and loving as you guys are. Huge thanks to all my friends over the years for making this journey so much nicer. You are the *bestest* and I love you. You have been an endless source of great times and random craziness, but also of energy and strength when I needed it the most. I'm especially grateful that you took care of me during these last months. I wouldn't have made it without all the conversations, moral support, hugs, tupperwares, gifts and cute emojis I got from you. A special mention to Anders, Dani, Jennah, Laura, Courtney, Andrea, Alejo, Kasia, Joran, Ana and Karl for proofreading parts of this thesis, and to my sister Marina for the cover illustration. To the ones who are no longer here, I know you would be proud.

Thanks to the patata family for all these years together. For being a fundamental pillar of the lamest and greatest team ever. Your support and love during these years is invaluable. Tack så mycket till min nya svenska familj för att ni alltid ger så mycket kärlek till mig.

Finalmente, gracias a toda mi familia, la original, por vuestro amor y apoyo incondicional. Especialmente a los mejores proges y sestras del universo. Gracias por estar siempre ahí. Por todos los mimos y gachas. Por el suministro continuo de pipas (gigantes con sal!!). Por ser una fuente infinita admiración. Por nuestras conversaciones sobre cualquier cosa y las risas. Por enseñarme tanto.

<3

References

Abby, S.S., Tannier, E., Gouy, M., and Daubin, V. (2012). Lateral gene transfer as a support for the tree of life. Proceedings of the National Academy of Sciences *109*, 4962-4967.

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., and Ikemura, T. (2003). Informatics for unveiling hidden genome signatures. Genome Res *13*, 693-702.

Abedon, S.T. (2013). Are archaeons incapable of being parasites or have we simply failed to notice? BioEssays *35*, 501-501.

Adam, P.S., Borrel, G., Brochier-Armanet, C., and Gribaldo, S. (2017). The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. ISME J *11*, 2407-2425.

Adl, S.M., Simpson, A.G.B., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V.*, et al.* (2012). The Revised Classification of Eukaryotes. Journal of Eukaryotic Microbiology *59*, 429-514.

Akerborg, O., Sennblad, B., Arvestad, L., and Lagergren, J. (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc Natl Acad Sci U S A *106*, 5714-5719.

Akil, C., and Robinson, R.C. (2018). Genomes of Asgard archaea encode profilins that regulate actin. Nature *562*, 439-443.

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., and Nielsen, P.H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol *31*, 533-538.

Albertsen, M., Karst, S.M., Ziegler, A.S., Kirkegaard, R.H., and Nielsen, P.H. (2015). Back to Basics--The Influence of DNA Extraction and Primer Choice on Phylogenetic Analysis of Activated Sludge Communities. PLoS One *10*, e0132783.

Ali, R.H., Bogusz, M., and Whelan, S. (2019). Identifying clusters of high confidence homologies in multiple sequence alignments. Mol Biol Evol.

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. Nat Methods *11*, 1144-1146.

Alneberg, J., Karlsson, C.M.G., Divne, A.-M., Bergin, C., Homa, F., Lindh, M.V., Hugerth, L.W., Ettema, T.J.G., Bertilsson, S., Andersson, A.F.*, et al.* (2018). Genomes from uncultivated prokaryotes: a comparison of

metagenome-assembled and single-amplified genomes. Microbiome *6*, 173.

Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B. (2003). Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics *19 Suppl 1*, i7-15.

Awad, S., Irber, L., and Titus Brown, C. (2017). Evaluating Metagenome Assembly on a Simple Defined Community with Many Strain Variants.

Ayling, M., Clark, M.D., and Leggett, R.M. (2019). New approaches for metagenome assembly with short reads. Brief Bioinform.

Bang, C., and Schmitz, R.A. (2015). Archaea associated with human surfaces: not to be underestimated. FEMS Microbiol Rev *39*, 631-648.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D.*, et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol *19*, 455-477.

Bates, S.T., Berg-Lyons, D., Caporaso, J.G., Walters, W.A., Knight, R., and Fierer, N. (2011). Examining the global distribution of dominant archaeal populations in soil. ISME J *5*, 908-917.

Baum, D.A., and Baum, B. (2014). An inside-out origin for the eukaryotic cell. BMC Biol *12*, 76.

Beiko, R.G., Doolittle, W.F., and Charlebois, R.L. (2008). The impact of reticulate evolution on genome phylogeny. Syst Biol *57*, 844-856.

Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A.H.Q., Kumar, M.S., Li, C., Dvornicic, M., Soldo, J.P., Koh, J.Y., Tong, C.*, et al.* (2019). Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nat Biotechnol *37*, 937-944.

Betts, H.C., Puttick, M.N., Clark, J.W., Williams, T.A., Donoghue, P.C.J., and Pisani, D. (2018). Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nature Ecology & Evolution *2*, 1556-1562.

Bishara, A., Moss, E.L., Kolmogorov, M., Parada, A.E., Weng, Z., Sidow, A., Dekas, A.E., Batzoglou, S., and Bhatt, A.S. (2018). High-quality genome sequences of uncultured microbes by assembly of read clouds. Nat Biotechnol.

Blackburne, B.P., and Whelan, S. (2013). Class of Multiple Sequence Alignment Algorithm Affects Genomic Analysis. Molecular Biology and Evolution *30*, 642-653.

Blanquart, S., and Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol *25*, 842-858.

Boetius, A., Ravenschlag, K., Schubert, C.J., Rickert, D., Widdel, F., Gieseke, A., Amann, R., Jorgensen, B.B., Witte, U., and Pfannkuche, O. (2000). A marine microbial consortium apparently mediating anaerobic oxidation of methane. Nature *407*, 623-626.

Bohlin, J., Skjerve, E., and Ussery, D.W. (2008). Investigations of oligonucleotide usage variance within and between prokaryotes. PLoS Comput Biol *4*, e1000057.

Bonk, F., Popp, D., Harms, H., and Centler, F. (2018). PCR-based quantification of taxa-specific abundances in microbial communities: Quantifying and avoiding common pitfalls. J Microbiol Methods *153*, 139-147.

Booth, A., and Doolittle, W.F. (2015). Eukaryogenesis, how special really? Proc Natl Acad Sci U S A *112*, 10278-10285.

Boussau, B., and Daubin, V. (2010). Genomes as documents of evolutionary history. Trends Ecol Evol *25*, 224-232.

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

Boussau, B., Szöllosi, G.J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. Genome Res *23*, 323-330.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A.*, et al.* (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature Biotechnology *35*, 725-731.

Boyce, K., Sievers, F., and Higgins, D.G. (2015). Instability in progressive multiple sequence alignment algorithms. Algorithms Mol Biol *10*, 26.

Branton, D., and Deamer, D. (2019). Nanopore Sequencing.

Breitwieser, F.P., Lu, J., and Salzberg, S.L. (2017). A review of methods and databases for metagenomic classification and assembly. Brief Bioinform.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D.*, et al.* (1996). Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Science *273*, 1058-1073.

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*, 1972-1973.

Castelle, C.J., Wrighton, K.C., Thomas, B.C., Hug, L.A., Brown, C.T., Wilkins, M.J., Frischkorn, K.R., Tringe, S.G., Singh, A., Markillie, L.M.*, et al.* (2015). Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr Biol *25*, 690-701.

Cavalier-Smith, T. (1983). Endosymbiotic origin of the mitochondrial envelope. Endocytobiology II, 265-279.

Cavalier-Smith, T. (1987). Eukaryotes with no mitochondria. Nature *326*, 332-333.

Cavalier-Smith, T. (2007). The chimaeric origin of mitochondria: photosynthetic cell enslavement, gene-transfer pressure, and compartmentation efficiency. In Origin of mitochondria and hydrogenosomes (Springer), pp. 161-199.

Chaban, B., Ng, S.Y., and Jarrell, K.F. (2006). Archaeal habitats--from the extreme to the ordinary. Can J Microbiol *52*, 73-116.

Chatzou, M., Magis, C., Chang, J.-M., Kemena, C., Bussotti, G., Erb, I., and Notredame, C. (2016). Multiple sequence alignment modeling: methods and applications. Brief Bioinform *17*, 1009-1023.

Chauve, C., Rafiey, A., Davin, A., Scornavacca, C., Veber, P., Boussau, B., Szollosi, G., Daubin, V., and Tannier, E. (2017a). MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene Transfers.

Chauve, C., Rafiey, A., Davín, A.A., Scornavacca, C., Veber, P., Boussau, B., Szöllősi, G.J., Daubin, V., and Tannier, E. (2017b). MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers.

Chernikova, D., Motamedi, S., Csuros, M., Koonin, E.V., and Rogozin, I.B. (2011). A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. Biol Direct *6*, 26.

Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E.*, et al.* (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods *10*, 563-569.

Clark, C.G., and Roger, A.J. (1995). Direct evidence for secondary loss of mitochondria in Entamoeba histolytica. Proc Natl Acad Sci U S A *92*, 6518-6521.

Collins, T.M., Wimberger, P.H., and Naylor, G.J.P. (1994). Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. Systematic Biology *43*, 482.

Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet *9*, 938-950.

Courties, C., Vaquer, A., Troussellier, M., Lautier, J., Chrétiennot-Dinet, M.J., Neveux, J., Machado, C., and Claustre, H. (1994). Smallest eukaryotic organism. Nature *370*, 255-255.

Cox, C.J., Foster, P.G., Hirt, R.P., Harris, S.R., and Embley, T.M. (2008). The archaebacterial origin of eukaryotes. Proc Natl Acad Sci U S A *105*, 20356-20361.

Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol *10*, 210.

Csurös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics *26*, 1910-1912.

Csűrös, M., and Miklós, I. (2006). A Probabilistic Model for Gene Content Evolution with Duplication, Loss, and Horizontal Transfer. Lecture Notes in Computer Science, 206-220.

Cunha, V.D., Da Cunha, V., Gaia, M., Gadelle, D., Nasir, A., and Forterre, P. (2017). Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLOS Genetics *13*, e1006810.

Dagan, T., and Martin, W. (2007). Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proc Natl Acad Sci U S A *104*, 870-875.

Daniel, R. (2005). The metagenomics of soil. Nat Rev Microbiol *3*, 470-478.

David, L.A., and Alm, E.J. (2011). Rapid evolutionary innovation during an Archaean genetic expansion. Nature *469*, 93-96.

Davin, A.A., Tannier, E., Williams, T.A., Boussau, B., Daubin, V., and Szollosi, G.J. (2018). Gene transfers can date the tree of life. Nat Ecol Evol *2*, 904-909.

Davín, A.A., Tannier, E., Williams, T.A., Boussau, B., Daubin, V., and Szöllősi, G.J. (2018). Gene transfers can date the tree of life. Nat Ecol Evol *2*, 904-909.

Desai, A., Marwah, V.S., Yadav, A., Jha, V., Dhaygude, K., Bangar, U., Kulkarni, V., and Jere, A. (2013). Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data. PLoS ONE *8*, e60204.

Dick, G.J., Andersson, A.F., Baker, B.J., Simmons, S.L., Thomas, B.C., Yelton, A.P., and Banfield, J.F. (2009). Community-wide analysis of microbial genome sequence signatures. Genome Biol *10*, R85.

Doolittle, W.F. (1999). Phylogenetic Classification and the Universal Tree. Science *284*, 2124-2128.

dos Reis, M., Donoghue, P.C., and Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet *17*, 71-80.

Elliott, T.A., and Ryan Gregory, T. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philosophical Transactions of the Royal Society B: Biological Sciences *370*, 20140331.

Eloe-Fadrosh, E.A., Ivanova, N.N., Woyke, T., and Kyrpides, N.C. (2016). Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. Nat Microbiol *1*, 15032.

Embley, T.M., and Martin, W. (2006). Eukaryotic evolution, changes and challenges. Nature *440*, 623-630.

Eme, L., Sharpe, S.C., Brown, M.W., and Roger, A.J. (2014). On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. Cold Spring Harb Perspect Biol *6*.

Eme, L., Spang, A., Lombard, J., Stairs, C.W., and Ettema, T.J.G. (2018). Archaea and the origin of eukaryotes. Nat Rev Microbiol *16*, 120.

Farrelly, V., Rainey, F.A., and Stackebrandt, E. (1995). Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. Appl Environ Microbiol *61*, 2798-2801.

Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. Systematic Zoology *27*, 401.

Felsenstein, J. (2004). Inferring Phylogenies (Sinauer Associates Incorporated).

Fitch, W.M., and Margoliash, E. (1967). Construction of phylogenetic trees. Science *155*, 279-284.

Foster, P.G., Cox, C.J., and Embley, T.M. (2009a). The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philos Trans R Soc Lond B Biol Sci *364*, 2197-2207.

Foster, P.G., Cox, C.J., and Martin Embley, T. (2009b). The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philosophical Transactions of the Royal Society B: Biological Sciences *364*, 2197-2207.

Foster, P.G., and Hickey, D.A. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol *48*, 284-290.

Frostegård, A., Courtois, S., Ramisse, V., Clerc, S., Bernillon, D., Le Gall, F., Jeannin, P., Nesme, X., and Simonet, P. (1999). Quantification of bias related to the extraction of DNA directly from soils. Appl Environ Microbiol *65*, 5409-5420.

Galtier, N. (2007). A model of horizontal gene transfer and the bacterial phylogeny problem. Syst Biol *56*, 633-642.

Galtier, N., and Daubin, V. (2008). Dealing with incongruence in phylogenomic analyses. Philos Trans R Soc Lond B Biol Sci *363*, 4023-4029.

Ghurye, J., and Pop, M. (2019). Modern technologies and algorithms for scaffolding assembled genomes. PLoS Comput Biol *15*, e1006994.

Ghurye, J.S., Cepeda-Espinoza, V., and Pop, M. (2016). Metagenomic Assembly: Overview, Challenges and Applications. Yale J Biol Med *89*, 353-362.

Gijzen, H.J., Broers, C.A., Barughare, M., and Stumm, C.K. (1991). Methanogenic bacteria as endosymbionts of the ciliate Nyctotherus ovalis in the cockroach hindgut. Appl Environ Microbiol *57*, 1630-1634.

Gilmore, J.A., McGann, L.E., Liu, J., Gao, D.Y., Peter, A.T., Kleinhans, F.W., and Critser, J.K. (1995). Effect of cryoprotectant solutes on water permeability of human spermatozoa. Biol Reprod *53*, 985-995.

Goyanes, V.J., Ron-Corzo, A., Costas, E., and Maneiro, E. (1990). Morphometric categorization of the human oocyte and early conceptus. Hum Reprod *5*, 613-618.

Grant, C.R., Wan, J., and Komeili, A. (2018). Organelle Formation in Bacteria and Archaea. Annu Rev Cell Dev Biol *34*, 217-238.

Gribaldo, S., and Brochier-Armanet, C. (2006). The origin and evolution of Archaea: a state of the art. Philos Trans R Soc Lond B Biol Sci *361*, 1007-1022.
Gribaldo, S., and Brochier-Armanet, C. (2012). Time for order in microbial systematics. Trends Microbiol *20*, 209-210.
Gribaldo, S., and Philippe, H. (2002). Ancient Phylogenetic Relationships. Theoretical Population Biology *61*, 391-408.
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? Nat Rev Microbiol *8*, 743-752.
Guy, L., and Ettema, T.J.G. (2011). The archaeal 'TACK' superphylum and the origin of eukaryotes. Trends in Microbiology *19*, 580-587.
Guy, L., Saw, J.H., and Ettema, T.J.G. (2014). The archaeal legacy of eukaryotes: a phylogenomic perspective. Cold Spring Harb Perspect Biol *6*, a016022.
Hampl, V., Čepička, I., and Eliáš, M. (2019). Was the Mitochondrion Necessary to Start Eukaryogenesis? Trends Microbiol *27*, 96-104.
Hartman, H., and Fedorov, A. (2002). The origin of the eukaryotic cell: a genomic investigation. Proc Natl Acad Sci U S A *99*, 1420-1425.
Heim, N.A., Payne, J.L., Finnegan, S., Knope, M.L., Kowalewski, M., Lyons, S.K., McShea, D.W., Novack-Gottshall, P.M., Smith, F.A., and Wang, S.C. (2017). Hierarchical complexity and the size limits of life. Proc Biol Sci *284*.
Heiss, A.A., Kolisko, M., Ekelund, F., Brown, M.W., Roger, A.J., and Simpson, A.G.B. (2018). Combined morphological and phylogenomic re-examination of malawimonads, a critical taxon for inferring the evolutionary history of eukaryotes. Royal Society Open Science *5*, 171707.
Henderson, G.P., Gan, L., and Jensen, G.J. (2007). 3-D Ultrastructure of O. tauri: Electron Cryotomography of an Entire Eukaryotic Cell. PLoS One *2*, e749.
Hernandez, A.M., and Ryan, J.F. (2019). Six-state amino acid recoding is not an effective strategy to offset the effects of compositional heterogeneity and saturation in phylogenetic analyses.
Ho, S.Y., and Duchêne, S. (2014). Molecular-clock methods for estimating evolutionary rates and timescales. Molecular ecology *23*, 5947-5965.
Höhna, S., and Drummond, A.J. (2012). Guided tree topology proposals for Bayesian phylogenetic inference. Syst Biol *61*, 1-11.
Holder, M., and Lewis, P.O. (2003). Phylogeny estimation: traditional and Bayesian approaches. Nature Reviews Genetics *4*, 275-284.
Hoshino, T., and Inagaki, F. (2019). Abundance and distribution of Archaea in the subseafloor sedimentary biosphere. ISME J *13*, 227-231.
Hossain, A.S.M.M., Blackburne, B.P., Shah, A., and Whelan, S. (2015). Evidence of Statistical Inconsistency of Phylogenetic Methods in the

Presence of Multiple Sequence Alignment Uncertainty. Genome Biol Evol *7*, 2102-2116.

Hou, W., Wang, S., Dong, H., Jiang, H., Briggs, B.R., Peacock, J.P., Huang, Q., Huang, L., Wu, G., Zhi, X.*, et al.* (2013). A comprehensive census of microbial diversity in hot springs of Tengchong, Yunnan Province China using 16S rRNA gene pyrosequencing. PLoS One *8*, e53350.

Hrdy, I., Hirt, R.P., Dolezal, P., Bardonová, L., Foster, P.G., Tachezy, J., and Embley, T.M. (2004). Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature *432*, 618-622.

Huber, H., Hohn, M.J., Rachel, R., Fuchs, T., Wimmer, V.C., and Stetter, K.O. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature *417*, 63-67.

Hugenholtz, P., Skarshewski, A., and Parks, D.H. (2016). Genome-Based Microbial Taxonomy Coming of Age. Cold Spring Harb Perspect Biol *8*.

Ip, C.L.C., Loose, M., Tyson, J.R., de Cesare, M., Brown, B.L., Jain, M., Leggett, R.M., Eccles, D.A., Zalunin, V., Urban, J.M.*, et al.* (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Res *4*, 1075.

Jacox, E., Chauve, C., Szöllősi, G.J., Ponty, Y., and Scornavacca, C. (2016). ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. Bioinformatics *32*, 2056-2058.

Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B., and Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. Nat Methods *12*, 351-356.

Jiang, Y.-X., Wu, J.-G., Yu, K.-Q., Ai, C.-X., Zou, F., and Zhou, H.-W. (2011). Integrated lysis procedures reduce extraction biases of microbial DNA from mangrove sediment. J Biosci Bioeng *111*, 153-157.

Jordan, G., and Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol Biol Evol *29*, 1125-1139.

Jorgensen, S.L., Hannisdal, B., Lanzen, A., Baumberger, T., Flesland, K., Fonseca, R., Ovreas, L., Steen, I.H., Thorseth, I.H., Pedersen, R.B.*, et al.* (2012). Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge. Proc Natl Acad Sci U S A *109*, E2846-2855.

Kamath, G.M., Shomorony, I., Xia, F., Courtade, T.A., and Tse, D.N. (2017). HINGE: long-read assembly achieves optimal repeat resolution. Genome Res *27*, 747-756.

Kandler, O., and Hippe, H. (1977). Lack of peptidoglycan in the cell walls of Methanosarcina barkeri. Arch Microbiol *113*, 57-60.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359.

Karin, E.L., Susko, E., and Pupko, T. (2014). Alignment Errors Strongly Impact Likelihood-Based Tests for Comparing Topologies. Molecular Biology and Evolution *31*, 3057-3067.

Karner, M.B., DeLong, E.F., and Karl, D.M. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature *409*, 507-510.

Karnkowska, A., Vacek, V., Zubacova, Z., Treitli, S.C., Petrzelkova, R., Eme, L., Novak, L., Zarsky, V., Barlow, L.D., Herman, E.K., *et al.* (2016). A Eukaryote without a Mitochondrial Organelle. Curr Biol *26*, 1274-1284.

Karst, S.M., Dueholm, M.S., McIlroy, S.J., Kirkegaard, R.H., Nielsen, P.H., and Albertsen, M. (2018). Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. Nat Biotechnol *36*, 190-195.

Karst, S.M., Kirkegaard, R.H., and Albertsen, M. (2016). mmgenome: a toolbox for reproducible genome extraction from metagenomes.

Keeling, P.J., and Slamovits, C.H. (2005). Causes and effects of nuclear genome reduction. Current Opinion in Genetics & Development *15*, 601-608.

Kelly, S., Wickstead, B., and Gull, K. (2011). Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. Proc Biol Sci *278*, 1009-1018.

Kletzin, A. (2007). General characteristics and important model organisms. In Archaea (American Society of Microbiology), pp. 14-92.

Klinger, C.M., Spang, A., Dacks, J.B., and Ettema, T.J. (2016). Tracing the Archaeal Origins of Eukaryotic Membrane-Trafficking System Building Blocks. Mol Biol Evol *33*, 1528-1541.

Konstantinidis, K.T., Rossello-Mora, R., and Amann, R. (2017). Uncultivated microbes in need of their own taxonomy. ISME J *11*, 2399-2406.

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

Koonin, E.V. (2015). Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? Philos Trans R Soc Lond B Biol Sci *370*, 20140333.

Koonin, E.V., and Yutin, N. (2014). The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. Cold Spring Harb Perspect Biol *6*, a016188.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res *27*, 722-736.

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*, 10139-10146.
Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008). A Bioinformatician's Guide to Metagenomics. Microbiol Mol Biol Rev *72*, 557-578.
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. Proc Natl Acad Sci U S A *81*, 3786-3790.
Lambowitz, A.M., and Zimmerly, S. (2004). Mobile group II introns. Annu Rev Genet *38*, 1-35.
Langworthy, T.A., Smith, P.F., and Mayberry, W.R. (1972). Lipids of Thermoplasma acidophilum. J Bacteriol *112*, 1193-1200.
Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol *7 Suppl 1*, S4.
Lartillot, N., and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol *21*, 1095-1109.
Lasek-Nesselquist, E., and Gogarten, J.P. (2013). The effects of model choice and mitigating bias on the ribosomal tree of life. Molecular Phylogenetics and Evolution *69*, 17-38.
Lasken, R.S. (2013). Single-cell sequencing in its prime. Nat Biotechnol *31*, 211-212.
Le, S.Q., and Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. Molecular Biology and Evolution *25*, 1307-1320.
Lerat, E., Daubin, V., Ochman, H., and Moran, N.A. (2005). Evolutionary origins of genomic repertoires in bacteria. PLoS Biol *3*, e130.
Levin, P.A., and Angert, E.R. (2015). Small but Mighty: Cell Size and Bacteria. Cold Spring Harb Perspect Biol *7*, a019216.
Li, D., Liu, C.M., Luo, R., Sadakane, K., and Lam, T.W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674-1676.
Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics *32*, 2103-2110.
Lin, H.H., and Liao, Y.C. (2016). Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. Sci Rep *6*, 24175.
Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M., and Pevzner, P.A. (2016). Assembly of long error-prone reads using de Bruijn graphs. Proc Natl Acad Sci U S A *113*, E8396-E8405.
Lind, A.E., Lewis, W.H., Spang, A., Guy, L., Embley, T.M., and Ettema, T.J.G. (2018). Genomes of two archaeal endosymbionts show convergent adaptations to an intracellular lifestyle. ISME J *12*, 2655-2667.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. J Biomed Biotechnol *2012*, 251364.

Lloyd, K.G., Steen, A.D., Ladau, J., Yin, J., and Crosby, L. (2018). Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. mSystems *3*.

Loman, N.J., Quick, J., and Simpson, J.T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods *12*, 733-735.

Lopez-Garcia, P., and Moreira, D. (2015). Open Questions on the Origin of Eukaryotes. Trends Ecol Evol *30*, 697-708.

Lu, Y.Y., Chen, T., Fuhrman, J.A., and Sun, F. (2017). COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. Bioinformatics *33*, 791-798.

Lynch, M., and Marinov, G.K. (2017). Membranes, energetics, and evolution across the prokaryote-eukaryote divide. Elife *6*.

Mahnert, A., Blohs, M., Pausan, M.-R., and Moissl-Eichinger, C. (2018). The human archaeome: methodological pitfalls and knowledge gaps. Emerging Topics in Life Sciences *2*, 469-482.

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 Res *33*, 4626-4638.

Martijn, J., and Ettema, T.J. (2013). From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. Biochem Soc Trans *41*, 451-457.

Martijn, J., Lind, A.E., Schön, M.E., Spiertz, I., Juzokaite, L., Bunikis, I., Pettersson, O.V., and Ettema, T.J.G. (2019). Confident phylogenetic identification of uncultured prokaryotes through long read amplicon sequencing of the 16S-ITS-23S rRNA operon. Environmental Microbiology.

Martijn, J., Vosseberg, J., Guy, L., Offre, P., and Ettema, T.J.G. (2018). Deep mitochondrial origin outside the sampled alphaproteobacteria. Nature *557*, 101-105.

Martin, W., Hoffmeister, M., Rotte, C., and Henze, K. (2001). An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. Biol Chem *382*, 1521-1539.

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

Martin, W.F., Roettger, M., Ku, C., Garg, S.G., Nelson-Sathi, S., and Landan, G. (2017). Late Mitochondrial Origin Is an Artifact. Genome Biology and Evolution *9*, 373-379.

McInerney, J.O., O'Connell, M.J., and Pisani, D. (2014). The hybrid nature of the Eukaryota and a consilient view of life on Earth. Nature Reviews Microbiology *12*, 449-455.
McLaughlin, P.J., and Dayhoff, M.D. (1970). Eukaryotes versus prokaryotes: an estimate of evolutionary distance. Science *168*, 1469-1471.
McLysaght, A., and Guerzoni, D. (2015). New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philos Trans R Soc Lond B Biol Sci *370*, 20140332.
Miller, J.R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics *95*, 315-327.
Miyamoto, M.M., and Fitch, W.M. (1996). Constraints on Protein Evolution and the Age of the Eubacteria/Eukaryote Split. Systematic Biology *45*, 568-575.
Moissl-Eichinger, C., and Huber, H. (2011). Archaeal symbionts and parasites. Curr Opin Microbiol *14*, 364-370.
Moreira, D., and Lopez-Garcia, P. (1998). Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. J Mol Evol *47*, 517-530.
Nakashima, H., Ota, M., Nishikawa, K., and Ooi, T. (1998). Genes from nine genomes are separated into their organisms in the dinucleotide composition space. DNA Res *5*, 251-259.
Nawrocki, E.P., Jones, T.A., and Eddy, S.R. (2018). Group I introns are widespread in archaea. Nucleic Acids Res *46*, 7970-7976.
Nicholls, S.M., Quick, J.C., Tang, S., and Loman, N.J. (2018). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. BioRxiv, 487033.
Nicholls, S.M., Quick, J.C., Tang, S., and Loman, N.J. (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience *8*.
Noble, P.A., Citek, R.W., and Ogunseitan, O.A. (1998). Tetranucleotide frequencies in microbial genomes. Electrophoresis *19*, 528-535.
Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res *27*, 824-834.
Offre, P., Spang, A., and Schleper, C. (2013). Archaea in biogeochemical cycles. Annu Rev Microbiol *67*, 437-457.
Ogden, T.H., Heath Ogden, T., and Rosenberg, M.S. (2006). Multiple Sequence Alignment Accuracy and Phylogenetic Inference. Systematic Biology *55*, 314-328.
Olson, N.D., Treangen, T.J., Hill, C.M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., and Pop, M. (2017). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief Bioinform.

Pace, N.R. (2009). Mapping the tree of life: progress and prospects. Microbiol Mol Biol Rev *73*, 565-576.

Pace, N.R., Stahl, D.A., Lane, D.J., and Olsen, G.J. (1986). The analysis of natural microbial populations by ribosomal RNA sequences. In Advances in microbial ecology (Springer), pp. 1-55.

Parfrey, L.W., Lahr, D.J., Knoll, A.H., and Katz, L.A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci U S A *108*, 13624-13629.

Park, S.J., Ghai, R., Martin-Cuadrado, A.B., Rodriguez-Valera, F., Chung, W.H., Kwon, K., Lee, J.H., Madsen, E.L., and Rhee, S.K. (2014). Genomes of two new ammonia-oxidizing archaea enriched from deep marine sediments. PLoS One *9*, e96449.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res *25*, 1043-1055.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol *2*, 1533-1542.

Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics *28*, 1420-1428.

Petitjean, C., Deschamps, P., Lopez-Garcia, P., and Moreira, D. (2014). Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. Genome Biol Evol *7*, 191-204.

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 Biology *9*, e1000602.

Philippe, H., de Vienne, D.M., Ranwez, V., Roure, B., Baurain, D., and Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. European Journal of Taxonomy.

Pittis, A.A., and Gabaldon, T. (2016). On phylogenetic branch lengths distribution and the late acquistion of mitochondria.

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

Poole, A.M., and Gribaldo, S. (2014). Eukaryotic origins: How and when was the mitochondrion acquired? Cold Spring Harb Perspect Biol *6*, a015990.

Pride, D.T., Meinersmann, R.J., Wassenaar, T.M., and Blaser, M.J. (2003). Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res *13*, 145-158.

Privman, E., Penn, O., and Pupko, T. (2012). Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol *29*, 1-5.
Quang, L.S., Gascuel, O., and Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics *24*, 2317-2323.
Raymann, K., Brochier-Armanet, C., and Gribaldo, S. (2015). The two-domain tree of life is linked to a new root for the Archaea. Proc Natl Acad Sci U S A *112*, 6670-6675.
Redeling, B.D., and Suchard, M.A. (2009). Robust Inferences from Ambiguous Alignments. Sequence AlignmentMethods, Models, Concepts, and Strategies, 208-270.
Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics *13*, 278-289.
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 *95*, 6239-6244.
Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Franz Lang, B., and Philippe, H. (2007a). Detecting and Overcoming Systematic Errors in Genome-Scale Phylogenies. Systematic Biology *56*, 389-399.
Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., and Philippe, H. (2007b). Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol *56*, 389-399.
Roger, A.J., Kolisko, M., and Simpson, A.G.B. (2013). Phylogenomic Analysis. Evolution of Virulence in Eukaryotic Microbes, 44-69.
Roger, A.J., Muñoz-Gómez, S.A., and Kamikawa, R. (2017). The Origin and Diversification of Mitochondria. Current Biology *27*, R1177-R1192.
Sagan, L. (1967). On the origin of mitosing cells. Journal of Theoretical Biology *14*, 225-IN226.
Salonen, A., Nikkilä, J., Jalanka-Tuovinen, J., Immonen, O., Rajilić-Stojanović, M., Kekkonen, R.A., Palva, A., and de Vos, W.M. (2010). Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. J Microbiol Methods *81*, 127-134.
Sandberg, R., Bränden, C.-I., Ernberg, I., and Cöster, J. (2003). Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G C content. Gene *311*, 35-42.
Schulz, H.N., Brinkhoff, T., Ferdelman, T.G., Mariné, M.H., Teske, A., and Jorgensen, B.B. (1999). Dense populations of a giant sulfur bacterium in Namibian shelf sediments. Science *284*, 493-495.
Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E.*, et al.* (2017). Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat Methods *14*, 1063-1071.
Shively, J.M. (2006). Complex Intracellular Structures in Prokaryotes (Springer Science & Business Media).

Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., and Banfield, J.F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature Microbiology *3*, 836-843.

Simon, C., and Daniel, R. (2011). Metagenomic analyses: past and future trends. Appl Environ Microbiol *77*, 1153-1161.

Smillie, C.S., Smith, M.B., Friedman, J., Cordero, O.X., David, L.A., and Alm, E.J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. Nature *480*, 241-244.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., Arrieta, J.M., and Herndl, G.J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A *103*, 12115-12120.

Som, A. (2015). Causes, consequences and solutions of phylogenetic incongruence. Brief Bioinform *16*, 536-548.

Somerville, V., Lutz, S., Schmid, M., Frei, D., Moser, A., Irmler, S., Frey, J.E., and Ahrens, C.H. (2019). Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. BMC Microbiology *19*, 143.

Song, W.-Z., and Thomas, T. (2017). Binning_refiner: improving genome bins through the combination of different binning programs. Bioinformatics *33*, 1873-1875.

Soubrier, J., Steel, M., Lee, M.S.Y., Der Sarkissian, C., Guindon, S., Ho, S.Y.W., and Cooper, A. (2012). The influence of rate heterogeneity among sites on the time dependence of molecular rates. Mol Biol Evol *29*, 3345-3358.

Soucy, S.M., Huang, J., and Gogarten, J.P. (2015). Horizontal gene transfer: building the web of life. Nat Rev Genet *16*, 472-482.

Spang, A., Caceres, E.F., and Ettema, T.J.G. (2017). Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. Science *357*.

Spang, A., Eme, L., Saw, J.H., Caceres, E.F., Zaremba-Niedzwiedzka, K., Lombard, J., Guy, L., and Ettema, T.J.G. (2018). Asgard archaea are the closest prokaryotic relatives of eukaryotes. PLoS Genet *14*, e1007080.

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 *521*, 173-179.

Spielman, S.J., Dawson, E.T., and Wilke, C.O. (2014). Limited utility of residue masking for positive-selection inference. Mol Biol Evol *31*, 2496-2500.

Stanier, R.Y., and Van Niel, C.B. (1962). The concept of a bacterium. Arch Mikrobiol *42*, 17-35.

Stepanauskas, R. (2012). Single cell genomics: an individual look at microbes. Curr Opin Microbiol *15*, 613-620.
Susko, E., and Roger, A.J. (2007). On reduced amino acid alphabets for phylogenetic inference. Mol Biol Evol *24*, 2139-2150.
Szöllosi, G.J., Boussau, B., Abby, S.S., Tannier, E., and Daubin, V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc Natl Acad Sci U S A *109*, 17513-17518.
Szöllősi, G.J., and Daubin, V. (2012). Modeling Gene Family Evolution and Reconciling Phylogenetic Discord. Methods in Molecular Biology, 29-51.
Szöllősi, G.J., Davín, A.A., Tannier, E., Daubin, V., and Boussau, B. (2015a). Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. Philos Trans R Soc Lond B Biol Sci *370*, 20140335.
Szöllõsi, G.J., Rosikiewicz, W., Boussau, B., Tannier, E., and Daubin, V. (2013). Efficient exploration of the space of reconciled gene trees. Syst Biol *62*, 901-912.
Szöllősi, G.J., Tannier, E., Daubin, V., and Boussau, B. (2015b). The inference of gene trees with species trees. Syst Biol *64*, e42-62.
Szöllősi, G.J., Tannier, E., Lartillot, N., and Daubin, V. (2013). Lateral Gene Transfer from the Dead. Systematic Biology *62*, 386-397.
Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol *56*, 564-577.
Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., and Dessimoz, C. (2015). Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. Syst Biol *64*, 778-791.
Tang, J. (2011). Microbial metabolomics. Curr Genomics *12*, 391-403.
Tekaia, F. (2016). Inferring Orthologs: Open Questions and Perspectives. Genomics Insights *9*, 17-28.
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 Biol Evol *4*, 466-485.
Torsvik, V., Ovreas, L., and Thingstad, T.F. (2002). Prokaryotic diversity--magnitude, dynamics, and controlling factors. Science *296*, 1064-1066.
Tovar, J., Fischer, A., and Clark, C.G. (1999). The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite Entamoeba histolytica. Mol Microbiol *32*, 1013-1021.
Tovar, J., Leon-Avila, G., Sanchez, L.B., Sutak, R., Tachezy, J., van der Giezen, M., Hernandez, M., Muller, M., and Lucocq, J.M. (2003). Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation. Nature *426*, 172-176.

Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S., and Turner, S.W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res *38*, e159.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The human microbiome project. Nature *449*, 804-810.

Tyakht, A.V., Kostryukova, E.S., Popenko, A.S., Belenikin, M.S., Pavlenko, A.V., Larin, A.K., Karpova, I.Y., Selezneva, O.V., Semashko, T.A., Ospanova, E.A.*, et al.* (2013). Human gut microbiota community structures in urban and rural populations in Russia. Nat Commun *4*, 2469.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., and Banfield, J.F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature *428*, 37-43.

Uritskiy, G.V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome *6*, 158.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W.*, et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science *304*, 66-74.

Viklund, J., Ettema, T.J.G., and Andersson, S.G.E. (2012). Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. Mol Biol Evol *29*, 599-615.

Vollmers, J., Wiegand, S., and Kaster, A.-K. (2017). Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! PLOS ONE *12*, e0169662.

von Dohlen, C.D., Kohler, S., Alsop, S.T., and McManus, W.R. (2001). Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. Nature *412*, 433-436.

von Wintzingerode, F., Göbel, U.B., and Stackebrandt, E. (1997). Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev *21*, 213-229.

Wagner, A., Whitaker, R.J., Krause, D.J., Heilers, J.-H., van Wolferen, M., van der Does, C., and Albers, S.-V. (2017). Mechanisms of gene flow in archaea. Nat Rev Microbiol *15*, 492-501.

Wang, H.-C., Li, K., Susko, E., and Roger, A.J. (2008). A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol *8*, 331.

Warwick-Dugdale, J., Solonenko, N., Moore, K., Chittick, L., Gregory, A.C., Allen, M.J., Sullivan, M.B., and Temperton, B. (2019). Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. PeerJ *7*, e6800.

Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., and Au, K.F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Res *6*.

Wen, D., and Nakhleh, L. (2018). Coestimating Reticulate Phylogenies and Gene Trees from Multilocus Sequence Data. Syst Biol *67*, 439-457.

Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol *18*, 691-699.

Williams, B.A., Hirt, R.P., Lucocq, J.M., and Embley, T.M. (2002). A mitochondrial remnant in the microsporidian Trachipleistophora hominis. Nature *418*, 865-869.

Williams, T.A., and Embley, T.M. (2014). Archaeal "dark matter" and the origin of eukaryotes. Genome Biol Evol *6*, 474-481.

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*, 231-236.

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. Proc Biol Sci *279*, 4870-4879.

Williams, T.A., Szollosi, G.J., Spang, A., Foster, P.G., Heaps, S.E., Boussau, B., Ettema, T.J.G., and Embley, T.M. (2017a). Integrative modeling of gene and genome evolution roots the archaeal tree of life. Proc Natl Acad Sci U S A *114*, E4602-E4611.

Williams, T.A., Szöllősi, G.J., Spang, A., Foster, P.G., Heaps, S.E., Boussau, B., Ettema, T.J.G., and Embley, T.M. (2017b). Integrative modeling of gene and genome evolution roots the archaeal tree of life. Proc Natl Acad Sci U S A *114*, E4602-E4611.

Woese, C.R. (1977). A comment on methanogenic bacteria and the primitive ecology. Journal of molecular evolution *9*, 369-371.

Woese, C.R. (1987). Bacterial evolution. Microbiol Rev *51*, 221-271.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A *74*, 5088-5090.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990a). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A *87*, 4576-4579.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990b). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A *87*, 4576-4579.

Wong, K.M., Suchard, M.A., and Huelsenbeck, J.P. (2008). Alignment Uncertainty and Genomic Analysis. Science *319*, 473-476.

Wu, Y.W., Simmons, B.A., and Singer, S.W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics *32*, 605-607.

Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G.J., and Woese, C.R. (1985). Mitochondrial origins. Proc Natl Acad Sci U S A *82*, 4443-4447.
Yang, Z. (1994a). Estimating the pattern of nucleotide substitution. J Mol Evol *39*, 105-111.
Yang, Z. (1994b). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol *39*, 306-314.
Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. Genetics *139*, 993-1005.
Yang, Z. (2014). Molecular Evolution: A Statistical Approach (OUP Oxford).
Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F.O., Ludwig, W., Schleifer, K.H., Whitman, W.B., Euzeby, J., Amann, R., and Rossello-Mora, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol *12*, 635-645.
Yutin, N., Makarova, K.S., Mekhedov, S.L., Wolf, Y.I., and Koonin, E.V. (2008). The deep archaeal roots of eukaryotes. Mol Biol Evol *25*, 1619-1630.
Zachar, I., and Szathmáry, E. (2017). Breath-giving cooperation: critical review of origin of mitochondria hypotheses : Major unanswered questions point to the importance of early ecology. Biol Direct *12*, 19.
Zillig, W., Stetter, K.O., and Janekovic, D. (1979). DNA-dependent RNA polymerase from the archaebacterium Sulfolobus acidocaldarius. Eur J Biochem *96*, 597-604.
Zuckerkandl, E., and Pauling, L. (1965). Molecules as documents of evolutionary history. J Theor Biol *8*, 357-366.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1861

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-393710

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019