Evolutionary Genomics of Dominant Bacterial and Archaeal Lineages in the Ocean

Carolina Alejandra Martinez Gutierrez

Thesis submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Biological Sciences

Committee
Frank O. Aylward, Chair
Lenwood S. Heath
Josef C. Uyeda
Zhaomin Yang

November 15, 2022
Blacksburg, Virginia

Keywords: Genome Evolution, Evolutionary Genomics, Genome Streamlining, Prokaryotic Tree of Life, Marine Microbial Diversification

Evolutionary Genomics of Dominant Bacterial and Archaeal Lineages in the Ocean

By Carolina Alejandra Martinez Gutierrez

Under the supervision of Professor Frank O. Aylward
at the Virginia Polytechnic Institute and State University

ABSTRACT

The ocean plays essential roles in Earth's biochemistry. Most of the nutrient transformations that fuel trophic webs in the ocean are mediated by microorganisms. The extent of phylogenetic and metabolic diversity of key culture and uncultured marine microbial clades started to be revealed due to progress in sequencing technologies, however we still lack a comprehensive understanding of the evolutionary processes that led to the microbial diversity we see in the ocean today. In this dissertation, I apply phylogenomic and comparative genomic methods to explore the evolutionary genomics of bacterial and archaeal clades that are relevant due to their abundance and biogeochemical activities in the ocean. In Chapter 1, I review relevant literature regarding the evolutionary genomics of marine bacteria and archaea, with emphasis on the origins of marine microbial diversity and the evolution of genome architecture. In Chapter 2, I use a comparative framework to get insights into the evolutionary forces driving genome streamlining in the *Ca.* Marinimicrobia, a clade widely distributed in the ocean. This project shows that differences in the environmental conditions found along the water column led to contrasting mechanisms of evolution and ultimately genome architectures. In Chapter 3, I assess the phylogenetic signal and congruence of marker genes commonly used for phylogenetic studies of bacteria and archaea and propose a pipeline and a set of genes that provide a robust phylogenetic signal for the reconstruction of multi-domain phylogenies. In Chapter 4, I apply a phylogeny-based statistical approach to evaluate how tightly genome size in bacteria and archaea is linked to evolutionary

history, including marine clades. I present evidence suggesting that phylogenetic history and environmental complexity are strong drivers of genome size in prokaryotes. Lastly, in Chapter 5, I estimate the emergence time of marine bacterial and archaeal clades in the context of the Prokaryotic Tree of Life and demonstrate that the diversification of these groups is linked to the three main oxygenation periods occurring throughout Earth's history. I also identify the metabolic novelties that likely led to the colonization of marine realms. Here I present methodological frameworks in the fields of comparative genomics and phylogenomics to study the evolution of marine microbial diversity and show evidence suggesting that the main evolutionary processes leading to the extant diversity seen in the ocean today are intimately linked to geological and biological innovations occurring throughout Earth's history.

Evolutionary Genomics of Dominant Bacterial and Archaeal Lineages in the Ocean

Carolina Alejandra Martinez Gutierrez

Bajo la supervisión del Profesor Frank O. Aylward
En el Virginia Polytechnic Institute and State University

RESUMEN

El océano juega un rol esencial en la biogeoquímica de nuestro Planeta. La mayoría de las transformaciones de nutrientes que controlan las redes tróficas en el océano son mediadas por microorganismos. La diversidad filogenética y metabólica de microorganismos marinos clave cultivados y no cultivados ha empezado a ser revelada gracias al progreso de tecnologías de secuenciación, sin embargo aún desconocemos los procesos evolutivos que han llevado a la diversidad microbiana que vemos en el océano. En esta tesis, aplique métodos filogenéticos y de genómica comparativa para explorar la evolución de clados de bacterias y arqueas que son relevantes en el océano debido a su abundancia y actividades biogeoquímicas. En el Capítulo 1, realice una revisión de la literatura más reciente y relevante en el campo de la evolución genómica de bacterias y arqueas marinas, con énfasis en los orígenes de su diversidad y la evolución de la arquitectura genómica. En el Capítulo 2, uso un marco comparativo para explorar las fuerzas evolutivas que han llevado al "*genome streamlining*" de *Ca.* Maninimicrobia, un clado ampliamente distribuido en el océano. Este proyecto muestra que cambios en las condiciones ambientales a lo largo de la columna de agua llevan a diferencias en los mecanismos de evolución y por tanto de la estructura del genoma. En el Capítulo 3, mido la congruencia y señal filogenética de marcadores moleculares comúnmente utilizados para estudios filogenéticos de bacterias y arqueas, proponiendo una estructura metodológica y un set de genes que proveen una señal filogenética robusta para la reconstrucción de filogenias de múltiples dominios. En el Capítulo 4,

aplico un método estadístico basado en relaciones filogenéticas para evaluar si la distribución del tamaño del genoma en bacterias y arqueas está ligado a su historia evolutiva, incluyendo clados marinos. En este estudio presento evidencia que sugiere que la historia filogenética y la complejidad ambiental son importantes predictores de la distribución del tamaño del genoma en procariotas. Finalmente, en el Capítulo 5 estimo la edad de la diversificación de clados de bacterias y arqueas marinos en el contexto del Árbol de la Vida, y demuestro que estos procesos de diversificación se pueden ligar a los tres principales periodos de oxigenación que han ocurrido a lo largo de la historia del Planeta. También identifico las novedades metabólicas que probablemente permitieron la colonización de estos clados a ambientes marinos. En este estudio presento marcos metodológicos en los campos de la genómica comparativa y filogenómica para el estudio de la diversidad microbiana marina, y muestro evidencia que sugiere que los principales procesos evolutivos que llevaron a la diversidad microbiana en el océano están íntimamente ligados a innovaciones geológicas y biológicas que han ocurrido en la historia de nuestro Planeta.

Evolutionary Genomics of Dominant Bacterial and Archaeal Lineages in the Ocean

By Carolina Alejandra Martinez Gutierrez

Under the supervision of Professor Frank O. Aylward
at the Virginia Polytechnic Institute and State University

GENERAL AUDIENCE ABSTRACT

The ocean plays essential roles in the functioning of our planet. Many of the nutrient's transformation happening in marine environments are mediated by microorganisms, whose metabolic activities underpin higher trophic levels. The identity of the most prevalent marine microbial groups has been reveled during the last two decades through sequencing technologies. Despite having a great progress in our understanding of the functions that these microorganisms have in the ocean; we still lack information about the evolutionary processes that allowed their diversification and colonization into marine realms. In this work, I developed and applied computational strategies to disentangle the evolutionary genomics of marine microorganisms. One particularity about most these marine groups is that they have very small genomes. To explore the evolutionary forces driving their genome reduction, I analyzed a broad set of genomes of Marinimicrobia, a bacterial group widely distributed in the ocean. This analysis shows that genome reduction in Marinimicrobia is driven by negative selection, an evolutionary force that allows the deletion of non-essential genes, leading to genome reduction. Moreover, I developed a benchmarked pipeline for the reconstruction of phylogenetic trees to study the evolutionary relationships of microorganisms. This pipeline allowed me to link the diversification of the main marine groups and the geological periods in which they first emerged. I discovered that the colonization of these groups happened during three different periods, which are coincident with the main oxygenation events occurring across Earth's history. Moreover, the diversification of

marine microbial groups was associated with the acquisition of genes to exploit the newly created niches that followed the oxygenation of the atmosphere and the ocean. Overall, my work shows that the diversification of the marine microbial clades that are essential for the functioning of the ocean today is intimately linked to the redox state of the ocean and the atmosphere throughout Earth's history.

DEDICATION

To my parents, Ana I. Gutierrez and Jose A. Martinez

Para mis padres, Ana I. Gutierrez y Jose A. Martinez

ACKNOWLEDGEMENTS

The scientific contributions presented in this document would not have been possible without the support of many people. First, I want to thank my advisor, Prof. Frank O. Aylward. I am very grateful for the opportunity of working in the lab and being advised by such a wonderful person and scientist. This has been a fun and enriching experience in all aspects. Thank you so much! And of course, a big thank you to all my supportive and amazing labmates and friends! The old ones: Mohammad Moniruzzaman (Monir), Nitin Nair, Ahn Ha, Sangita Karki, Claudia Perez, and Josh Stanton, as well as the new ones: Roxanna Farzad, Paula Erazo, Uri Sheyn, Rubayet Alam, and Abdeali Jivaji. I would like to especially thank my friend Alaina Weinheimer; having you as a labmate and friend throughout this journey has been great!

I would like to thank my committee for their support and guidance during my Ph.D. I would like to thank Prof. Lenwood S. Heath for teaching me Python, Prof. Zhaomin Yang for that amazing course in which I learned about Microbial Genomics, and Prof. Josef Uyeda for transmitting me his knowledge about phylogenomics. Thanks so much for your support!

Thank you to Virginia Tech and the Biological Sciences Department for the opportunity of being part of their Graduate Program. I would also like to thank all the professors that enriched my knowledge during the courses I took. Navigating through graduate school would not have been possible without the help of Rebecca Zimmerman and Valerie Sutherland, thanks so much for your kind support. I want to thank the funding agencies that made my research possible: The Simons Foundation, the Institute for Critical Technology and Applied Science, and the National Science Foundation. Thank you so much for supporting the study of the evolution of marine microbes.

The process of leaving home would have been so much harder without the support of my incredible friends in Blacksburg, especially of those who were there when I first came. My friends Amrita Chakraborty and Rathsara Herath, thanks so much for everything! Also, my friends Camilo Alfonso Cuta, Karla Tellez, Matthew Flores, and many others.

Thanks so much to my family in Mexico: my parents Ana Gutierrez and Jose Martinez, as well as my siblings Aremy Altamirano and Carlos Martinez for their support despite the distance. Muchas gracias por todo su amor!

Lastly but not least, I would like to thank my loving husband Paul Oehlmann for his support, company, and encouragement throughout this journey.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

LIST OF TABLES

Chapter 3

Chapter 4

Chapter 5

Chapter 1. Literature Review

Evolutionary Genomics of Marine Bacteria and Archaea

Previously published: Martinez-Gutierrez CA and Aylward FO. 2022. Evolutionary Genomics of Marine Bacteria and Archaea. In Stal J.L. and Cretoiu MS (eds) The Marine Microbiome: An Untapped Source of Biodiversity and Biotechnological Potential, Second Edition. Springer 327-354.

Co-authors contributed in the following ways: Conceived and designed this work: CAMG and FOA. Wrote the paper: CAMG and FOA.

1.1 Abstract

The ocean harbors an enormous diversity of bacteria and archaea whose metabolic activities have global biogeochemical impacts. Evolutionary genomic studies play an important role in marine microbiology by providing insights into the selective pressures that influence the diversity and functioning of marine microbial lineages, shed light on their niche specialization, and better define their broader ecological roles. The scope of evolutionary genomic studies in the ocean has dramatically expanded in recent years owing to advances in DNA sequencing technology, metagenomics, and single-cell sequencing. Historically the collection of sequenced genomes has been composed of a small number of cultured microbes, but recent advances have opened a window into the genomics of a broad diversity of lineages throughout the biosphere. This broad genomic representation of lineages across the tree of life has enabled comparative genomic analyses that help to identify the ecological and evolutionary forces that drive the diversification of microorganisms in the ocean. In this chapter we review some of the salient themes that have emerged in the last few decades of these studies. We discuss the main findings of these

evolutionary genomic studies and their implications for our understanding of the diversity and functioning of microbial life in the ocean.

1.2 Introduction

The ocean plays a central role in Earth's biogeochemistry (Falkowski, 1998; Falkowski et al., 2008). Due to their overwhelming abundance and high activity, microorganisms mediate the majority of nutrient transformations in the marine environment and are therefore key engines of Earth's biogeochemical cycles (Arrigo, 2005; Brown et al., 2014; Bunse and Pinhassi, 2017). A major fraction of the oxygen produced globally is the result of the activity of marine phototrophs (Kasting, 2002), and higher trophic levels derive their energy and nutrients thanks to the activities of marine microorganisms (Azam, 2004; Brown et al., 2014; Falkowski et al., 2008; Pomeroy et al., 2007).

Despite their important roles, marine microorganisms were not always recognized for their influence on global processes. Because they are not readily visible, progress in understanding ocean microbes has often depended on the development of methods that allowed for their examination directly in the environment (Sherr and Sherr, 2008). Before sequence-based surveys of microbial diversity became common practice, most of the research that aimed at understanding microbial evolution and diversity was based on the study of a few cultured microorganisms and relied on methods derived from clinical microbiology (Salazar and Sunagawa, 2017; Sherr and Sherr, 2008). Early studies of microbial population genetics were based on culture-dependent methodologies such as electrophoretic profiles of well-studied cultured microbes and rarely included marine microorganisms (Witzel, 1990). Throughout the mid- to late 20th Century it was recognized that the number of microbial colonies that grew on solid agar media when plating environmental samples was significantly lower than the number of cells observed through direct

microscopic counts (Staley and Konopka, 1985). Now known as the "Great Plate Count Anomaly", this discrepancy is often used to emphasize the importance of culture-independent methods for studying microbial life in the biosphere, and especially the ocean (Salazar and Sunagawa, 2017).

In recent years the study of microbial communities has been aided by advances in DNA sequencing. This has subsequently led to breakthroughs in marker gene surveys of microbial diversity, community metagenomics, and single-cell genomics (Lasken and McLean, 2014; Moran, 2008; Tringe and Hugenholtz, 2008). It is now possible to routinely recover nearly complete microbial genomes directly from environmental samples (Sharon and Banfield, 2013). These approaches have generated a plethora of new data that are publicly available, which allow for the analysis of the phylogenetic diversity, genome evolution, and physiology of widespread marine microbial groups (Sherr and Sherr, 2008). Progress in the evolutionary genomics of marine bacteria and archaea has often been hampered by the lack of sequenced genomes because it relies on genomic comparisons within and between different lineages (Heidelberg et al., 2010). Hence, these newly available genomic data have spurred numerous insights into the how and what of marine microbial diversity.

Rather than attempting to provide a review of the enormous number of genomic studies of marine bacteria and archaea that have been published to date, we have instead focused in this chapter on five major themes that highlight important insights in marine evolutionary genomics. These are: 1) the origin of genomic diversity in marine bacteria and archaea, 2) genome streamlining, 3) ecological factors that determine genome composition, 4) genomics of the dark ocean, and 5) virus-host interactions as drivers of genome evolution. Much of the insight into marine microbial genomics to date has been garnered through the analysis of certain cosmopolitan marine clades that are particularly dominant in the ocean and distributed across the tree of life (Fig.

1.1). We provide particular emphasis on these clades, although recent studies have begun to shed light on numerous other lineages that will likely be an important emphasis of future work in this field (Yilmaz et al., 2015).

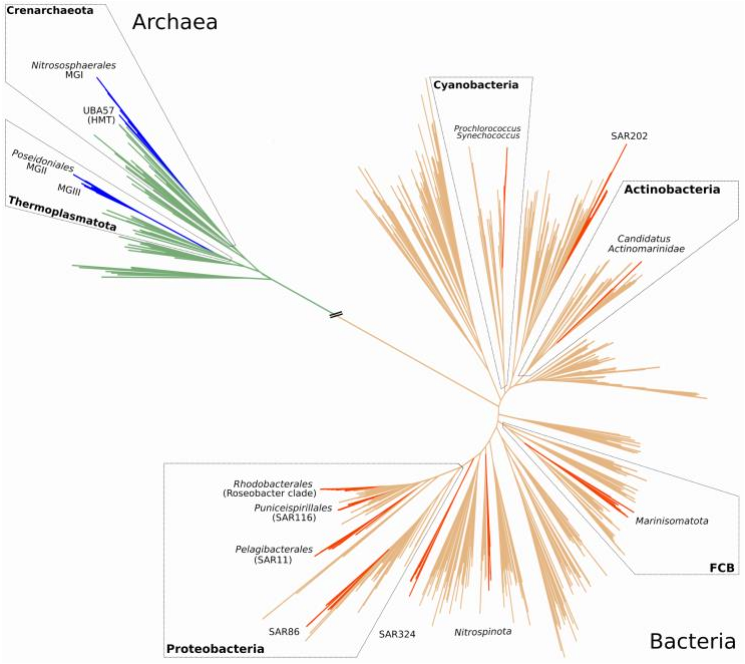

Figure 1.1 Phylogenetic relationships of the major planktonic archaeal and bacterial clades Phylogenetic tree reconstructed based on the representative genomes from the Genome Taxonomy Database (GTDB; v89) (Parks et al., 2018) using a Maximum Likelihood approach and a concatenated alignment of 30 ribosomal proteins and RNA polymerase subunits. Taxonomy shown is based on the GTDB. Abbreviations: MG; Marine Group, HMT; heterotrophic marine Thaumarchaeota, FCB; superphylum named after the phyla Fibrobacteres, Chlorobi, and Bacteroidetes. Green and orange branches represent Archaeal and Bacterial clades, respectively, and dark branches the major marine groups.

1.3 The origins of genomic diversity in marine microbial populations

Microorganisms inhabiting the ocean harbor an enormous functional and phylogenetic diversity (DeLong, 1997; Eguíluz et al., 2019). The microevolutionary processes that lead to the origin and maintenance of this diversity are driven by the interplay of four primary evolutionary forces: mutation, recombination or horizontal gene transfer (HGT), genetic drift, and selection (Feil, 2004;

Futuyma, 1986; Lynch, 2007). HGT is particularly frequent in bacteria and archaea, and it can occur even between distantly-related groups, thereby playing a central role in the acquisition of novel traits and the overall genome evolution of microbial lineages (Bansal et al., 2011; Gogarten et al., 2005). High levels of HGT can blur population boundaries and thereby complicate application of the species concept to bacteria and archaea (Rosselló-Mora and Amann, 2001). Therefore, ecological populations or "ecotypes" are usually used to distinguish between groups of individuals that are clustered based on genotypic and phenotypic criteria as well as ecological similarities (Cohan, 2006; Cohan and Perry, 2007; Cordero and Polz, 2014). For simplicity, in this chapter we will refer to these clusters as populations.

New genes can be introduced into a population through the duplication of existing genes or the introduction of exogenous genes via HGT. In the case of gene duplication, new paralogs can subsequently diverge from the original sequence and evolve new functions (i.e., neofunctionalization), thereby generating genomic novelty (Kunin and Ouzounis, 2003; McDaniel et al., 2010; Snel et al., 2002). Although genes acquired through duplication may display an important role for the survival of microorganisms under changing environmental conditions (Sanchez-Perez et al., 2008), paralogs are often not retained simply because newly acquired gene copies are quickly eliminated before they evolve a new function (Kirchberger et al., 2020; Mira et al., 2001). This is especially true for marine lineages that exhibit genome streamlining and thereby have few paralogs in their genome (Giovannoni et al., 2014). Evidence therefore suggests that HGT is generally a more prominent force introducing novelty into bacterial and archaeal genomes (Koonin et al., 2001; Treangen and Rocha, 2011).

HGT is a particularly important force that drives microbial diversification in marine systems (Sobecky and Hazen, 2009). In addition to the well-studied routes of HGT such as those

mediated by plasmids, temperate viruses, and other mobile genetic elements, Gene Transfer Agents (GTAs) and extracellular vesicles are ubiquitous in marine systems and may constitute an important and previously unrecognized route of gene exchange (Biller et al., 2014; McDaniel et al., 2010). GTAs are virus-derived gene clusters that package diverse DNA fragments and can potentially transfer them between disparate lineages (Lang et al., 2017). They have been found in diverse bacterial and archaeal lineages, but the ability to transfer DNA has only been confirmed for a few cases (Lang et al., 2017; Shakya et al., 2017). Extracellular vesicles are another potential means of HGT; these vesicles are produced by a wide variety of marine bacteria and archaea (Deatherage and Cookson, 2012) and contain diverse DNA fragments, although the mechanism of DNA incorporation remains unclear (Biller et al., 2014, 2017). One study found that most vesicles did not contain microscopically detectable DNA when using fluorescent stains, suggesting that only a few vesicles may contain large quantities of DNA (Biller et al., 2017). Nonetheless, extracellular vesicles have been shown to facilitate HGT on some occasions (Klieve et al., 2005; Renelli et al., 2004; Yaron et al., 2000). Together with GTAs, extracellular vesicles comprise a potentially important route of gene exchange in marine systems.

HGT plays a prominent role in influencing the composition of bacterial and archaeal genomes over long timescales, and it has likely promoted the diversification of many marine lineages by allowing them to occupy their current niches. Prevalent HGT was found in pelagic ammonia oxidizing archaea (Marine Group I: MGI) and Euryarchaeota (Marine Groups II and III: MGII and MGIII, respectively), where a fosmid-based analysis has shown that up to 25% of the genes analyzed may have been acquired through HGT (Brochier-Armanet et al., 2011). Most of the putatively transferred genes are predicted to be involved in processes such as energy metabolism and transport of metabolites across membranes (Brochier-Armanet et al., 2011;

Deschamps et al., 2014). Another study showed that the transfer of genes from bacteria to ammonia oxidizing archaea (AOA) may have played a role in the transition of this group from terrestrial environments to the ocean (Ren et al., 2019). The genes analyzed here may have facilitated the adaptation to different environmental conditions along the water column. In another example, a genomic analysis of the nitrite oxidizer *Nitrospina gracilis* revealed that a large fraction of its encoded proteins show homology with *Deltaproteobacteria, Gammaproteobacteria,* and *Nitrospira,* suggesting frequent HGT among these groups. Genes involved in anaerobic nitrite oxidation in *N. gracilis* were probably obtained from anaerobic ammonia oxidizers (Lücker et al., 2013). This suggests that the capability to oxidize nitrite was acquired under anaerobic conditions and subsequently distributed to other areas of the ocean. Lastly, comparative genomic analyses of *Ca.* Marinimicrobia, a candidate phylum comprising lineages specialized in epipelagic or mesopelagic environments, has revealed a central role of HGT in the diversification of this marine candidatus phylum. Distinct marinimicrobial clades that inhabited the same environment had convergently acquired similar genomic repertoires, indicating that parallel HGT events had played a role in their niche partitioning (Getz et al., 2018).

Although bacteria and archaea can acquire genes from different sources, their genomes tend to remain small and compact over time (Bobay and Ochman, 2017). Mutations in the genomes of bacteria and archaea are biased towards deletions (Mira et al., 2001). Thus, newly acquired genes resulting from duplications or HGT are typically fixed and retained in populations only if they provide a fitness benefit, whereas genes that do not represent a significant advantage are expected to be lost or pseudogenized within a few generations. The fraction of successfully incorporated genes due to HGT or duplications is therefore low (Batut et al., 2014; Kuo and Ochman, 2009). The overall effect of this deletional bias is related to the strength of selection

relative to genetic drift, which is ultimately determined by the effective population size (N_e) (Box 1) (Bobay and Ochman, 2017). Given their broad distribution, most marine bacteria and archaea have extremely large N_e when compared with multicellular life, and thus selection can be considered a strong evolutionary force acting on their populations (Brockhurst et al., 2019). This is in stark contrast to organisms with a small N_e, such as obligate parasites and endosymbionts, where selection is weak relative to genetic drift. In these cases, deleterious mutations may reach fixation through stochastic processes (Fig. 1.2) (Mira et al., 2001).

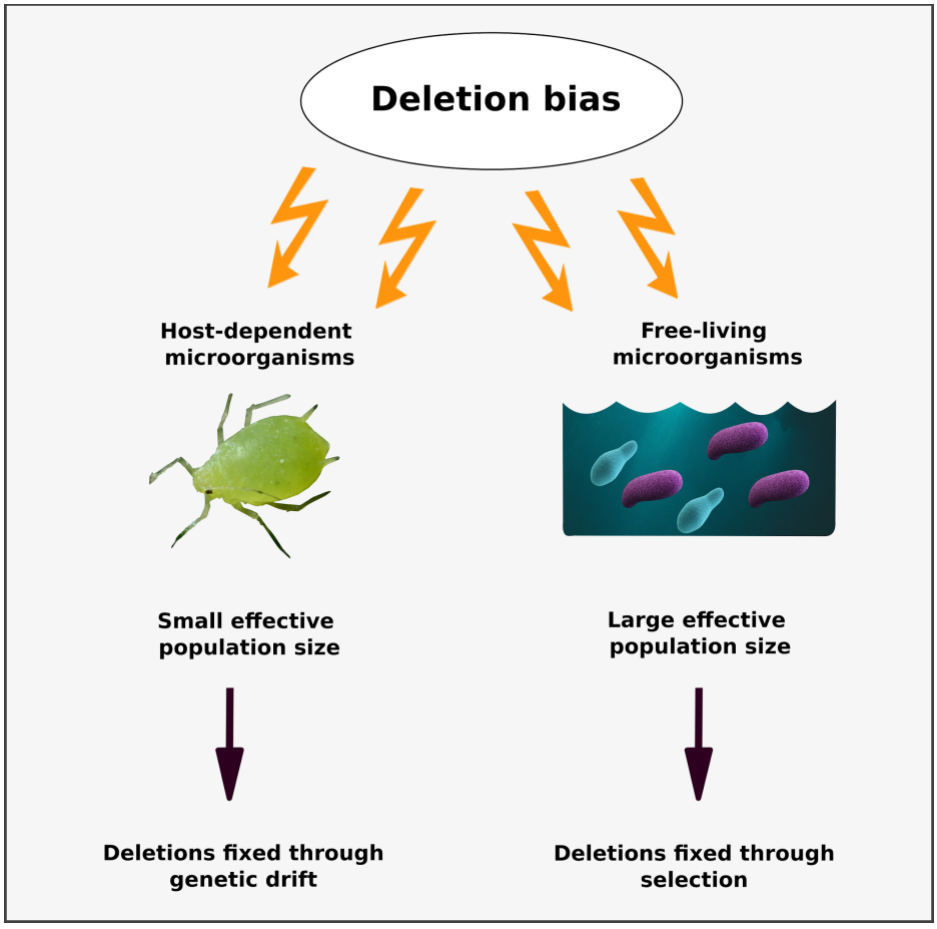

Figure 1.2 Distinct evolutionary paths driving genome reduction in marine vs endosymbiotic bacteria. Adapted from Giovannoni et al., (Giovannoni et al., 2014).

1.4 Streamlining: genome simplification in the open ocean

Many bacteria and archaea that live in the surface waters of the ocean are adapted to the characteristic low-nutrient state of these epipelagic environments (Azam and Malfatti, 2007; Giovannoni and Stingl, 2005). The genomes of many epipelagic microorganisms share some peculiar features including small size (typically < 2 Mbp), short intergenic regions, low %GC content, and few paralogous genes. The evolution of these features is thought to be driven by a process of genome streamlining, whereby natural selection in nutrient-depleted environments favors cellular economization leading to small genomes with little extraneous coding material and simple gene regulation (Giovannoni et al., 2014). Although streamlining was first noted in well-studied cultured lineages like *Prochlorococcus marinus* and *Pelagibacter ubique,* these genomic features were soon found in other abundant marine groups such as *Oceanospirillales, Euryarchaeota, Thaumarchaeota, Ca.* Marinimicrobia*, Puniceispirillales,* the Roseobacter clade, and marine members of the *Actinobacteria*, *Sphingomonadaceae*, and *Dadabacteria* (Dupont et al., 2012; Getz et al., 2018; Ghai et al., 2013; Graham and Tully, 2020; Lauro et al., 2009; Luo et al., 2014b; Martin-Cuadrado et al., 2015; Orellana et al., 2019; Santoro et al., 2015; Swan et al., 2013). These comparative genomics studies have demonstrated that streamlined genomic features have evolved independently in a wide variety of marine lineages, providing a remarkable example of inter domain convergent evolution. Some clades, such as the *Pelagibacterales* appear to contain exclusively streamlined genomes, while others, such as the Roseobacter and *Ca.* Marinimicrobia groups, consist of a mixture of both streamlined and non-streamlined representatives (Getz et al., 2018; Giovannoni, 2017; Luo et al., 2014b).

There has been some debate over the drivers of genome streamlining in marine bacteria and archaea, especially regarding the question whether all genomic features resulting from this

process can be viewed as purely adaptive (reviewed in (Batut et al., 2014)). This debate has been motivated in part by the observation that many parasitic and endosymbiotic bacteria exhibit genomic features in some ways akin to streamlined genomes, including small size and low %GC content. The high level of genetic drift experienced by endosymbionts and parasites is a consequence of their small effective population size (Box 1), which is driven in part by the extreme population bottlenecks experienced during transmission (Moran, 1996). This leads to the fixation of deleterious mutations, including the loss of genes that provide an adaptive benefit and an overall pattern of genomic erosion (Fig. 1.2). The loss of functional genes observed in host-dependent microorganisms often does not occur in parallel with other features indicative of streamlining; the reduced genomes of host-dependent microorganisms is therefore not the product of adaptive processes (Wolf and Koonin, 2013). Marine bacteria and archaea span vast ocean basins and have enormous effective population sizes, and thus genome simplification in this group of microorganisms is unlikely due to high genetic drift (Wolf and Koonin, 2013). Although some work has shown that weakly deleterious mutations and low rates of recombination may lower the effective population size of bacteria compared to their actual population size (Price and Arkin, 2015), it is unlikely that this decreases N_e to a size where genetic drift is the dominant force in genome evolution.

Indeed, several studies have indicated that streamlined features of marine genomes can at least in part be attributed to adaptation. One study concluded that low %GC content in SAR11 is maintained by strong purifying selection (Luo et al., 2015), while another study showed that streamlined *Ca.* Marinimicrobia experiences higher levels of purifying selection compared to their non-streamlined relatives (Martinez-Gutierrez and Aylward, 2019). A similar result was found between high-light adapted *Prochlorococcus* and its close relative *Sychechococcus* (Hu and

Blanchard, 2009). This indicates that the inherent deletional bias of bacterial genomes causes genes that are not essential to undergo relaxed selection and subsequent erosion and deletion from the population, while genes that provide a selective benefit are retained (Bobay and Ochman, 2017; Mira et al., 2001). Hence, genome streamlining is expected to be prevalent in stable environments such as oceanic gyres where environmental conditions and nutrient concentrations do not change substantially throughout the year. The loss of costly regulatory machinery used to respond to changing conditions would not be selected against in such environments (Giovannoni et al., 2014).

Although genetic drift in marine bacterial and archaeal populations is unlikely to give rise to genome streamlining, there are two other non-adaptive processes that could potentially give rise to some of these genomic features: population bottlenecks that occurred in the distant past and evolutionary constraints that may cause traits to covary. Regarding population bottlenecks, some studies have suggested that population reduction events in the distant past may have contributed to the current features of streamlined genomes irrespective of current selective pressures (Luo et al., 2017; Luo, Swan, et al., 2014b). For example, one study evaluating radical substitutions found evidence for population bottlenecks in the early evolution of SAR11 and *Prochlorococcus* (Luo et al., 2017). Adaptive radiations are often accompanied by relaxed selection, and it is therefore plausible that many marine lineages experienced higher levels of genetic drift and, hence, more non-adaptive deletions during the early stages of their radiation into the ocean. The inference of selective pressures in ancient adaptive radiations is tantalizingly difficult to ascertain, however, and some studies have come to the opposite conclusion that early genome reduction in picocyanobacteria was adaptive (Sun and Blanchard, 2014). Regardless, the hypothesis of high genetic drift during early marine radiations need not contradict the adaptive nature of many streamlined genomic features: it is also possible that many such features evolved during a period

of relaxed selection but were subsequently selected for as effective population sizes increased and purifying selection strengthened.

Another complication with ascertaining the adaptive benefit of streamlined genomic features is that many of them covary and may therefore be the product of evolutionary constraints rather than *bona fide* adaptations. One example of linked genomic traits is %GC content and the nitrogen content of encoded amino acids (often referred to as the N-ARSC: Nitrogen Atoms per Residue Side Chain). Codons with low %GC content preferentially code for amino acids with low nitrogen content and selection for either %GC content or protein nitrogen content will therefore alter the other (Bragg and Hyder, 2004). Several studies have shown that both N-ARSC and %GC content are significantly lower in microbial communities that reside in N-depleted waters. One study demonstrated that these features were significantly lower in pelagic compared to coastal marine microbial communities (Grzymski and Dussaq, 2012) and another study showed that N-ARSC and %GC were significantly lower in surface water communities compared to those sampled below the nutricline in the North Pacific Subtropical Gyre (Mende et al., 2017). G-C base-pairs contain one more nitrogen than A-T pairs, and a transition to low %GC content genomes would therefore also lead to nitrogen savings. It is therefore difficult to disentangle the selective pressures acting on these traits. Given that proteins make up a larger portion of the nitrogen content of cells it is plausible that the primarily selective force acting on N-depleted microbial populations is a lowering of the N-ARSC, thereby leading to the decrease of %GC content as an evolutionary by-product. In support of this view, it was shown that *Prochlorococcus marinus* shifts its transcriptome during N depletion to produce proteins with a lower N content (Read et al., 2017). Since *Prochlorococcus* is already highly streamlined, this shift in transcription can be considered

as evidence for an additional selective pressure to lower the proteome N content during nitrogen limitation.

Lastly, it has been postulated that interactions between sympatric microbes may be a driver of reductive genome evolution. The Black Queen Hypothesis (BQH) (Morris et al., 2012) posits that gene losses in free-living marine microorganisms may lead to a dependence on co-existing microbes to substitute for lost metabolic capabilities. The BQH proposes that under nutrient-limited conditions, as is usually the case in the open ocean, streamlined microorganisms may lose functions when these are provided by other microorganisms in the community ("common goods") thereby decreasing the cost of gene maintenance, transcription, and translation. The BQH is supported by the absence of the gene that encodes catalase-peroxidase (*katG)* in *Prochlorococcus*, which is essential for this organism because it removes hydrogen peroxide (HOOH), a by-product of oxygenic photosynthesis that causes oxidative stress. Supposedly, other microorganisms may act as a sink of HOOH, relieving phototrophs like *Prochlorococcus* from the necessity to maintain the genetic potential to protect themselves from oxidative stress (Morris et al., 2012). According to the BQH, positive selection for the loss of redundant genetic material could be an important driver of genome minimization in oligotrophic environments (Mas et al., 2016; Morris et al., 2012).

1.5 Ecological factors influencing genome composition

The ecological niche of a microbe plays a central role in determining its genomics, both in terms of gene content and overall genome composition. Given the ubiquity of HGT and the fast rates at which microbial populations can evolve, closely related marine microorganisms that inhabit different niches often harbor substantially different gene complements that provide niche-specific adaptations. Conversely, distantly related microbes inhabiting the same environment can sometimes evolve convergently and thereby acquire similar genomic traits. By examining both

trends of niche partitioning and parallel evolution it is often possible to gain valuable insight into the ecological factors and selective pressures that a microbial population may experience.

Perhaps the most dramatic example of convergent evolution of gene content involves proteorhodopsin (PR), a light-driven proton pump found in diverse microorganisms in the ocean. First identified in the SAR86 lineage of *Gammaproteobacteria* (Beja, 2000), PRs were subsequently found in diverse epipelagic microbial lineages (Frigaard et al., 2006; Giovannoni et al., 2005; Torre et al., 2003; Yutin and Koonin, 2012). Experimental work has provided evidence that PR aids in survival of microorganisms that contain it and may increase their growth rates at low nutrient concentrations (Gómez-Consarnau et al., 2007, 2010). This likely provides an adaptive benefit when thriving in the prevailing oligotrophic conditions of many marine habitats (DeLong and Béjà, 2010). PR genes appear to be transferred readily across large phylogenetic distances (Frigaard et al., 2006) likely because of the simple modular structure of these gene cassettes, which typically include the PR gene itself and sometimes genes involved in carotenoid biosynthesis (Pinhassi et al., 2016). It has been estimated that between 17-80% of marine bacteria and archaea in the sunlit ocean encode a PR in their genome (DeLong and Béjà, 2010; Moran and Miller, 2007), emphasizing how divergent lineages have repeatedly converged on the same trait presumably because it both provides a sufficient adaptive benefit and it is readily transferred into new genomes without disrupting existing metabolic or regulatory networks.

Another example of genomic changes that occur concomitant with niche specialization is highlighted in marine picocyanobacteria. The genera *Synechococcus* and *Prochlorococcus*, both abundant picocyanobacteria in the ocean, have distinct niche specializations despite their close phylogenetic relationship (Sánchez-Baracaldo et al., 2019). *Synechococcus* dominates in coastal and temperate waters whereas *Prochlorococcus* prevails in warm-oligotrophic waters

(Zwirglmaier et al., 2008). Such niche differentiation results in greater phenotypic flexibility and regulatory capacity in *Synechococcus*, while *Prochlorococcus* exhibits larger population sizes and strong specialization to nutrient depleted environments (Moore et al., 1995; Partensky and Garczarek, 2010; Rocap et al., 2003). *Prochlorococcus* has lost the large extrinsic antenna complexes (phycobilisomes) that are present in *Synechococcus*; instead, *Prochlorococcus* possesses simpler compact antennae within thylakoid membranes that are in direct contact with the photosystems (Berube et al., 2018; Partensky and Garczarek, 2010; Sánchez-Baracaldo et al., 2019). This difference may be responsible for the discrepancy in light absorption properties of both clades, since *Prochlorococcus* cells are specialized for the capture of blue wavelengths that prevail in oligotrophic waters (Ting et al., 2002). In addition to physiological changes, genomic disparities have accompanied the specialization to different habitats in these picocyanobacteria. High-light adapted *Prochlorococcus* ecotypes have undergone a reduction in genome size and %GC content as part of streamlining and adaptation to the oligotrophic environment, while these changes have not been observed in *Synechococcus* (Partensky and Garczarek, 2010; Rocap et al., 2003). *Prochlorococcus* likely experiences higher purifying selection than *Synechococcus* (Hu and Blanchard, 2009; Wolf and Koonin, 2013), which may explain why these genomic trends of genome reduction and lower %GC are more prominent in the former. Different ecotypes partitioned based on light availability are also present within *Prochlorococcus*; the high-light adapted ecotype (HL) is more abundant in surface waters and tends to have small genomes and low %GC content whereas low-light adapted *Prochlorococcus* are dominant in deeper waters and tend to have larger genomes with higher %GC content (Kettler et al., 2007; Rocap et al., 2003). Both ecotypes differ in light absorption properties, and within each ecotype there are groups that are different in terms of growth temperature and abundance (Rocap et al., 2003). Analyses have

shown that the HL-adapted ecotype has likely evolved from deeper waters (Martiny et al., 2009; Partensky and Garczarek, 2010; Paul et al., 2010).

Besides *Prochlorococcus*, many other marine microbial lineages exhibit clear changes in genomic properties that are linked with the depth in which they live. Several studies examining shifts in microbial community composition as well as genomic features across a depth gradient have been conducted at Station ALOHA in the North Pacific Subtropical Gyre (DeLong, 2006; Konstantinidis et al., 2009; Mende et al., 2017). Mende et al. (2017) studied the evolutionary and ecological processes affecting the genomes of bacteria and archaea at Station ALOHA. This study revealed a marked vertical transition in which organisms of the same clade showed a decrease in %GC content above the deep chlorophyll maximum (DCM), as well as reduction in genome size and intergenic spacer length. These genomic traits are therefore correlated with the physicochemical features that vary with depth. Another study at Station ALOHA that compared the microbial communities residing in the surface to those present at 4,000 m suggested that %GC content as well as codon usage changes may be associated with differences in the strength of purifying selection and ultimately disparities in the effective population sizes (Konstantinidis et al., 2009). This suggests that surface and deep water environments are subjected to substantially different evolutionary regimes, both in terms of selective pressures themselves as well as their overall strength. A similar trend was found in a study that compared the codon usage of coastal and open ocean microorganisms (Grzymski and Dussaq, 2012), suggesting that nitrogen limitation may represent a strong selective pressure driving adaptations towards lower %GC content and the minimization of protein synthesis costs in many marine systems.

Genomic changes that occur coincident with ecological transitions across depth have been studied in detail in *Ca.* Marinimicrobia, in which streamlining occurred multiple times

independently in distantly related epipelagic clades. This emphasizes that similar environmental features can drive convergent evolution of genomic features such as low %GC content, short intergenic regions, and low encoded nitrogen content (Getz et al., 2018) (Fig. 1.3). Moreover, epipelagic *Ca.* Marinimicrobia experienced higher levels of purifying selection than their mesopelagic relatives, which is consistent with the strong purifying selection often associated with streamlining (Martinez-Gutierrez and Aylward, 2019). In addition to differences in bulk genomic features, epipelagic and mesopelagic *Ca.* Marinimicrobia also differed in their encoded functional repertoires and cellular bioenergetics. Whereas epipelagic *Ca.* Marinimicrobia often encoded proteorhodopsins and a Na^+-translocating NADH dehydrogenase (NQR), their mesopelagic counterparts acquired a nitrate respiratory machinery, a H^+-translocating NADH dehydrogenase complex (NDH), and different cytochromes possibly involved in the use of alternative electron acceptors (Getz et al., 2018). The repeated acquisition of similar genes by distantly related *Ca.* Marinimicrobia emphasizes the important role of HGT in facilitating parallel niche diversification across large phylogenetic distances.

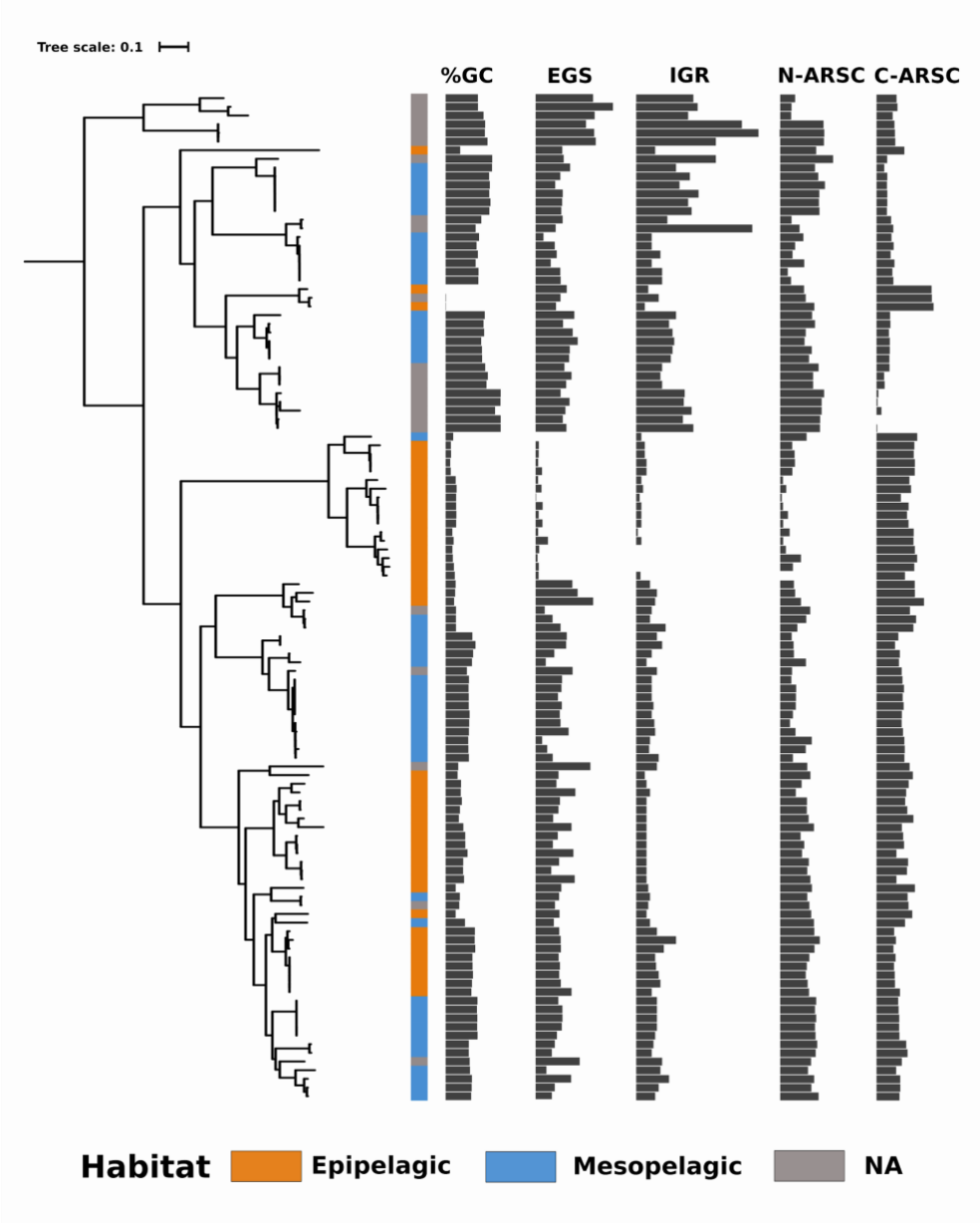

Figure 1.3 Habitat transitions in *Ca.* Marinimicrobia clades. Maximum likelihood phylogenetic tree reconstructed using a concatenated alignment of 30 ribosomal proteins and RNA polymerase subunits. The colored strip shows the habitat for each tip on the tree. Abbreviations: %GC, %GC content (range, 27.16-50.55%); EGS, estimated genome size (range, 1.06-3.53 Mbp); IGR, median intergenic region length (range, 7-78 bp); N-ARSC, Nitrogen atoms per residue side chain (range, 0.3-0.35); C-ARSC, Carbon atoms per residue side chain (range, 2.97-3.20).

Further studies have provided insight into ecologically driven genomic changes involving other habitats. Using a comparative genomic approach, one study reconstructed the genomic consequences of habitat transitions in the *Methylophilaceae* group in the *Betaproteobacteria* (Salcher et al., 2019). This study was able to identify two key transitions in this group: one between freshwater sediment to a pelagic freshwater habitat and another from a pelagic freshwater to a marine habitat. The authors found evidence for a sudden and pronounced loss of genes coincident with the transition from sediment to freshwater pelagic habitat. The subsequent transition from freshwater to marine pelagic environment was not accompanied by further genome reduction, but rather by the acquisition of genes that facilitate adaptation to higher salinities. Another study used a similar approach to reconstruct the genomic changes in *Flavobacteria* that took place coincident with ecological transitions from marine to terrestrial systems (Zhang et al., 2019). This study found similar patterns of gene loss and gain, notably the loss of genes involved in the tolerance of salinity in the ocean. Similar to the *Ca.* Marinimicrobia's example discussed above, this study also found that the presence of either the NQR or NDH dehydrogenase was strongly correlated with the environment in which the *Flavobacteria* were found. These studies demonstrate that divergent microbial lineages will often evolve convergent genomic features depending on their habitat, underscoring how environment-specific selective pressures are a major force that can determine how marine microbial populations evolve.

Although transitions between environments are important factors that determine microbial diversification, it is also possible that ecological transitions drive niche specialization within a single environment. Due to the high dispersibility and large population size of marine bacteria and archaea, sympatric speciation may occur as the result of initial ecological adaptations followed by a decrease of interpopulation gene flow caused by separation of microhabitats (Shapiro et al.,

2012). It is thought that population-specific mutations and the introduction of new genes through HGT trigger the differentiation of populations until they become distinct genotypic clusters. This has been experimentally explored among coastal *Vibrionaceae* strains; here numerous sympatric but ecologically-distinct populations of *Vibrio splendidus* have been found, possibly as a result of adaptations to planktonic vs particle-associated lifestyles (Hunt et al., 2008). Given the large heterogeneity that can be found within small volumes of seawater, it is likely that this route of niche specialization and concomitant sympatric differentiation may be a major force that drives diversification in marine microbes that reside in the same environment.

1.6 Genome evolution in the dark ocean

The dark ocean represents the vast majority of the ocean by volume and comprises a wide variety of marine habitats, including sediments, oceanic crust, cold seeps, and hydrothermal vents (Arístegui et al., 2009; Orcutt et al., 2011). The water column below the photic zone is considered to be an extreme environment due to low temperature, high hydrostatic pressure, and the absence of solar radiation (Orcutt et al., 2011). Moreover, the environmental conditions in the dark ocean vary dramatically in terms of temperature, hydrostatic pressure, and chemical composition (Orcutt et al., 2011) and consequently metabolically diverse bacterial and archaeal lineages reside in this environment. Given the metabolic diversity of microbes that reside in the dark ocean, breakthroughs in genomics have been highly valuable in examining the predicted physiology and biogeochemical roles of bacteria and archaea in this environment. For purposes of this section, we will focus on recent research on several planktonic lineages inhabiting the water column of the dark ocean. Several reviews discuss other deep sea habitats and provide more detail into some of the groups discussed here (Corinaldesi, 2015; Dick, 2019; Lauro and Bartlett, 2008; Orcutt et al., 2011; Orsi, 2018; Santoro et al., 2019).

Several recent studies focusing on heterotrophic lineages of bacteria and archaea in the dark ocean have begun to unveil their genomic and physiological diversity. One study found that the broadly-distributed SAR202 clade in the *Chloroflexi* phylum represents up to 30% of the microbial community in the dark ocean and is broadly distributed in mesopelagic and bathypelagic waters (Mehrshad et al., 2018). This study identified two major clades of SAR202 that were partitioned by depth and had genome sizes ranging from 1.2-3.4 Mbp and %GC content ranging from 41-55%. These authors revealed the capacity to metabolize organosulfur compounds and oxidize sulfite in many of the genomes of SAR202, suggesting that this lineage may play an important role in sulfur cycling in the dark ocean. Other studies examining single-cell genomes belonging to this clade identified numerous catabolic enzymes including monooxygenases, dioxygenases, and racemases that were predicted to target recalcitrant carbon compounds. Paralogous expansion of these catabolic enzymes was linked to SAR202 subclades that reside in different locales, suggesting that duplication of these genes has played an important role in the niche expansion in this group (Landry et al., 2017; Saw et al., 2020). Saw et al. (2020) also recovered proteorhodopsin-encoding genes in SAR202 genomes from surface waters, indicating that this group is not exclusively found in the dark ocean.

The Marine Group II Euryarchaeota (MGII), recently named *Ca.* Poseidoniia, are also abundant constituents of planktonic communities in both the sunlit and dark ocean, where they are thought to be major contributors to organic matter remineralization (Rinke et al., 2019; Zhang et al., 2015). MGII was first discovered in coastal marine environments (DeLong, 1992; Fuhrman et al., 1992), but this group was subsequently found in pelagic waters where it is consistently associated with the deep chlorophyll maximum and below (Santoro et al., 2019). MGII is a group of aerobic heterotrophs with the genomic potential to degrade a broad range of substrates as

evidenced by genes encoding di- and oligo-peptidases as well as enzymes for the degradation of amino acids and fatty acids (Tully, 2019). Members belonging to MGII show a remarkable spatial partitioning in terms of their genomic properties; representatives from surface waters show shorter genomes and harbor photolyase and proteorhodopsin-encoding genes whereas those thriving in the deep ocean encode the potential for flagellar-based adhesion and respiration of nitrate, possibly as an adaptation to oxygen limitation (Rinke et al., 2019; Tully, 2019). The acquisition of diverse genes from different sources has facilitated the niche adaptation of different MGII lineages, underscoring the importance of HGT for the diversification of this group (Rinke et al., 2019; Tully, 2019).

Several studies have noted that genomes of bacteria and archaea from the mesopelagic are generally larger than those from the epipelagic, but there are exceptions to this trend. One notable example is a lineage of heterotrophic marine *Thaumarchaeota* (HMT) that branches within a sister lineage to the ammonia oxidizing archaea (AOA) (Aylward and Santoro, 2020; Reji and Francis, 2020). This group, sometimes also referred to as the psl12-like group, is widespread in the dark ocean and has been found in waters ranging from a depth of 150 m near Monterey Bay (Reji and Francis, 2020) to bathypelagic and abyssopelagic waters of the South Pacific and South Atlantic (Aylward and Santoro, 2020). This HMT group has some genomic characteristics of streamlining including notably small genomes of near 1 Mbp and a low %GC content of 33%. Unlike their AOA relatives, the HMT lack ammonia monooxygenase and hydroxypropionate/hydroxybutyrate carbon fixation pathway. Instead they encode a type iii-a RuBisCO and diverse pyrroloquinoline quinone (PQQ)-dependent dehydrogenases that are highly expressed and likely play a key role in their heterotrophic lifestyle. PQQ-dehydrogenases comprise as much as 3% of the HMT genomes and their prevalence is likely the product of a combination of paralogous expansion and

widespread HGT (Aylward and Santoro, 2020). More features of this lineage are likely to become uncovered in the future given its widespread occurrence throughout the dark ocean.

Due to the low energy flux that characterizes the dark ocean, chemolithoautotrophic microorganisms that derive energy from reduced inorganic compounds are among the most abundant at this depth (Bach et al., 2006; Orcutt et al., 2011). Perhaps the most well-studied example is the Marine Group I (MGI) ammonia-oxidizing archaea, which can comprise up to 40% of the cells in some marine habitats (Karner et al., 2001). Genome sizes of complete MGI genomes in the NCBI database range from 1.23-2.17 Mbp with the smaller genomes exhibiting signatures of genome streamlining (Santoro et al., 2015). Several studies based on single cell genomes found that MGI populations are strongly depth-stratified (Luo et al., 2014; Swan et al., 2014) confirming previous observations based on marker genes (Francis et al., 2005; Hallam et al., 2006; Nicol et al., 2014). A large-scale comparative analysis of the radiation of ammonia oxidizing archaea throughout terrestrial and marine habitats revealed that the genomic repertoires of this group varied with environmental variables such as hydrostatic pressure and the availability of ammonia and phosphorus, suggesting that the genomic content is largely structured by partitioning into spatially-defined niches (Qin et al., 2020). Another study found that the acquisition of a V-type ATPase through HGT was a salient feature in deep water adapted MGI genomes including several that were recovered from hadopelagic waters (Wang et al., 2019). Although the physiological implications of the V-Type ATPase in deep water AOA remain unclear, it is noteworthy that this was also found in some acidophilic AOA that may use it for pumping protons at low pH (Wang et al., 2019).

Another group that contributes to carbon fixation and nitrogen cycling in the deep ocean is the nitrite-oxidizing bacteria (NOB), a polyphyletic group involved in the oxidation of nitrite to

nitrate (Bock and Wagner, 2006). Although most of the NOB representatives only have the metabolic capability to perform the last step of nitrification, *Nitrospira* strains able to oxidize ammonia to nitrate have been isolated (Comammox) (Daims et al., 2015; van Kessel et al., 2015). The most abundant marine members of the NOB belong to the phylum *Nitrospinae*, and it has been estimated that they fix 15 to 45% of the inorganic carbon in the mesopelagic waters of the western North Atlantic and are therefore a major contributor to nitrification in deep waters (Pachiadaki et al., 2017). Analysis of the *Nitrospina gracilis* genome revealed signatures of extensive HGT from distantly-related groups; it is thought that the subunit A of the nitrite oxidoreductase (NxrA) found in *N. gracilis* was obtained from anaerobic ammonium oxidizers (Lücker et al., 2013). This as well as the presence of rTCA cycle for carbon fixation suggests that the *Nitrospinae* may have originated from an anaerobic or microaerophilic environment and subsequently colonized aerobic habitats (Lücker et al., 2013). This is consistent with another study that suggests that oxygen levels played an important role in the diversification of *Nitrospinae* into different marine habitats (Sun et al., 2019).

Lastly, carbon fixation in the dark ocean is also associated with sulfur oxidation as shown by metagenomic reconstructions of the ubiquitous SAR324 clade in the *Deltaproteobacteria* (Swan et al., 2011). In addition to autotrophy, metabolic predictions revealed the presence of genes that encode enzymes involved in the oxidation of carbon monoxide and hydrocarbons, as well as methylotrophy (Brown et al., 2014; Sheik et al., 2014; Swan et al., 2011). Furthermore, genes involved in motility may also play a role in the attachment to particulate organic matter. These observations suggest that the high abundance and ability to inhabit diverse environments observed in SAR324 may be associated with their broad metabolic capabilities (Giovannoni and Vergin, 2012; Swan et al., 2011). Regarding genomic composition, SAR324 genomes tend to be

small (1.4-2.9 Mbp) and harbor a large amount of repetitive sequences and signatures of prophage integration, indicating that interactions with viruses may play an important role in SAR324 genome evolution (Cao et al., 2016).

1.7 Virus-host interactions influencing genome evolution in bacteria and archaea

Viruses are omnipresent biological entities that play a key role in genome evolution and determine the abundance of microbial lineages in the ocean (Breitbart et al., 2018; Wommack et al., 2000). It is estimated that every second 10^{23} viral infections occur in the ocean, killing approximately 20% of the marine microbial biomass per day. Therefore, marine viruses exert a strong selective pressure on their hosts (Suttle, 2007). As a consequence, viruses alter the flow of energy and nutrients and determine the selective pressures experienced by their hosts (Brum et al., 2015; Fuhrman, 1999; Lindell et al., 2007; Suttle, 2007; Zimmerman et al., 2020).

Host evolution is driven by the constant arms race with their associated viruses that lead to counter-adaptations and the increase of diversity within populations (Lindell et al., 2007). This phenomenon is known as the Red Queen hypothesis (Stern and Sorek, 2011). Many studies have reported long-term stability of viral genotypes in the ocean, often across broad geographic distances (Aylward et al., 2017; Breitbart et al., 2004; Breitbart and Rohwer, 2005; Hevroni et al., 2020; Marston and Martiny, 2016; Short and Suttle, 2005); this indicates that viruses co-exist with host populations for long periods in a pervasive arms race dynamics. A long-term analysis performed on marine microbial and viral communities indicated that, although viral composition is stable throughout the year, there is a rise and fall of variants within populations (Ignacio-Espinoza et al., 2020). This is probably the result of the fluctuating selection from virus-host defenses and counter-defenses as proposed by the Red Queen hypothesis. Additionally, experimental evidence indicates that co-occurrence of viruses and hosts speeds up the evolution

of both groups (Thingstad et al., 2014). For example, phages promote microdiversity in cultures of the marine bacteria *Flavobacterium* and *Prochlorococcus* by driving strain succession towards resistance (Avrani et al., 2011; Middelboe et al., 2009). Marine populations under phage pressure are composed of subpopulations with different viral susceptibility. It is thought that the diversity generated by phage-driven selective pressure may facilitate adaptations in fluctuating environments (Williams, 2013). For example, this selective pressure may drive the alteration of polysaccharides, glycoproteins, or outer membrane proteins because these structures can serve as recognition sites for phages. In *Alteromonas* populations, these genes are located in flexible genomic islands notable for their microdiversity (López-Pérez and Rodriguez-Valera, 2016; Rodriguez-Valera et al., 2009).

Viruses mobilize 10^{25}-10^{28} bp of DNA per year in the world's oceans (Sandaa, 2008). They mediate the transfer of genetic material across microbial populations and thereby provide the host with potentially new adaptive traits that may aid subsequent diversification (Brum et al., 2015; Rodriguez-Valera et al., 2009; Sandaa, 2008; Touchon et al., 2017). Viral genomes often contain auxiliary metabolic genes (AMGs) derived from their host (Breitbart et al., 2007; Hurwitz et al., 2013; Roux et al., 2016). Some of the most well-studied examples are T4-like phages that infect *Prochlorococcus* and *Synechococcus* (Lindell et al., 2007; Sullivan et al., 2010). Contigs of T4-like phages recovered from marine metagenomes often contain photosystem I genes (PSI), which are thought to form a monomeric PSI complex that can tunnel reducing power from the electron transport chain of the host to this PSI-related function complex during the infection process, possibly contributing to phage fitness (Sullivan et al., 2010). Fridman et al. (2017) successfully cultivated a *Prochlorococcus*-infecting myovirus that encoded genes involved in both photosystem I and II, emphasizing the complex manipulation of host physiology by viruses during infection.

The evolution of host photosystem genes may also be driven in part by cyanophages because viral genes are able to recombine and transfer back into the host's gene pool (Lindell et al., 2004; Sullivan et al., 2006, 2010). Aside from photosystem genes, a wide variety of other phage-encoded AMGs have been identified, including genes involved in carbon, nitrogen, and sulfur metabolism (Ahlgren et al., 2019; Anantharaman et al., 2014; Hurwitz et al., 2013). Also, ammonia monooxygenase genes (*amoC)* have been identified in globally distributed viruses that infect marine AOA (Ahlgren et al., 2019). Ahlgren et al. (2019) found viral *amoC* expression providing evidence that viruses play a role in nitrogen cycling.

Although phages are considered to be the major cause of death of bacteria and archaea in the ocean, not all phages immediately lyse their host upon infection. Lysogenic phages can integrate into the host chromosome and be maintained until specific environmental conditions trigger induction (Breitbart, 2012). Culture-based and bioinformatic approaches have estimated that about half of all marine bacteria harbor prophages, which can mediate defense against future infections by providing prophage-induced viral immunity (Bondy-Denomy et al., 2016; Breitbart, 2012; Zhao et al., 2019). Prophages may also provide novel functions to the host and can become fixed in a population if they provide fitness benefits to the recipient cell (Bondy-Denomy and Davidson, 2014; Paul, 2008; Rodriguez-Valera et al., 2009). For example, a region similar to the *Escherichia coli* defective prophage CP4-57 involved in biofilm development has been found in most *Alteromonas* genomes described so far (López-Pérez et al., 2014), suggesting that the prevalence and maintenance of this prophage in *Alteromonas* species may provide an advantage for survival. Prophages and prophage-like sequences have also been found in the streamlined organisms SAR11 and *Prochlorococcus*, respectively (Malmstrom et al., 2013; Morris et al.,

2020). This suggests that viral integration can also influence the evolution of even the smallest streamlined genomes.

1.8 Outlook

Comparative genomics has provided an important window into the ecology, evolution, and physiology of bacterial and archaeal lineages in the ocean. Here we have discussed some common themes in the evolutionary genomics of a selected number of marine bacteria and archaea. Many more clades are expected to be described (Yilmaz et al., 2015), and each of them will have unique features and evolutionary histories that must be taken into account to develop a comprehensive understanding of their evolution, ecology, and activities in the ocean. Metagenomics, single-cell sequencing, and culture-based methods will continue to increase the genomic representation of clades of marine bacteria and archaea in the future. This will provide an important basis for addressing key questions in microbial oceanography. These include a better understanding of the drivers and consequences of microdiversity in microbial populations, the phylogenetic relationships between major marine clades in the Tree of Life, reconstruction of key evolutionary innovations that have allowed for successful lineages to occupy their current ecological niches, and the extent to which marine viruses determine the ecologies and evolutionary trajectories of their hosts, among many others. Given the wide-ranging nature of these questions, continued research into the evolutionary genomics of marine microbes will be a critical component of future research.

1.9 Box. 1 Effective population size and its effects on microbial evolution

Scientists apply two different metrics when studying microbial population dynamics in nature or in culture: census population size (N_c) and effective population size (N_e). N_c refers to the number of organisms within a given population. N_e is a complex concept that reflects neutral diversity

within a population, that is, the size of a population evolving in the absence of selection that would generate as much neutral diversity as is actually observed in the population (Fraser et al., 2009; Luo et al., 2014a). Empirical evidence suggests that there is a mismatch of several orders of magnitude between both metrics for most of the archaea and bacteria clades studied so far. Such discrepancy is associated with processes like bottlenecks, in which neutral diversity is low due to a drastic decrease of the population in a short period of time. Other processes like selective sweeps may decrease neutral diversity because the population adapts to a new environment and rapidly fixes genes correlated with such conditions (Fraser et al., 2009; Luo, Swan, et al., 2014a). N_e is an important evolutionary concept because it determines the way evolution in a population occurs. Populations with small N_e will tend to evolve under stronger genetic drift relative to selection, whereas in large N_e microorganisms evolve predominantly through selection. Although some marine microbial populations have been described as enormous, all populations in nature are finite and evolve through the interplay of selection and drift (Bobay and Ochman, 2017). Estimates of N_e are difficult in microorganisms because its calculation requires mutation rate measurements and the detection of intraspecific nucleotide diversity at synonymous sites. Both values are unknown for most microbial species, particularly those that have not been cultured (Fraser et al., 2009; Luo et al., 2014a). Several studies performed on marine cultures suggest that abundant marine oligotrophs like SAR11 and *Prochlorococcus* have large N_e (Giovannoni, 2017; Kashtan et al., 2014; Wilhelm et al., 2007) and a study based on metagenomic data showed high intrapopulation sequence diversity in SAR11 (López-Pérez et al., 2020). These evolutionary traits suggest a prolonged habitat specialization and the absence of drastic population reductions in the recent past (Batut et al., 2014). Abundant bacterioplankton clades evolve mostly through selection, indicating a potential to purge deleterious mutations and spread favorable variants (Fraser et al., 2009; Luo et al., 2014a; Martinez-Gutierrez and Aylward, 2019). However, N_e estimations are still

needed for a broad range of microorganisms in order to obtain a comprehensive picture of the way marine archaea and bacteria evolve.

1.10 References

Ahlgren NA, Fuchsman CA, Rocap G, Fuhrman JA. 2019. Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. ISME J. 13:618–631.

Anantharaman K et al. 2014. Sulfur Oxidation Genes in Diverse Deep-Sea Viruses. Science. 344:757–760. doi: 10.1126/science.1252229.

Arístegui J, Gasol JM, Duarte CM, Herndld GJ. 2009. Microbial oceanography of the dark ocean's pelagic realm. Limnology and Oceanography. 54:1501–1529. doi: 10.4319/lo.2009.54.5.1501.

Avrani S, Wurtzel O, Sharon I, Sorek R, Lindell D. 2011. Genomic island variability facilitates Prochlorococcus-virus coexistence. Nature. 474:604–608.

Aylward FO et al. 2017. Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. Proceedings of the National Academy of Sciences. 114:11446–11451. doi: 10.1073/pnas.1714821114.

Aylward FO, Santoro AE. 2020. Heterotrophic Thaumarchaea with Small Genomes Are Widespread in the Dark Ocean. mSystems. 5. doi: 10.1128/mSystems.00415-20.

Azam F. 2004. OCEANOGRAPHY: Microbes, Molecules, and Marine Ecosystems. Science. 303:1622–1624. doi: 10.1126/science.1093892.

Azam F, Malfatti F. 2007. Microbial structuring of marine ecosystems. Nature Reviews Microbiology. 5:782–791. doi: 10.1038/nrmicro1747.

Bach W et al. 2006. Energy in the dark: Fuel for life in the deep ocean and beyond. Eos, Transactions American Geophysical Union. 87:73. doi: 10.1029/2006eo070002.

Bansal MS, Banay G, Peter Gogarten J, Shamir R. 2011. Detecting Highways of Horizontal Gene Transfer. Journal of Computational Biology. 18:1087–1114. doi: 10.1089/cmb.2011.0066.

Batut B, Knibbe C, Marais G, Daubin V. 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. Nat. Rev. Microbiol. 12:841–850.

Beja O. 2000. Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea. Science. 289:1902–1906. doi: 10.1126/science.289.5486.1902.

Berube PM et al. 2018. Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. Scientific Data. 5. doi: 10.1038/sdata.2018.154.

Biller SJ et al. 2014. Bacterial vesicles in marine ecosystems. Science. 343:183–186.

Biller SJ et al. 2017. Membrane vesicles in sea water: heterogeneous DNA content and implications for viral abundance estimates. ISME J. 11:394–404.

Bobay L-M, Ochman H. 2017. The Evolution of Bacterial Genome Architecture. Frontiers in Genetics. 8. doi: 10.3389/fgene.2017.00072.

Bock E, Wagner M. 2006. Oxidation of Inorganic Nitrogen Compounds as an Energy Source. The Prokaryotes. 457–495. doi: 10.1007/0-387-30742-7_16.

Bondy-Denomy J et al. 2016. Prophages mediate defense against phage infection through diverse mechanisms. The ISME Journal. 10:2854–2866. doi: 10.1038/ismej.2016.79.

Bondy-Denomy J, Davidson AR. 2014. When a virus is not a parasite: the beneficial effects of prophages on bacterial fitness. Journal of Microbiology. 52:235–242. doi: 10.1007/s12275-014-4083-3.

Bragg JG, Hyder CL. 2004. Nitrogen versus carbon use in prokaryotic genomes and proteomes. Proc. Biol. Sci. 271 Suppl 5:S374–7.

Breitbart M. 2012. Marine Viruses: Truth or Dare. Annual Review of Marine Science. 4:425–448. doi: 10.1146/annurev-marine-120709-142805.

Breitbart M, Bonnain C, Malki K, Sawaya NA. 2018. Phage puppet masters of the marine microbial realm. Nat Microbiol. 3:754–766.

Breitbart M, Miyake JH, Rohwer F. 2004. Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiol. Lett. 236:249–256.

Breitbart M, Rohwer F. 2005. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 13:278–284.

Breitbart M, Thompson L, Suttle C, Sullivan M. 2007. Exploring the Vast Diversity of Marine Viruses. Oceanography. 20:135–139. doi: 10.5670/oceanog.2007.58.

Brochier-Armanet C et al. 2011. Complete-fosmid and fosmid-end sequences reveal frequent horizontal gene transfers in marine uncultured planktonic archaea. ISME J. 5:1291–1302.

Brockhurst MA et al. 2019. The Ecology and Evolution of Pangenomes. Curr. Biol. 29:R1094–R1103.

Brown MV, Ostrowski M, Grzymski JJ, Lauro FM. 2014. A trait based perspective on the biogeography of common and abundant marine bacterioplankton clades. Mar. Genomics. 15:17–28.

Brum JR et al. 2015. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science. 348:1261498.

Bunse C, Pinhassi J. 2017. Marine Bacterioplankton Seasonal Succession Dynamics. Trends Microbiol. 25:494–505.

Cao H et al. 2016. Delta-proteobacterial SAR324 group in hydrothermal plumes on the South Mid-Atlantic Ridge. Sci. Rep. 6:22842.

Cohan FM. 2006. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. Philos. Trans. R. Soc. Lond. B Biol. Sci. 361:1985–1996.

Cohan FM, Perry EB. 2007. A systematics for discovering the fundamental units of bacterial diversity. Curr. Biol. 17:R373–86.

Cordero OX, Polz MF. 2014. Explaining microbial genomic diversity in light of evolutionary ecology. Nat. Rev. Microbiol. 12:263–273.

Corinaldesi C. 2015. New perspectives in benthic deep-sea microbial ecology. Frontiers in Marine Science. 2. doi: 10.3389/fmars.2015.00017.

Daims H et al. 2015. Complete nitrification by Nitrospira bacteria. Nature. 528:504–509.

Deatherage BL, Cookson BT. 2012. Membrane vesicle release in bacteria, eukaryotes, and archaea: a conserved yet underappreciated aspect of microbial life. Infect. Immun. 80:1948–1957.

DeLong EF. 1992. Archaea in coastal marine environments. Proc. Natl. Acad. Sci. U. S. A. 89:5685–5689.

DeLong EF. 2006. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. Science. 311:496–503. doi: 10.1126/science.1120250.

DeLong EF. 1997. Marine microbial diversity: the tip of the iceberg. Trends Biotechnol. 15:203–207.

DeLong EF, Béjà O. 2010. The light-driven proton pump proteorhodopsin enhances bacterial survival during tough times. PLoS Biol. 8:e1000359.

Deschamps P, Zivanovic Y, Moreira D, Rodriguez-Valera F, López-García P. 2014. Pangenome evidence for extensive interdomain horizontal transfer affecting lineage core and shell genes in uncultured planktonic thaumarchaeota and euryarchaeota. Genome Biol. Evol. 6:1549–1563.

Dick GJ. 2019. The microbiomes of deep-sea hydrothermal vents: distributed globally, shaped locally. Nat. Rev. Microbiol. 17:271–283.

Dupont CL et al. 2012. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J. 6:1186–1199.

Eguíluz VM et al. 2019. Scaling of species distribution explains the vast potential marine prokaryote diversity. Sci. Rep. 9:18710.

Falkowski PG. 1998. Biogeochemical Controls and Feedbacks on Ocean Primary Production. Science. 281:200–206. doi: 10.1126/science.281.5374.200.

Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science. 320:1034–1039.

Feil EJ. 2004. Small change: keeping pace with microevolution. Nat. Rev. Microbiol. 2:483–495.

Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB. 2005. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. Proc. Natl. Acad. Sci. U. S. A. 102:14683–14688.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. Science. 323:741–746.

Fridman S et al. 2017. A myovirus encoding both photosystem I and II proteins enhances cyclic electron flow in infected Prochlorococcus cells. Nat Microbiol. 2:1350–1357.

Frigaard N-U, Martinez A, Mincer TJ, DeLong EF. 2006. Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. Nature. 439:847–850.

Fuhrman JA. 1999. Marine viruses and their biogeochemical and ecological effects. Nature. 399:541–548.

Fuhrman JA, McCallum K, Davis AA. 1992. Novel major archaebacterial group from marine plankton. Nature. 356:148–149.

Futuyma DJ. 1986. *Evolutionary biology*. Sinauer Associates Inc.

Getz EW, Tithi SS, Zhang L, Aylward FO. 2018. Parallel Evolution of Genome Streamlining and Cellular Bioenergetics across the Marine Radiation of a Bacterial Phylum. MBio. 9. doi: 10.1128/mBio.01089-18.

Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. 2013. Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. Scientific Reports. 3. doi: 10.1038/srep02471.

Giovannoni SJ et al. 2005. Proteorhodopsin in the ubiquitous marine bacterium SAR11. Nature. 438:82–85.

Giovannoni SJ. 2017. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. Ann. Rev. Mar. Sci. 9:231–255.

Giovannoni SJ, Cameron Thrash J, Temperton B. 2014. Implications of streamlining theory for microbial ecology. ISME J. 8:1553–1565.

Giovannoni SJ, Stingl U. 2005. Molecular diversity and ecology of microbial plankton. Nature. 437:343–348. doi: 10.1038/nature04158.

Giovannoni SJ, Vergin KL. 2012. Seasonality in ocean microbial communities. Science. 335:671–676.

Gogarten JP, Peter Gogarten J, Townsend JP. 2005. Horizontal gene transfer, genome innovation and evolution. Nature Reviews Microbiology. 3:679–687. doi: 10.1038/nrmicro1204.

Gómez-Consarnau L et al. 2007. Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. Nature. 445:210–213. doi: 10.1038/nature05381.

Gómez-Consarnau L et al. 2010. Proteorhodopsin Phototrophy Promotes Survival of Marine Bacteria during Starvation. PLoS Biology. 8:e1000358. doi: 10.1371/journal.pbio.1000358.

Grzymski JJ, Dussaq AM. 2012. The significance of nitrogen cost minimization in proteomes of marine microorganisms. ISME J. 6:71–80.

Hallam SJ et al. 2006. Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. PLoS Biol. 4:e95.

Heidelberg KB, Gilbert JA, Joint I. 2010. Marine genomics: at the interface of marine microbial ecology and biodiscovery. Microb. Biotechnol. 3:531–543.

Hu J, Blanchard JL. 2009. Environmental Sequence Data from the Sargasso Sea Reveal That the Characteristics of Genome Reduction in Prochlorococcus Are Not a Harbinger for an Escalation in Genetic Drift. Molecular Biology and Evolution. 26:1191–1191. doi: 10.1093/molbev/msn299.

Hunt DE et al. 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. 320:1081–1085.

Hurwitz BL, Hallam SJ, Sullivan MB. 2013. Metabolic reprogramming by viruses in the sunlit and dark ocean. Genome Biol. 14:R123.

Ignacio-Espinoza JC, Ahlgren NA, Fuhrman JA. 2020. Long-term stability and Red Queen-like strain dynamics in marine viruses. Nat Microbiol. 5:265–271.

Karner MB, DeLong EF, Karl DM. 2001. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature. 409:507–510.

Kashtan N et al. 2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science. 344:416–420.

Kasting JF. 2002. Life and the Evolution of Earth's Atmosphere. Science. 296:1066–1068. doi: 10.1126/science.1071184.

van Kessel MAHJ et al. 2015. Complete nitrification by a single microorganism. Nature. 528:555–559.

Kettler GC et al. 2007. Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. PLoS Genet. 3:e231.

Kirchberger PC, Schmidt M, Ochman H. 2020. The Ingenuity of Bacterial Genomes. Annu. Rev. Microbiol. doi: 10.1146/annurev-micro-020518-115822.

Klieve AV et al. 2005. Naturally occurring DNA transfer system associated with membrane vesicles in cellulolytic Ruminococcus spp. of ruminal origin. Appl. Environ. Microbiol. 71:4248–4253.

Konstantinidis KT, Braff J, Karl DM, DeLong EF. 2009. Comparative Metagenomic Analysis of a Microbial Community Residing at a Depth of 4,000 Meters at Station ALOHA in the North Pacific Subtropical Gyre. Applied and Environmental Microbiology. 75:5345–5355. doi: 10.1128/aem.00473-09.

Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu. Rev. Microbiol. 55:709–742.

Kunin V, Ouzounis CA. 2003. The balance of driving forces during genome evolution in prokaryotes. Genome Res. 13:1589–1594.

Kuo C-H, Ochman H. 2009. The fate of new bacterial genes. FEMS Microbiology Reviews. 33:38–43. doi: 10.1111/j.1574-6976.2008.00140.x.

Landry Z, Swan BK, Herndl GJ, Stepanauskas R, Giovannoni SJ. 2017. SAR202 Genomes from the Dark Ocean Predict Pathways for the Oxidation of Recalcitrant Dissolved Organic Matter. MBio. 8. doi: 10.1128/mBio.00413-17.

Lang AS, Westbye AB, Beatty JT. 2017. The Distribution, Evolution, and Roles of Gene Transfer Agents in Prokaryotic Genetic Exchange. Annu Rev Virol. 4:87–104.

Lauro FM et al. 2009. The genomic basis of trophic strategy in marine bacteria. Proc. Natl. Acad. Sci. U. S. A. 106:15527–15533.

Lauro FM, Bartlett DH. 2008. Prokaryotic lifestyles in deep sea habitats. Extremophiles. 12:15–25.

Lindell D et al. 2007. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature. 449:83–86.

Lindell D et al. 2004. Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proc. Natl. Acad. Sci. U. S. A. 101:11013–11018.

López-Pérez M, Gonzaga A, Ivanova EP, Rodriguez-Valera F. 2014. Genomes of Alteromonas australica, a world apart. BMC Genomics. 15:483.

López-Pérez M, Rodriguez-Valera F. 2016. Pangenome Evolution in the Marine Bacterium Alteromonas. Genome Biol. Evol. 8:1556–1570.

Lücker S, Nowka B, Rattei T, Spieck E, Daims H. 2013. The Genome of Nitrospina gracilis Illuminates the Metabolism and Evolution of the Major Marine Nitrite Oxidizer. Frontiers in Microbiology. 4. doi: 10.3389/fmicb.2013.00027.

Luo H, Tolar BB, et al. 2014. Single-cell genomics shedding light on marine Thaumarchaeota diversification. ISME J. 8:732–736.

Luo H, Huang Y, Stepanauskas R, Tang J. 2017. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. Nat Microbiol. 2:17091.

Luo H, Swan BK, Stepanauskas R, Hughes AL, Moran MA. 2014a. Comparing effective population sizes of dominant marine alphaproteobacteria lineages. Environ. Microbiol. Rep. 6:167–172.

Luo H, Swan BK, Stepanauskas R, Hughes AL, Moran MA. 2014b. Evolutionary analysis of a streamlined lineage of surface ocean Roseobacters. ISME J. 8:1428–1439.

Luo H, Thompson LR, Stingl U, Hughes AL. 2015. Selection Maintains Low Genomic GC Content in Marine SAR11 Lineages. Mol. Biol. Evol. 32:2738–2748.

Lynch M. 2007. *The Origins of Genome Architecture*. Sinauer Associates Incorporated.

Malmstrom RR et al. 2013. Ecology of uncultured Prochlorococcus clades revealed through single-cell genomics and biogeographic analysis. The ISME Journal. 7:184–198. doi: 10.1038/ismej.2012.89.

Marston MF, Martiny JBH. 2016. Genomic diversification of marine cyanophages into stable ecotypes. Environ. Microbiol. 18:4240–4253.

Martin-Cuadrado A-B et al. 2015. A new class of marine Euryarchaeota group II from the Mediterranean deep chlorophyll maximum. ISME J. 9:1619–1634.

Martinez-Gutierrez CA, Aylward FO. 2019. Strong Purifying Selection Is Associated with Genome Streamlining in Epipelagic Marinimicrobia. Genome Biol. Evol. 11:2887–2894.

Martiny AC, Tai APK, Veneziano D, Primeau F, Chisholm SW. 2009. Taxonomic resolution, ecotypes and the biogeography of Prochlorococcus. Environ. Microbiol. 11:823–832.

Mas A, Jamshidi S, Lagadeuc Y, Eveillard D, Vandenkoornhuyse P. 2016. Beyond the Black Queen Hypothesis. ISME J. 10:2085–2091.

McDaniel LD et al. 2010. High frequency of horizontal gene transfer in the oceans. Science. 330:50.

Mehrshad M, Rodriguez-Valera F, Amoozegar MA, López-García P, Ghai R. 2018. The enigmatic SAR202 cluster up close: shedding light on a globally distributed dark ocean lineage involved in sulfur cycling. ISME J. 12:655–668.

Mende DR et al. 2017. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nature Microbiology. 2:1367–1373. doi: 10.1038/s41564-017-0008-3.

Middelboe M, Holmfeldt K, Riemann L, Nybroe O, Haaber J. 2009. Bacteriophages drive strain diversification in a marineFlavobacterium: implications for phage resistance and physiological

properties. Environmental Microbiology. 11:1971–1982. doi: 10.1111/j.1462-2920.2009.01920.x.

Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17:589–596.

Moore LR, Goericke R, Chisholm SW. 1995. Comparative physiology of Synechococcus and Prochlorococcus: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. Marine Ecology Progress Series. 116:259–275. doi: 10.3354/meps116259.

Moran MA. Genomics and Metagenomics of Marine Prokaryotes. Microbial Ecology of the Oceans. 91–129. doi: 10.1002/9780470281840.ch4.

Moran MA, Miller WL. 2007. Resourceful heterotrophs make the most of light in the coastal ocean. Nature Reviews Microbiology. 5:792–800. doi: 10.1038/nrmicro1746.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proceedings of the National Academy of Sciences. 93:2873–2878. doi: 10.1073/pnas.93.7.2873.

Morris JJ, Lenski RE, Zinser ER. 2012. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio. 3. doi: 10.1128/mBio.00036-12.

Morris RM, Cain KR, Hvorecny KL, Kollman JM. 2020. Lysogenic host–virus interactions in SAR11 marine bacteria. Nature Microbiology. doi: 10.1038/s41564-020-0725-x.

Nicol GW, Leininger S, Schleper C. 2014. Distribution and Activity of Ammonia-Oxidizing Archaea in Natural Environments. Nitrification. 157–178. doi: 10.1128/9781555817145.ch7.

Orcutt BN, Sylvan JB, Knab NJ, Edwards KJ. 2011. Microbial Ecology of the Dark Ocean above, at, and below the Seafloor. Microbiology and Molecular Biology Reviews. 75:361–422. doi: 10.1128/mmbr.00039-10.

Orellana LH et al. 2019. Niche differentiation among annually recurrent coastal Marine Group II Euryarchaeota. ISME J. 13:3024–3036.

Orsi WD. 2018. Ecology and evolution of seafloor and subseafloor microbial communities. Nat. Rev. Microbiol. 16:671–683.

Pachiadaki MG et al. 2017. Major role of nitrite-oxidizing bacteria in dark ocean carbon fixation. Science. 358:1046–1051.

Parks DH et al. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36:996–1004.

Partensky F, Garczarek L. 2010. Prochlorococcus: advantages and limits of minimalism. Ann. Rev. Mar. Sci. 2:305–331.

Paul JH. 2008. Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? The ISME Journal. 2:579–589. doi: 10.1038/ismej.2008.35.

Paul S, Dutta A, Bag SK, Das S, Dutta C. 2010. Distinct, ecotype-specific genome and proteome signatures in the marine cyanobacteria Prochlorococcus. BMC Genomics. 11:103.

Pomeroy L, Williams PL, Azam F, Hobbie J. 2007. The Microbial Loop. Oceanography. 20:28–33. doi: 10.5670/oceanog.2007.45.

Price MN, Arkin AP. 2015. Weakly Deleterious Mutations and Low Rates of Recombination Limit the Impact of Natural Selection on Bacterial Genomes. MBio. 6:e01302–15.

Qin W et al. 2020. Alternative strategies of nutrient acquisition and energy conservation map to the biogeography of marine ammonia-oxidizing archaea. ISME J. doi: 10.1038/s41396-020-0710-7.

Read RW et al. 2017. Nitrogen cost minimization is promoted by structural changes in the transcriptome of N-deprived Prochlorococcus cells. ISME J. 11:2267–2278.

Reji L, Francis CA. 2020. Metagenome-assembled genomes reveal unique metabolic adaptations of a basal marine Thaumarchaeota lineage. ISME J. 14:2105–2115.

Renelli M, Matias V, Lo RY, Beveridge TJ. 2004. DNA-containing membrane vesicles of Pseudomonas aeruginosa PAO1 and their genetic transformation potential. Microbiology. 150:2161–2169.

Ren M et al. 2019. Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. ISME J. 13:2150–2161.

Rinke C et al. 2019. A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (Ca. Poseidoniales ord. nov.). ISME J. 13:663–675.

Rocap G et al. 2003. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature. 424:1042–1047.

Rodriguez-Valera F et al. 2009. Explaining microbial population genomics through phage predation. Nat. Rev. Microbiol. 7:828–836.

Roux S et al. 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature. 537:689–693.

Salazar G, Sunagawa S. 2017. Marine microbial diversity. Curr. Biol. 27:R489–R494.

Salcher MM, Schaefle D, Kaspar M, Neuenschwander SM, Ghai R. 2019. Evolution in action: habitat transition from sediment to the pelagial leads to genome streamlining in Methylophilaceae. ISME J. 13:2764–2777.

Sánchez-Baracaldo P, Bianchini G, Di Cesare A, Callieri C, Chrismas NAM. 2019. Insights Into the Evolution of Picocyanobacteria and Phycoerythrin Genes (mpeBA and cpeBA). Frontiers

in Microbiology. 10. doi: 10.3389/fmicb.2019.00045.

Sanchez-Perez G, Mira A, Nyiro G, Pasić L, Rodriguez-Valera F. 2008. Adapting to environmental changes using specialized paralogs. Trends Genet. 24:154–158.

Sandaa R-A. 2008. Burden or benefit? Virus–host interactions in the marine environment. Research in Microbiology. 159:374–381. doi: 10.1016/j.resmic.2008.04.013.

Santoro AE et al. 2015. Genomic and proteomic characterization of 'Candidatus Nitrosopelagicus brevis': an ammonia-oxidizing archaeon from the open ocean. Proc. Natl. Acad. Sci. U. S. A. 112:1173–1178.

Santoro AE, Richter RA, Dupont CL. 2019. Planktonic Marine Archaea. Ann. Rev. Mar. Sci. 11:131–158.

Saw JHW et al. 2020. Pangenomics Analysis Reveals Diversification of Enzyme Families and Niche Specialization in Globally Abundant SAR202 Bacteria. MBio. 11. doi: 10.1128/mBio.02975-19.

Shakya M, Soucy SM, Zhaxybayeva O. 2017. Insights into origin and evolution of α-proteobacterial gene transfer agents. Virus Evol. 3:vex036.

Shapiro BJ et al. 2012. Population genomics of early events in the ecological differentiation of bacteria. Science. 336:48–51.

Sheik CS, Jain S, Dick GJ. 2014. Metabolic flexibility of enigmatic SAR324 revealed through metagenomics and metatranscriptomics. Environ. Microbiol. 16:304–317.

Sherr E, Sherr B. Understanding Roles of Microbes in Marine Pelagic Food Webs: A Brief History. Microbial Ecology of the Oceans. 27–44. doi: 10.1002/9780470281840.ch2.

Short CM, Suttle CA. 2005. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. Appl. Environ. Microbiol. 71:480–486.

Snel B, Bork P, Huynen MA. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res. 12:17–25.

Sobecky PA, Hazen TH. 2009. Horizontal gene transfer and mobile genetic elements in marine systems. Methods Mol. Biol. 532:435–453.

Staley JT, Konopka A. 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu. Rev. Microbiol. 39:321–346.

Stern A, Sorek R. 2011. The phage-host arms race: shaping the evolution of microbes. Bioessays. 33:43–51.

Sullivan MB et al. 2010. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environ. Microbiol. 12:3035–3056.

Sullivan MB et al. 2006. Prevalence and Evolution of Core Photosystem II Genes in Marine Cyanobacterial Viruses and Their Hosts. PLoS Biology. 4:e234. doi: 10.1371/journal.pbio.0040234.

Sun X et al. 2019. Uncultured Nitrospina-like species are major nitrite oxidizing bacteria in oxygen minimum zones. ISME J. 13:2391–2402.

Sun Z, Blanchard JL. 2014. Strong Genome-Wide Selection Early in the Evolution of Prochlorococcus Resulted in a Reduced Genome through the Loss of a Large Number of Small Effect Genes. PLoS ONE. 9:e88837. doi: 10.1371/journal.pone.0088837.

Suttle CA. 2007. Marine viruses — major players in the global ecosystem. Nature Reviews Microbiology. 5:801–812. doi: 10.1038/nrmicro1750.

Swan BK et al. 2014. Genomic and Metabolic Diversity of Marine Group I Thaumarchaeota in the Mesopelagic of Two Subtropical Gyres. PLoS ONE. 9:e95380. doi: 10.1371/journal.pone.0095380.

Swan BK et al. 2011. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. Science. 333:1296–1300.

Swan BK et al. 2013. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proc. Natl. Acad. Sci. U. S. A. 110:11463–11468.

Thingstad TF, Våge S, Storesund JE, Sandaa R-A, Giske J. 2014. A theoretical analysis of how strain-specific viruses can control microbial species diversity. Proc. Natl. Acad. Sci. U. S. A. 111:7813–7818.

Ting CS, Rocap G, King J, Chisholm SW. 2002. Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. Trends Microbiol. 10:134–142.

Torre JR de la et al. 2003. Proteorhodopsin genes are distributed among divergent marine bacterial taxa. Proceedings of the National Academy of Sciences. 100:12830–12835. doi: 10.1073/pnas.2133554100.

Touchon M, Moura de Sousa JA, Rocha EP. 2017. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. Curr. Opin. Microbiol. 38:66–73.

Treangen TJ, Rocha EPC. 2011. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLoS Genetics. 7:e1001284. doi: 10.1371/journal.pgen.1001284.

Tully BJ. 2019. Metabolic diversity within the globally abundant Marine Group II Euryarchaea offers insight into ecological patterns. Nat. Commun. 10:271.

Wang B et al. 2019. Expansion of Thaumarchaeota habitat range is correlated with horizontal

transfer of ATPase operons. ISME J. 13:3067–3079.

Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ. 2007. Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. Biol. Direct. 2:27.

Williams HTP. 2013. Phage-induced diversification improves host evolvability. BMC Evol. Biol. 13:17.

Witzel K-P. 1990. Approaches to Bacterial Population Dynamics. Aquatic Microbial Ecology. 96–128. doi: 10.1007/978-1-4612-3382-4_5.

Wolf YI, Koonin EV. 2013. Genome reduction as the dominant mode of evolution. Bioessays. 35:829–837.

Wommack KE, Eric Wommack K, Colwell RR. 2000. Virioplankton: Viruses in Aquatic Ecosystems. Microbiology and Molecular Biology Reviews. 64:69–114. doi: 10.1128/mmbr.64.1.69-114.2000.

Yaron S, Kolling GL, Simon L, Matthews KR. 2000. Vesicle-mediated transfer of virulence genes from Escherichia coli O157:H7 to other enteric bacteria. Appl. Environ. Microbiol. 66:4414–4420.

Yilmaz P, Yarza P, Rapp JZ, Glöckner FO. 2015. Expanding the World of Marine Bacterial and Archaeal Clades. Front. Microbiol. 6:1524.

Yutin N, Koonin EV. 2012. Proteorhodopsin genes in giant viruses. Biology Direct. 7:34. doi: 10.1186/1745-6150-7-34.

Zhang CL, Xie W, Martin-Cuadrado A-B, Rodriguez-Valera F. 2015. Marine Group II Archaea, potentially important players in the global ocean carbon cycle. Front. Microbiol. 6:1108.

Zhang H et al. 2019. Repeated evolutionary transitions of flavobacteria from marine to non-marine habitats. Environ. Microbiol. 21:648–666.

Zhang L et al. 2022. DNA barcoding of Cymbidium by genome skimming: call for next-generation nuclear barcodes. Mol. Ecol. Resour. doi: 10.1111/1755-0998.13719.

Zhao Y et al. 2019. Pelagiphages in the Podoviridae family integrate into host genomes. Environ. Microbiol. 21:1989–2001.

Zimmerman AE et al. 2020. Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. Nat. Rev. Microbiol. 18:21–34.

Zwirglmaier K et al. 2008. Global phylogeography of marine Synechococcus and Prochlorococcus reveals a distinct partitioning of lineages among oceanic biomes. Environ. Microbiol. 10:147–161.

Chapter 2. Research Project

Strong Purifying Selection is Associated with Genome Streamlining in Epipelagic Marinimicrobia

Previously published: Martinez-Gutierrez, C.A. and Aylward F.O. (2019) Strong Purifying Selection is Associated with Genome Streamlining in Epipelagic Marinimicrobia. Genome Biology and Evolution, 11(10), 2887-2894.

Co-authors contributed in the following ways: Conceived and designed this work: CAMG and FOA. Wrote the paper: CAMG and FOA.

2.1 Abstract

Marine microorganisms inhabiting nutrient-depleted waters play critical roles in global biogeochemical cycles due to their abundance and broad distribution. Many of these microbes share similar genomic features including small genome size, low % G+C content, short intergenic regions, and low nitrogen content in encoded amino acid residue side chains (N-ARSC), but the evolutionary drivers of these characteristics are unclear. Here we compared the strength of purifying selection across the *Marinimicrobia,* a candidate phylum which encompasses a broad range of phylogenetic groups with disparate genomic features, by estimating the ratio of non-synonymous and synonymous substitutions (dN/dS) in conserved marker genes. Our analysis reveals that epipelagic *Marinimicrobia* that exhibit features consistent with genome streamlining have significantly lower dN/dS values when compared to the mesopelagic counterparts. We also found a significant positive correlation between median dN/dS values and % G+C content, N-ARSC, and intergenic region length. We did not identify a significant correlation between dN/dS ratios and estimated genome size, suggesting the strength of selection is not a primary factor shaping genome size in this group. Our findings are generally consistent with genome streamlining theory, which postulates that many genomic features of abundant epipelagic bacteria are the result

of adaptation to oligotrophic nutrient conditions. Our results are also in agreement with previous findings that genome streamlining is common in epipelagic waters, suggesting that genomes inhabiting this region of the ocean have been shaped by strong selection together with prevalent nutritional constraints characteristic of this environment.

2.2 Main Text

Bacteria and Archaea play key roles in marine biogeochemical cycles and are a dominant force that drives global nutrient transformations (Azam et al., 1983; Falkowski et al. 2008). Our understanding of microbial diversity in the ocean has been transformed in the last few decades due to the discovery of several globally-abundant marine microbial lineages that are among the most numerically abundant life forms on Earth (Giovannoni & Stingl, 2005). Work on some of these abundant lineages succeeded in culturing representatives that could then be studied extensively in the laboratory, such as *Prochlorococcus marinus* (Chisholm et al., 1992) and heterotrophic bacterioplankton belonging to the *Pelagibacteriales* (Rappé et al., 2002), and *Roseobacter* groups (Luo & Moran, 2014), but many other dominant microbial lineages have not been brought into pure culture and require cultivation-independent methods for analysis (DeLong & Karl, 2005).

Previous research of *Prochlorococcus marinus* and *Pelagibacter ubique* genomes provided some of the earliest insights into the ecology and evolution of these dominant planktonic microbial lineages (Giovannoni et al., 2005; Rocap et al., 2003). It was quickly noted that both groups had small genomes that contained short intergenic regions and encoded among the fewest genes of any free-living organism (Giovannoni et al., 2005). These characteristics were explained through the proposed theory of genome streamlining, which states that genome simplification is an adaptation to consistently oligotrophic conditions, and that the loss of unnecessary genes and their corresponding transcriptional, translational, and regulatory burdens is advantageous (Giovannoni

et al., 2014). Genome streamlining theory is supported by the observation that many streamlined genomes also have lower % GC content and subsequently contain fewer codons encoding nitrogen-rich amino acids (Grzymski & Dussaq, 2012; Mende et al., 2017), which is expected to be advantageous in nutrient depleted conditions found in the open ocean (Giovannoni et al., 2014). Genome streamlining therefore corresponds to multiple characteristics, and recent cultivation-independent studies have confirmed that many of them are present in the genomes of a variety of marine lineages in addition to *Prochlorococcus* and *Pelagibacter* (Dupont et al., 2012; Ghai et al., 2013; Swan et al., 2013; Luo et al., 2014; Getz et al., 2018), suggesting that common evolutionary drivers shape diverse bacterioplankton groups in the ocean.

Although the term "genome streamlining" implies adaptation under oligotrophic nutrient conditions, it remains a possibility that these genomic signatures are non-adaptive or potentially the result of genetic drift (Batut et al., 2014). For example, it has long been known that many endosymbiotic bacteria contain small genomes with short intergenic regions and low % GC content, but in these cases a small effective population size (*Ne*) and correspondingly high genetic drift are likely responsible for these features (Kuo et al., 2009; Charlesworth, 2009). While it remains unlikely that marine free-living bacteria have small effective population sizes comparable to those of endosymbiotic bacteria, it has been argued that population bottlenecks in the distant evolutionary past of some marine lineages may be responsible for aspects of their present genomic architecture (Luo et al., 2017). Moreover, recent work has also shown that weakly deleterious mutations and low recombination rates can substantially lower the efficacy of purifying selection in bacterial genomes (Price & Arkin, 2015), implying that the large abundances of marine bacteria may not translate directly into high selection.

In this study we focused our analyses on the candidate phylum *Marinimicrobia*, a predominantly marine group that comprises diverse globally abundant lineages involved in distinct biogeochemical processes (Hawley et al., 2017; Getz et al., 2018). Formerly referred to as clade SAR406 or Marine Group A, the *Marinimicrobia* span a broad range of distinct marine lineages that are poorly understood, in part due to difficulties in cultivating representatives of this phylum. Advances in metagenomics and single-cell sequencing have yielded a large number of draft genomes from this group, however, and several recent studies have shed light on the important role of different *Marinimicrobia* lineages to carbon and nitrogen cycling in the ocean (Thrash et al., 2017; Zhang et al., 2016; Wright et al., 2014; Bertagnolli et al., 2017; Aylward et al., 2015; Plominsky et al., 2018). In a recent study we compiled a set of draft *Marinimicrobia* genomes that have been sequenced using cultivation-independent methods (Getz et al., 2018), and we leverage this set here to analyze the evolutionary genomics of this group. The 211 genomes we used here belong to *Marinimicrobia* that inhabit both epipelagic and mesopelagic waters across the global ocean and comprise a broad range of phylogenetic diversity (38% average amino acid identity among marker genes in the CheckM marker set; see Methods).

The *Marinimicrobia* are an ideal group to test genome streamlining theory because streamlined genomic traits have evolved multiple times independently in this phylum (Getz et al., 2018). Moreover, streamlined genomic characteristics are linked to the environment in which *Marinimicrobia* are found; epipelagic *Marinimicrobia* tend to have genomes with low % GC content, short intergenic spacers, and relatively low nitrogen and high carbon encoded amino acids, while mesopelagic *Marinimicrobia* generally lack these features (Fig. 2.1). Higher levels of purifying selection in epipelagic *Marinimicrobia* would therefore be consistent with genome streamlining theory because it would indicate that these genomic features are not due to genetic

drift. Conversely, lower levels of purifying selection in epipelagic *Marinimicrobia* would suggest that their genomes are shaped by a process analogous to that experienced by endosymbiotic bacteria.

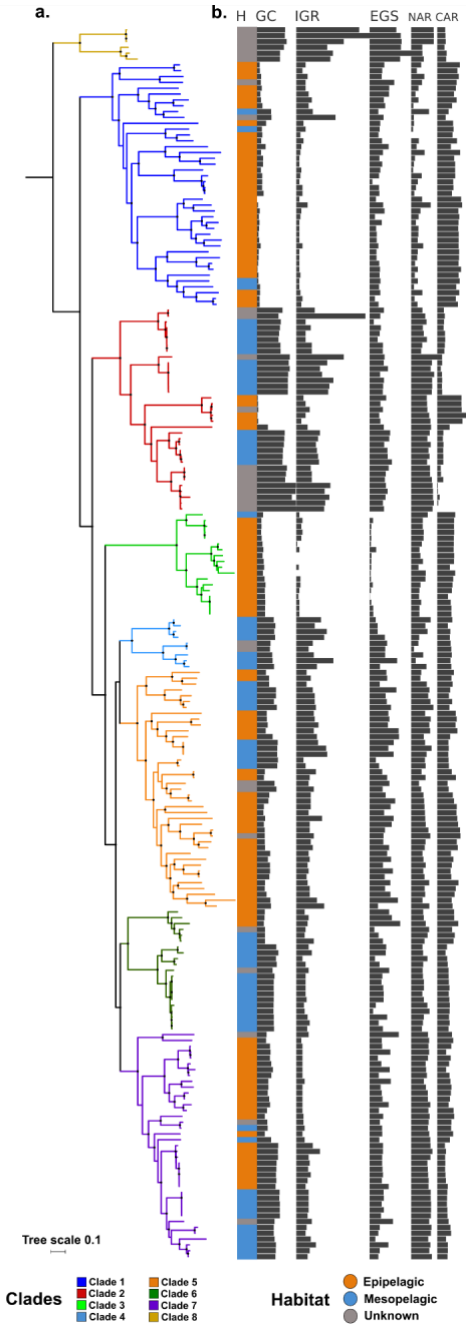

Figure 2.1 Representation of phylogeny, habitat classification, and genomic features. (a) Maximum likelihood phylogenetic tree of the 211 genomes constructed using amino acid sequences of 120 highly conserved marker genes. (b) Habitat classification based on Getz et al.

(2018) and genomic features of *Marinimicrobia* genomes. Abbreviations: H, Habitat; GC, % GC content (range, 27 to 55%); IGR, median intergenic region length (range, 7 to 78 nucleotides); EGS, estimated genome size (range 1-4.4 Mbp); N-ARSC (range, 0.3 to 0.34); C-ARSC (range, 3 to 3.2). Black points on branches represent support values > 0.95.

To test our hypothesis, we estimated the ratio of nonsynonymous and synonymous substitutions (dN/dS) of conserved marker genes in *Marinimicrobia*. In general, dN/dS values < 1 are indicative of purifying selection, and the relative strength of selection can be compared across groups using this metric, with lower values implying higher levels of purifying selection (Kryazhimskiy & Plotkin, 2008; Kuo et al., 2009). To ensure that our results could be accurately compared across divergent clades, we used two sets of marker genes that are broadly shared among Bacteria, which we refer to here as the EMBL (Sunagawa et al., 2013) and CheckM (Parks et al., 2015) gene sets. We observed a general trend in which epipelagic genomes exhibited lower median dN/dS values (Fig. 2.2, Supplemental File 1). Although less pronounced, we also observed higher median dS values in epipelagic *Marinimicrobia* (Supplemental File 3). The dN/dS values we obtained are far lower than one, which is consistent with the expectation that conserved phylogenetic marker genes experience purifying selection to maintain protein function. Our observation of lower median dN/dS values in epipelagic *Marinimicrobia* was strongly supported by statistical analyses of both the CheckM and the EMBL marker gene sets (Mann-Whitney U Test, $P < 0.005$ in both cases; Fig. 2.2). Our findings suggest that *Marinimicrobia* found in epipelagic waters experience higher levels of purifying selection than those inhabiting mesopelagic waters.

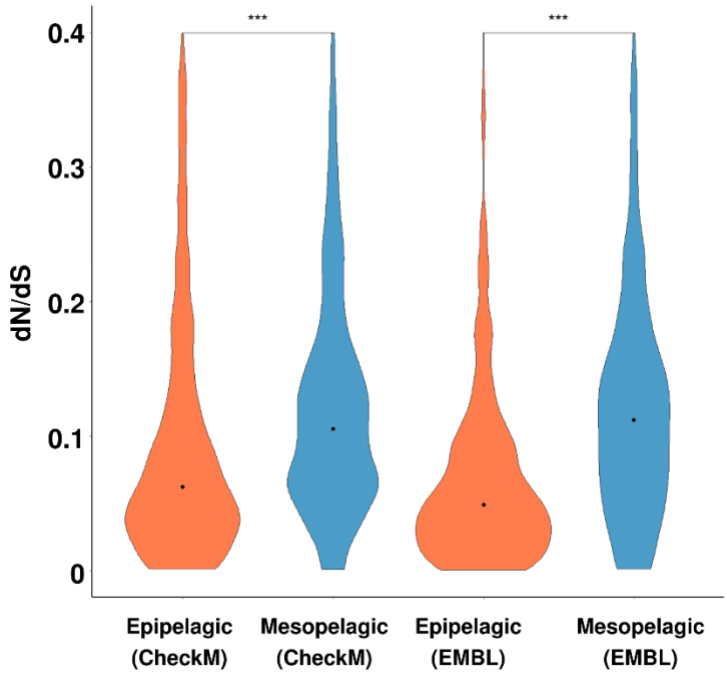

Figure 2.2 Violin plot representing median dN/dS values of epipelagic and mesopelagic *Marinimicrobia.* Statistical significance of differences between dN/dS values of the compared groups according to a non-paired, one-sided Mann Whitney-Wilcoxon test is denoted by: (***) for $P < 0.005$.

It has been shown that dN/dS values are dependent on the time scale in which comparisons are performed (Balbi et al., 2009; Rocha et al., 2006). To test if our results were consistent across different time scales, we created two sets of dN/dS values based on their corresponding dS values, which reflects sequence divergence; one set corresponded to dS values greater than the mean (more divergent comparisons) while the other set corresponded to those lower than the mean (less divergent comparisons). We compared epipelagic vs mesopelagic dN/dS values for both marker sets and found that epipelagic *Marinimicrobia* had lower median dN/dS values in all cases ($P < 0.005$), indicating that the time dependence of dN/dS values is not responsible for our findings.

We also explored the correlation between the strength of selection and several genomic features associated with streamlining. For this purpose, we generated several habitat-specific clusters of closely-related *Marinimicrobia* and plotted their median dN/dS ratio against average genomic characteristics within that cluster (see Methods). In general, we found that features consistent with genome streamlining were correlated with low dN/dS values (Fig. 2.3, Supplemental File 4), with the strongest correlations observed for % GC content ($\rho = 0.71$, Fig. 2.3a), nitrogen content in amino acid residue side chains (N-ARSC; $\rho = 0.54$, Fig. 2.3b), and median intergenic region length ($\rho = 0.68$, Fig. 2.3e). The low N-ARSC values in epipelagic *Marinimicrobia* are consistent with previous findings and are likely a product of nutrient limitation in surface waters (Getz et al. 2018). Similarly, we identified a negative correlation between dN/dS values and the carbon content of amino acid residue side chains (C-ARSC; $\rho = -0.59$, Fig. 2.3d), which is also consistent with higher carbon availability in epipelagic waters. The weakest correlation we observed was between dN/dS values and estimated genome size ($\rho = 0.24$, Fig. 2.3e). To confirm these trends, we also performed a multivariate analysis of the dN/dS values and genomic features of the *Marinimicrobia* genome clusters, and the results of this analysis confirmed the tendency of streamlined epipelagic genomes to have lower dN/dS values (Fig. 2.4, Supplemental File 5).

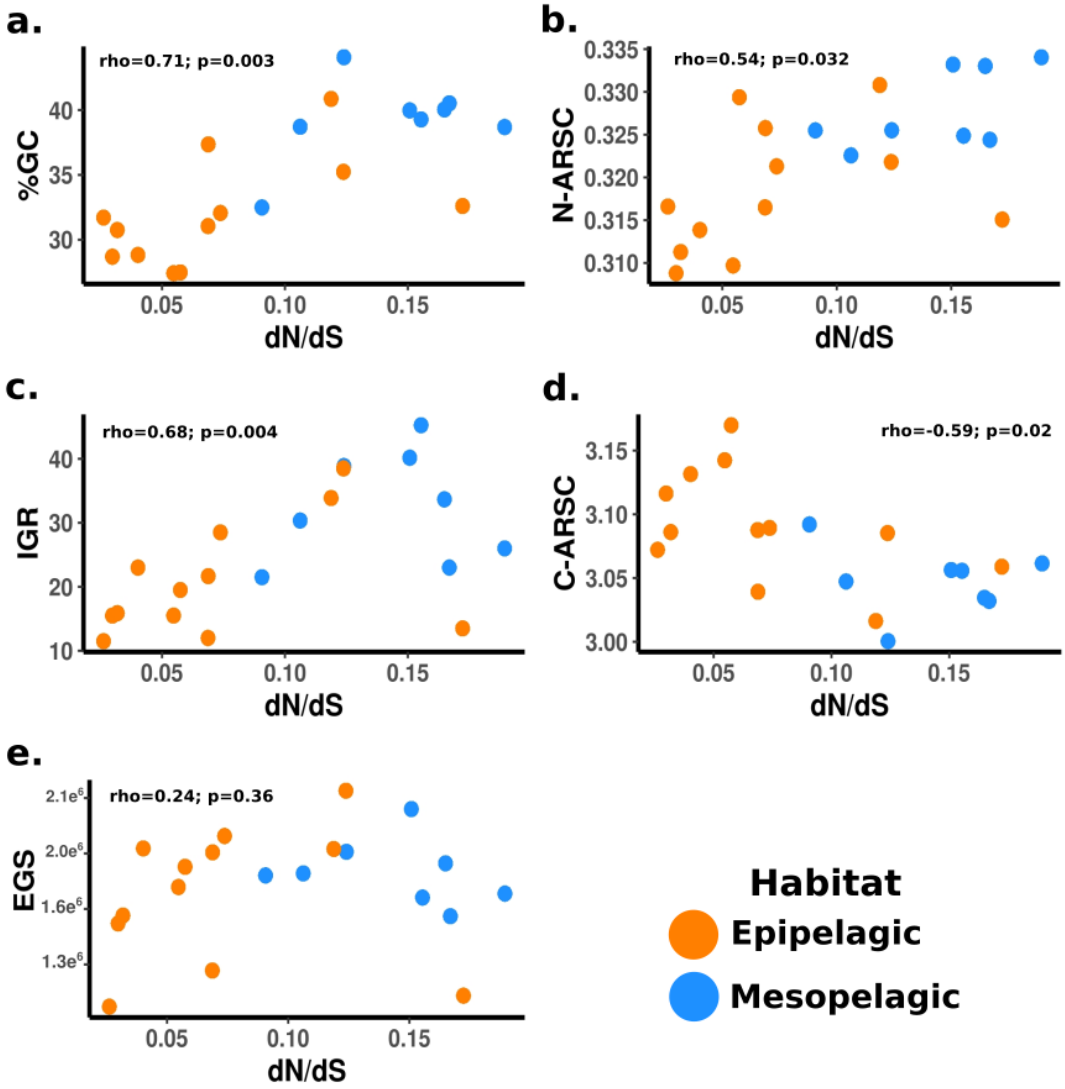

Figure 2.3 Scatter plots showing the relationship between median dN/dS values and streamlined genomic features of the *Marinimicrobia* genome clusters. Median dN/dS values were calculated using the CheckM marker gene set. a. GC content vs dN/dS; b. N-ARSC vs dN/dS; c. Median intergenic regions length (bp) vs dN/dS; d. C-ARSC vs dN/dS; e. estimated genome size (log bp) vs dN/dS. Spearman correlations were performed for each variable pair and details can be found on the main text. Details for the genome clusters can be found in Supplemental File 2.

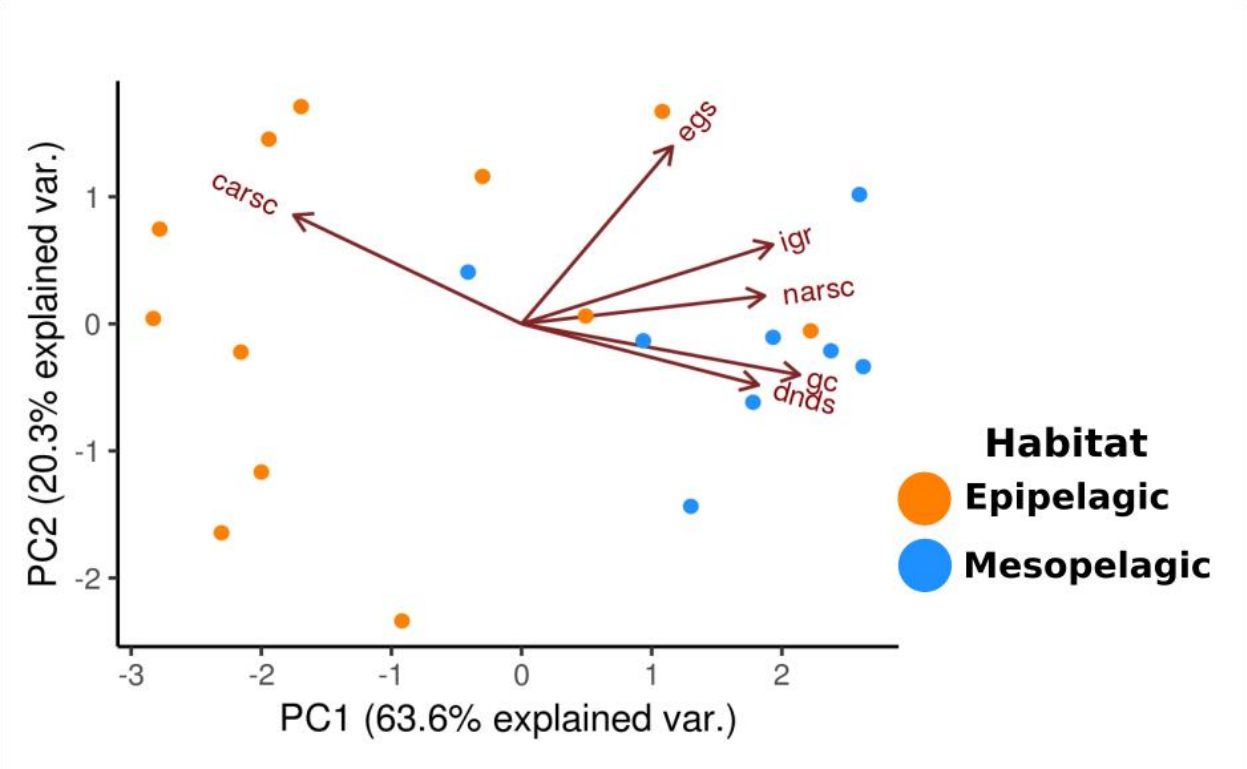

Figure 2.4 PCA analysis displaying the euclidean distance among *Marinimicrobia* genomes. Abbreviations: gc, %GC content; narsc: N-ARSC; igr, intergenic regions length; dNdS, dN/dS ratio; egs, estimated genome size; CARSC, C-ARSC.

Several previous studies compared a broad array of bacterial lineages and found that dN/dS ratios are generally higher in smaller genomes, suggesting that genetic drift is a prominent evolutionary force in genome reduction (Sela et al., 2016; Kuo et al., 2009; Novichkov et al., 2009). We did not identify a strong relationship between estimated genome size and dN/dS ratios in the Marinimicrobia, suggesting that this trend may not hold for this group. Many abundant marine lineages remain poorly represented in sequenced genome repositories due to difficulties in cultivation, and the evolutionary factors shaping their genomes are therefore relatively unexplored. There is evidence of widespread genome streamlining in abundant marine lineages in the ocean (Swan et al., 2013), and as more genomes become available it will be possible to rigorously evaluate the strength of purifying selection on these groups and its possible impact on genome

size. The theory of genome streamlining predicts that streamlining may occur across a range of genome sizes due to different genetic repertoires that are necessary to thrive in different environments or ecological niches (Giovannoni et al., 2014), and it is therefore unclear if we would expect streamlined epipelagic Marinimicrobia to have substantially smaller genomes overall. An additional complication of the present study is that the genomes analyzed are incomplete owing to their sequencing via metagenomic or single-cell sequencing efforts, and we extrapolated genome sizes from completeness estimates. The lack of a significant correlation between dN/dS values and genome size must therefore be interpreted with caution, and further studies will be needed to examine this in more detail.

It is important to note that the results we present here do not imply that strong selection directly leads to low % GC content, low N-ARSC, or other streamlined features, since the genomic changes that result from strong selection depend on prevailing environmental factors. For example, strong selection on genomes in mesopelagic waters would not be predicted to lead to a decrease in the N-ARSC of encoded proteins, since nitrogen is more abundant in deeper waters and this evolutionary transition would not be advantageous. The disparate genomic features of epipelagic and mesopelagic Marinimicrobia are therefore likely the result of differential nutrient availability and environmental factors along the water column; surface waters are relatively depleted in nitrogen and phosphorus, while mesopelagic waters contain more of these nutrients but less photosynthetically-derived carbon (Karl, 2002; Moore et al., 2013). Other factors in addition to selection under different environmental conditions may also play a role in genome streamlining; for example, it has been hypothesized that an increase in mutation rate may lead to some of the genomic features of both endosymbiotic bacteria and abundant marine bacteria (Marais et al., 2008). In our phylogeny of the Marinimicrobia, genomes that contain streamlined genomic

features in Clades 1 and 2 (red and light green in Fig. 2.1, respectively) are associated with long branches that may be indicative of increased mutation rates. This link between long branches and genome streamlining must be made with caution, however, because long branches do not definitively demonstrate increased mutation rates, and even so it is unclear if this would lead to other genomic features of streamlining. Nevertheless, given the complexity of these genome evolutionary processes it is likely that multiple factors are responsible for the trends we observe here.

An important caveat of the dN/dS ratio is that it only provides insight into the strength of recent selective pressure and therefore cannot be used to infer the selective strength experienced by lineages in the past. Other streamlined lineages such as the *Pelagibacterales* and *Prochlorococcus* are thought to have undergone genome reduction in the distant past, and it is therefore difficult to assess the strength of selection on these ancestral genomes during these transitions. Some studies have suggested that genetic drift due to possible population bottlenecks drove these genomic changes (Luo et al., 2017), while other studies have argued that strong purifying selection was the primary driver (Sun & Blanchard, 2014). In contrast to these other streamlined groups, the Marinimicrobia appear to have experienced multiple independent genome transition events relatively recently in their evolutionary history (Getz et al., 2018), and comparison of the selective pressures across disparate clades with similar genomic features therefore provides insight into more recent selective regimes that led to current genomic architectures. Overall, our results suggest that although parasitic and endosymbiotic bacteria share some genomic features with streamlined bacteria, these features are the product of distinct evolutionary paths.

2.3 Materials and Methods

2.3.1 *Marinimicrobia* genomes used

We analyzed a set of 211 *Marinimicrobia* genomes derived from a previous study (Getz et al., 2018). This data set included genomes from GenBank (Sayers et al., 2019), the Integrated Microbial Genomes database (IMG (Markowitz & Kyrpides, 2007)), and from two different studies in which Metagenome-Assembled Genomes (MAGs) were generated (Tully et al., 2018; Delmont et al., 2018). The data set employed by Getz *et al.* was complemented with the genomes SCGC_AD-604-D17, SCGC_AD-606-A07, SCGC_AD-615_E22 from another recent study (Plominsky et al. 2018). Methods for quality filtering, estimation of genome completeness and contamination, and the calculation of genomic features have been described previously (Getz et al., 2018).

2.3.2 Phylogenetic reconstruction

To reconstruct the *Marinimicrobia* phylogeny, we predicted proteins from genomes using Prodigal v2.6.2 (Hyatt et al., 2010) and identified phylogenetic marker genes using HMMER3 (Eddy, 2011). We constructed a phylogeny from an amino acid alignment created from the concatenation of 120 marker genes that have been previously used for phylogenetic reconstructions of Bacteria (Parks et al., 2015). The trusted cutoffs were used in all HMMER3 searches with the "cut_tc" option in hmmsearch. We used the standard_fasttree workflow included in the ETE Toolkit which includes ClustalOmega for alignment (Sievers & Higgins, 2018), trimAl for alignment trimming (Capella-Gutierrez et al., 2009), and FastTree for phylogenetic estimation (Price et al., 2010). The different branches obtained were classified into clades based on previously published results (Getz et al., 2018). We visualized the resulting tree in the interactive Tree of Life (iTOL (Letunic & Bork, 2016); https://itol.embl.de/tree/4537914239725156208868 3)).

2.3.3 dN/dS ratio calculation and filtering

To estimate the strength of purifying selection we used the ratio of nonsynonymous and synonymous substitutions (dN/dS). When considering values < 1, lower values are a sign of higher purifying selection while higher values are a sign of higher genetic drift (low purifying selection). To calculate genome-wide dN/dS ratios we used two sets of conserved marker genes, that would be expected to be found in most genomes. The first one consists of 120 phylogenetic marker genes that are highly conserved in Bacteria, which we also used for phylogenetic reconstruction (Parks et al., 2015). The second set consists of 40 phylogenetic marker genes used in phylogenetic reconstructions, which we refer to as the EMBL set due to its development in the European Molecular Biology Laboratory (Sunagawa et al., 2013).

For both marker gene sets, we predicted proteins from each genome using Prodigal and then annotated the marker genes of interest using the hmmsearch tool of HMMER3 with model-specific cutoffs. We aligned the amino acid sequences for each annotated gene coming from *Marinimicrobia* genomes separately using ClustalOmega, and the resulting alignments converted into codon alignments using PAL2NAL (Suyama et al., 2006). Maximum-likelihood approximation (codeML) within the PAML 4.9h package (Yang, 2007) was used through Biopython in order to perform dN/dS pairwise comparisons within the clades previously established (Getz et al., 2018). We removed dN/dS values with dS >= 1, which implies that synonymous substitutions are near saturation. Moreover, to avoid comparing sequences from genomes that may be part of the same population, we also excluded comparisons for which dN = 0 and dS <= 0.01. Additionally, we discarded all dN/dS values >= 10 on the grounds that these were largely artifactual. Lastly, because we wished to compare dN/dS values from *Marinimicrobia* that reside in different habitats, we only included dN/dS values where the pair of compared

genomes were from the same habitat (epipelagic and mesopelagic). All values used can be found in Supplemental File 1.

2.3.4 Genomes Clustering

To compare dN/dS values with other genomic features, it was first necessary to generate clusters of closely related genomes. For this, *Marinimicrobia* genomes were compared using the MASH program (Ondov et al., 2016), which rapidly identifies similarities in the k-mer profiles of genomes and provides statistical measures of nucleotide similarity. Comparisons that yielded MASH e-values < 1e-100 were retained and used to link closely-related genomes, and final genome clusters were generated using a single-linkage clustering algorithm in R. Median dN/dS values for all clusters were calculated and then plotted against average genome features within that cluster (% GC content, estimated genome size, median intergenic region length, estimated genome size, N-ARSC, and C-ARSC; see Figure 3 and 4). Clusters that had fewer than 10 total dN/dS measurements were excluded from further analyses. Details for the genome clusters can be found in Supplemental File 2. This approach of using clusters of related genomes to estimate group-specific dN/dS values is similar to previously used methods (Novichkov et al., 2009; Kuo et al., 2009).

2.3.5 Statistical analyses

To investigate the strength of selection acting on epipelagic and mesopelagic *Marinimicrobia*, genomes were classified into epipelagic and mesopelagic based on their biogeographic distribution (Getz et al., 2018). For statistical analysis, we loaded the filtered dN/dS values into R and performed comparisons using the Mann-Whitney U test (wilcox.test() function). Additionally, to investigate the relationship between median dN/dS and genomic features associated with each cluster, we applied the "cor.test" function using the Spearman method. Comparisons and correlation plots were visualized through the ggplot2 package (Wilkinson, 2011). We also

explored the distance between epipelagic and mesopelagic *Marinimicrobia* genomes employing the genomic features and median dN/dS values through a PCA analysis with the "prcomp" function available on R. Euclidean distance was visualized using the "ggbiplot" function within the ggplot2 package (Wilkinson, 2011).

2.4 Data availability

The supplemental files of this chapter are available of the Figshare collection: https://doi.org/10.6084/m9.figshare.c.6240783.v1

2.5 Acknowledgements

We acknowledge the use of the Virginia Tech Advanced Research Computing Center for bioinformatic analyses performed in this study. This work was supported by grants from the Institute for Critical Technology and Applied Science at Virginia Tech, a Sloan Research Fellowship, and a Simons Early Career Award in Marine Microbial Ecology and Evolution to FOA.

2.6 References

Aylward FO et al. 2015. Microbial community transcriptional networks are conserved in three domains at ocean basin scales. Proceedings of the National Academy of Sciences. 112:5443–5448. doi: 10.1073/pnas.1502883112.

Azam F et al. 1983. The Ecological Role of Water-Column Microbes in the Sea. Marine Ecology Progress Series. 10:257–263. doi: 10.3354/meps010257.

Balbi KJ, Rocha EPC, Feil EJ. 2009. The temporal dynamics of slightly deleterious mutations in Escherichia coli and Shigella spp. Mol. Biol. Evol. 26:345–355.

Batut B, Knibbe C, Marais G, Daubin V. 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. Nat. Rev. Microbiol. 12:841–850.

Bertagnolli AD, Padilla CC, Glass JB, Thamdrup B, Stewart FJ. 2017. Metabolic potential and in situ activity of marine Marinimicrobia bacteria in an anoxic water column. Environmental Microbiology. 19:4392–4416. doi: 10.1111/1462-2920.13879.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment

trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972–1973. doi: 10.1093/bioinformatics/btp348.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics. 10:195–205. doi: 10.1038/nrg2526.

Chisholm SW et al. 1992. Prochlorococcus marinus nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll a and b. Archives of Microbiology. 157:297–300. doi: 10.1007/bf00245165.

Delmont TO et al. 2018. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat Microbiol. 3:804–813.

DeLong EF, Karl DM. 2005. Genomic perspectives in microbial oceanography. Nature. 437:336–342.

Dupont CL et al. 2012. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J. 6:1186–1199.

Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput. Biol. 7:e1002195.

Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science. 320:1034–1039.

Getz EW, Tithi SS, Zhang L, Aylward FO. 2018. Parallel Evolution of Genome Streamlining and Cellular Bioenergetics across the Marine Radiation of a Bacterial Phylum. MBio. 9. doi: 10.1128/mBio.01089-18.

Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. 2013. Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. Sci. Rep. 3:2471.

Giovannoni SJ et al. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science. 309:1242–1245.

Giovannoni SJ, Stingl U. 2005. Molecular diversity and ecology of microbial plankton. Nature. 437:343–348. doi: 10.1038/nature04158.

Giovannoni SJ, Thrash CJ, Temperton B. 2014. Implications of streamlining theory for microbial ecology. ISME J. 8:1553–1565.

Grzymski JJ, Dussaq AM. 2012. The significance of nitrogen cost minimization in proteomes of marine microorganisms. ISME J. 6:71–80.

Hawley AK et al. 2017. Diverse Marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients. Nat. Commun. 8:1507.

Hyatt D et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 11. doi: 10.1186/1471-2105-11-119.

Karl DM. 2002. Nutrient dynamics in the deep blue sea. Trends Microbiol. 10:410–418.

Kryazhimskiy S, Plotkin JB. 2008. The Population Genetics of dN/dS. PLoS Genetics. 4:e1000304. doi: 10.1371/journal.pgen.1000304.

Kuo C-H, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res. 19:1450–1454.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44:W242–5.

Luo H, Huang Y, Stepanauskas R, Tang J. 2017. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. Nat Microbiol. 2:17091.

Luo H, Moran MA. 2014. Evolutionary ecology of the marine Roseobacter clade. Microbiol. Mol. Biol. Rev. 78:573–587.

Luo H, Swan BK, Stepanauskas R, Hughes AL, Moran MA. 2014. Evolutionary analysis of a streamlined lineage of surface ocean Roseobacters. ISME J. 8:1428–1439.

Marais GAB, Calteau A, Tenaillon O. 2008. Mutation rate and genome reduction in endosymbiotic and free-living bacteria. Genetica. 134:205–210.

Markowitz VM, Kyrpides NC. 2007. Comparative genome analysis in the integrated microbial genomes (IMG) system. Methods Mol. Biol. 395:35–56.

Mende DR et al. 2017. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nature Microbiology. 2:1367–1373. doi: 10.1038/s41564-017-0008-3.

Moore CM et al. 2013. Processes and patterns of oceanic nutrient limitation. Nature Geoscience. 6:701–710. doi: 10.1038/ngeo1765.

Novichkov PS, Ratnere I, Wolf YI, Koonin EV, Dubchak I. 2009. ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. Nucleic Acids Research. 37:D448–D454. doi: 10.1093/nar/gkn684.

Novichkov PS, Wolf YI, Dubchak I, Koonin EV. 2009. Trends in Prokaryotic Evolution Revealed by Comparison of Closely Related Bacterial and Archaeal Genomes. J. Bacteriol. 191: 65-73. doi:10.1128/JB.01237-08.

Ondov BD et al. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17:132.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25:1043–1055.

Plominsky AM et al. 2018. Metabolic potential and in situ transcriptomic profiles of previously uncharacterized key microbial groups involved in coupled carbon, nitrogen and sulfur cycling in anoxic marine zones. Environ. Microbiol. 20:2727–2742.

Price MN, Arkin AP. 2015. Weakly Deleterious Mutations and Low Rates of Recombination Limit

the Impact of Natural Selection on Bacterial Genomes. MBio. 6:e01302–15.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE. 5:e9490. doi: 10.1371/journal.pone.0009490.

Rappé MS, Connon SA, Vergin KL, Giovannoni SJ. 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature. 418:630–633.

Rocap G et al. 2003. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature. 424:1042–1047.

Rocha EPC et al. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J. Theor. Biol. 239:226–235.

Sayers EW et al. 2019. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 47:D23–D28.

Sela I, Wolf YI, Koonin EV. 2016. Theory of prokaryotic genome evolution. Proceedings of the National Academy of Sciences. 113:11399–11407. doi: 10.1073/pnas.1614083113.

Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Science. 27:135–145. doi: 10.1002/pro.3290.

Sunagawa S et al. 2013. Metagenomic species profiling using universal phylogenetic marker genes. Nat. Methods. 10:1196–1199.

Sun Z, Blanchard JL. 2014. Strong Genome-Wide Selection Early in the Evolution of Prochlorococcus Resulted in a Reduced Genome through the Loss of a Large Number of Small Effect Genes. PLoS ONE. 9:e88837. doi: 10.1371/journal.pone.0088837.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609–12.

Swan BK et al. 2013. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proc. Natl. Acad. Sci. U. S. A. 110:11463–11468.

Thrash JC et al. 2017. Metabolic Roles of Uncultivated Bacterioplankton Lineages in the Northern Gulf of Mexico 'Dead Zone'. mBio. 8. doi: 10.1128/mbio.01017-17.

Tully BJ, Graham ED, Heidelberg JF. 2018. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data. 5:170203.

Wilkinson L. 2011. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. Biometrics. 67:678–679. doi: 10.1111/j.1541-0420.2011.01616.x.

Wright JJ et al. 2014. Genomic properties of Marine Group A bacteria indicate a role in the marine sulfur cycle. The ISME Journal. 8:455–468. doi: 10.1038/ismej.2013.152.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Zhang W et al. 2016. Genomic and Transcriptomic Evidence for Carbohydrate Consumption among Microorganisms in a Cold Seep Brine Pool. Frontiers in Microbiology. 7. doi: 10.3389/fmicb.2016.01825.

Chapter 3. Research Project

Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea

Previously published: Martinez-Gutierrez CA and Aylward FO. 2021. Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. Molecular Biology and Evolution 38(12), 5514–5527.

Co-authors contributed in the following ways: Conceived and designed this work: CAMG and FOA. Wrote the paper: CAMG and FOA.

3.1 Abstract

Reconstruction of the Tree of Life is a central goal in biology. Although numerous novel phyla of bacteria and archaea have recently been discovered, inconsistent phylogenetic relationships are routinely reported, and many inter-phylum and inter-domain evolutionary relationships remain unclear. Here, we benchmark different marker genes often used in constructing multidomain phylogenetic trees of bacteria and archaea and present a set of marker genes that perform best for multidomain trees constructed from concatenated alignments. We use recently developed Tree Certainty metrics to assess the confidence of our results and to obviate the complications of traditional bootstrap-based metrics. Given the vastly disparate number of genomes available for different phyla of bacteria and archaea, we also assessed the impact of taxon sampling on multidomain tree construction. Our results demonstrate that biases between the representation of different taxonomic groups can dramatically impact the topology of resulting trees. Inspection of our highest-quality tree supports the division of most bacteria into *Terrabacteria* and Gracilicutes, with Thermatogota and Synergistota branching earlier from these superphyla. This tree also supports the inclusion of the *Patescibacteria* within the *Terrabacteria* as a sister group to the Chloroflexota instead of as a basal-branching lineage. For the Archaea, our tree supports three monophyletic lineages (DPANN, Euryarchaeota, and TACK/Asgard), although we note the basal

placement of the DPANN may still represent an artifact caused by biased sequence composition. Our findings provide a robust and standardized framework for multidomain phylogenetic reconstruction that can be used to evaluate inter-phylum relationships and assess uncertainty in conflicting topologies of the Tree of Life.

3.2 Introduction

Due to the lack of informative morphological characters and a limited fossil record, phylogenies of bacteria and archaea have historically relied on molecular sequences (Altermann and Kazmierczak 2003; Battistuzzi et al. 2004). Woese and collaborators proposed the use of the small subunit ribosomal RNA genes (SSU) due to their "molecular chronometer" nature and fast- and slow- evolving positions (Woese and Fox 1977; Doolittle 1999). This allowed the reconstruction of a universal Tree of Life that included bacteria, archaea, and eukaryotes (Woese 1987). Although single genes like 16S rRNA have had a tremendous value for the study of prokaryotes phylogeny over the last decades, their use is often problematic owing to PCR-amplification bias, saturation derived from the use of nucleotides, and a limited number of alignment positions that may be insufficient for resolving evolutionary relationships among divergent lineages (Lerat et al. 2003; Konstantinidis and Tiedje 2007; Anon 2011). Recently, the application of high-throughput sequencing methodologies has allowed the recovery of a vast amount of genomic data that have improved taxonomic sampling across bacteria and archaea and enabled for "whole-genome phylogenies", i.e., trees inferred from the concatenation of numerous marker genes. These advances, together with improvements in computational power now permit analyses of concatenated alignments that include thousands of characters belonging to a broad diversity of taxa (Ciccarelli et al. 2006; Klenk and Göker 2010; Segata et al. 2013; Hug et al. 2016; Parks et al. 2017; Coleman et al. 2020).

Despite these advances, it remains unclear if the inclusion of more genes and genomes necessarily improves the quality of resulting trees, and, if not, which marker gene sets and taxon sampling strategies produce the most robust phylogenies. The concatenation of multiple genes may improve accuracy due to an increase in the number of phylogenetically informative characters in relation to noise sites (Gadagkar et al. 2005; de Queiroz and Gatesy 2007). Still, some genes may have undergone horizontal gene transfer (HGT) and will therefore have an evolutionary history distinct from other genes in the alignment, introducing phylogenetic noise and complicating the interpretation of results. In addition, different genes evolve at different rates, and the inclusion of many disparate protein families can introduce a heterotacheous signal that may lead to long-branch attraction and other phylogenetic artefacts (Philippe and Laurent 1998; Gribaldo and Philippe 2002; Bleidorn 2017). Moreover, traditional approaches used to assess phylogenetic confidence, such as the bootstrap, were developed for the analysis of single-gene trees and often provide misleadingly high support when applied to trees constructed from multigene concatenations because support increases artificially with alignment length (Delsuc et al. 2005; Jeffroy et al. 2006; Salichos et al. 2014; Simmons and Gatesy 2016; Simon 2020; Stott and Bobay 2020). Lastly, given the highly biased taxonomic composition of the sequenced genome collection, different taxonomic groups can be sampled to dramatically different depths. Several studies have noted that taxon sampling can impact phylogenetic results (Rokas and Carroll 2005; Nasir et al. 2016; Cunha et al. 2017), but the overall impact of markedly different taxon sampling between phylogenetic groups remains unclear.

Given the complications associated with the construction of phylogenetic trees from concatenated alignments, it is not surprising that many recent studies have reported conflicting results concerning the placement of deep-branching groups of bacteria and archaea. For example,

several studies have reported the *Patescibacteria* (also known as the Candidate Phyla Radiation, or CPR) as basal-branching in bacteria, but recent studies have suggested that this group is a sister phyla to the *Chloroflexota* (Coleman et al. 2020; Taib et al. 2020). Controversy has also surrounded the placement of the Asgard archaea, with some studies showing they are placed near the TACK superphylum (comprised of the *Thaumarchaeota, Aigarchaeota, Crenarchaeota,* and *Korarchaeota)* and are closely related to eukaryotes, and other studies reporting placement within the Euryarchaeota (Cunha et al. 2017; Zaremba-Niedzwiedzka et al. 2017; Da Cunha et al. 2018; Williams et al. 2020). Further, some studies have even suggested that the long branch between bacteria and archaea precludes any robust generation of a multi-domain TOL, further complicating the identification of basal-branching groups from any domain (Gaucher et al. 2010; Coleman et al. 2020).

In this study we benchmarked different single-copy marker genes (SCMs) commonly used in multi-domain bacterial and archaeal phylogenetics and using recently-developed tree certainty metrics we identify a set of SCMs that performs best for phylogenetic trees derived from concatenated alignments (workflow in Fig. 3.1). Moreover, we benchmark different taxon sampling strategies and demonstrate that uneven representation of phyla can dramatically impact the resulting trees and lower their overall tree certainty. Using the best-performing marker gene set and balanced taxon sampling across bacteria and archaea, we then reconstructed a high-resolution tree that clarifies the phylogenetic relationships between several phyla and identifies several deep-branching nodes where the true topology remains unclear. Our results provide a robust and standardized framework for phylogenetic reconstruction of bacteria and archaea that quantifies the certainty and limitations of concatenated gene trees for resolving deep branching nodes.

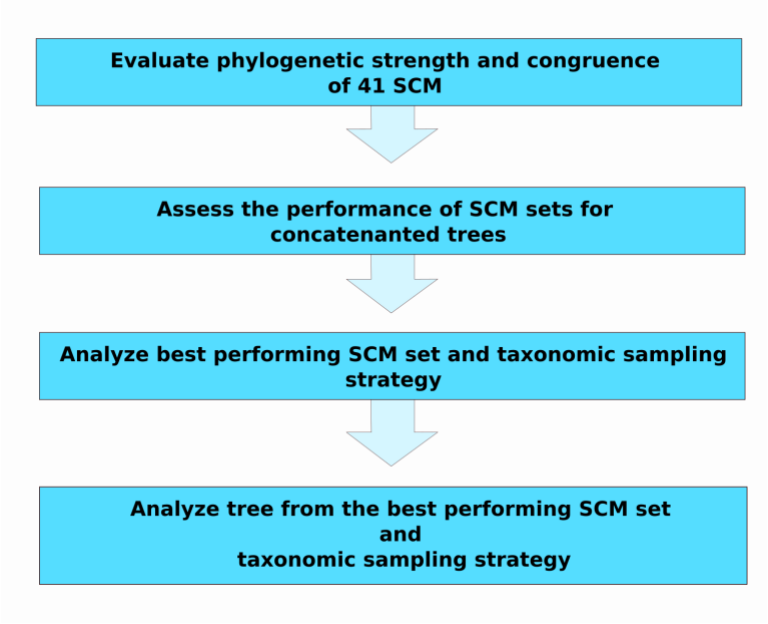

Figure 3.1 Schematic summary of the methodological workflow used in this study. Abbreviations: SCM; Single-copy marker.

3.3 Results and discussion

3.3.1 Evaluating the phylogenetic congruence of individual marker genes used for TOL reconstruction

Given the large phylogenetic distance encompassed by bacteria and archaea, there are few single-copy marker genes (SCMs) that are suitable for inter-domain phylogenetic reconstruction (Berkemer and McGlynn 2020). Nevertheless, several independent studies have found 30-40 orthologous protein families that can be used for this purpose (Ciccarelli et al. 2006; Wu and Eisen 2008; Williams et al. 2012); these include RNA polymerase subunits, ribosomal proteins, tRNA synthetases, and proteins involved in intracellular trafficking. Using a set of 41 SCMs that encompasses this set and has been previously used for this purpose (Sunagawa et al. 2013), we first evaluated the occurrence of these SCMs in a curated set of 1,650 bacterial and archaeal genomes derived from the Genome Taxonomy Database (GTDB, see Methods) (Chaumeil et al. 2019). Our results confirmed that these SCMs are broadly found in diverse bacterial and archaeal

lineages as a single-copy genes: RNA polymerase subunits, ribosomal proteins, tRNA synthetases, and intracellular trafficking proteins showed a high occurrence (83-97%) and low presence of multiple copies (0.1-0.8%) (Supplemental File 15). In contrast, other genes that have been used in the past as SCMs were found in either a lower fraction of the genomes (e.g., *recA*, found in only 71% of the genomes surveyed), or were often not found as single copy (e.g., *EF-Tu*, found as multi-copy in 26% of the genome surveyed) (Supplemental File 15). The β and β' subunits of RNA Polymerase (RNAP, COG0085 and COG0086, respectively) are known to be fragmented into multiple individual genes (6.42 and 3.52% for COG0085 and COG0086, respectively) (Werner and Grohmann 2011), which can lead to the erroneous conclusion that paralogs of these genes are present, but concatenation of the gene fragments ameliorates this issue (Supplemental File 16, see Methods).

We developed a bioinformatic tool called MarkerFinder to easily identify different marker gene sets from bacterial and archaeal genomes and produce a concatenated alignments that can be used for phylogenetic reconstruction (https://github.com/faylward/markerfinder). MarkerFinder also identifies fragmented RNAP subunits and concatenates them together, thereby obviating the difficulty in including genomes with fragmented RNAP subunits (see Methods). Using this bioinformatic framework we first benchmarked the phylogenetic signal and congruence of each SCM individually using the Tree Certainty metric (TC). The TC represents the mean of all the "Internode Certainty" values (IC), an estimate that assesses the degree of conflict of each internal node in a given tree (Salichos and Rokas 2013a; Kobert et al. 2016a). In contrast to other support estimates like bootstrap or posterior probabilities, the IC index reflects the conflict of a given bipartition by comparing its frequency with a set of conflicting bipartitions in a collection of replicate trees (Kobert et al. 2016a). Our results show a clear relationship between SCM length

and TC estimates (Fig. 3.2), consistent with the view that longer SCMs tend to have higher phylogenetic signal. The β and β' subunits of RNAP have the highest phylogenetic signal and represent the longest genes, followed by several tRNA-synthetases (Fig. 3.2, Supplemental File 2). Ribosomal proteins (RPs) were among the shortest SCMs in our analysis and tended to have low phylogenetic signal, indicating that these genes, when used individually, generally perform poorly as phylogenetic markers.

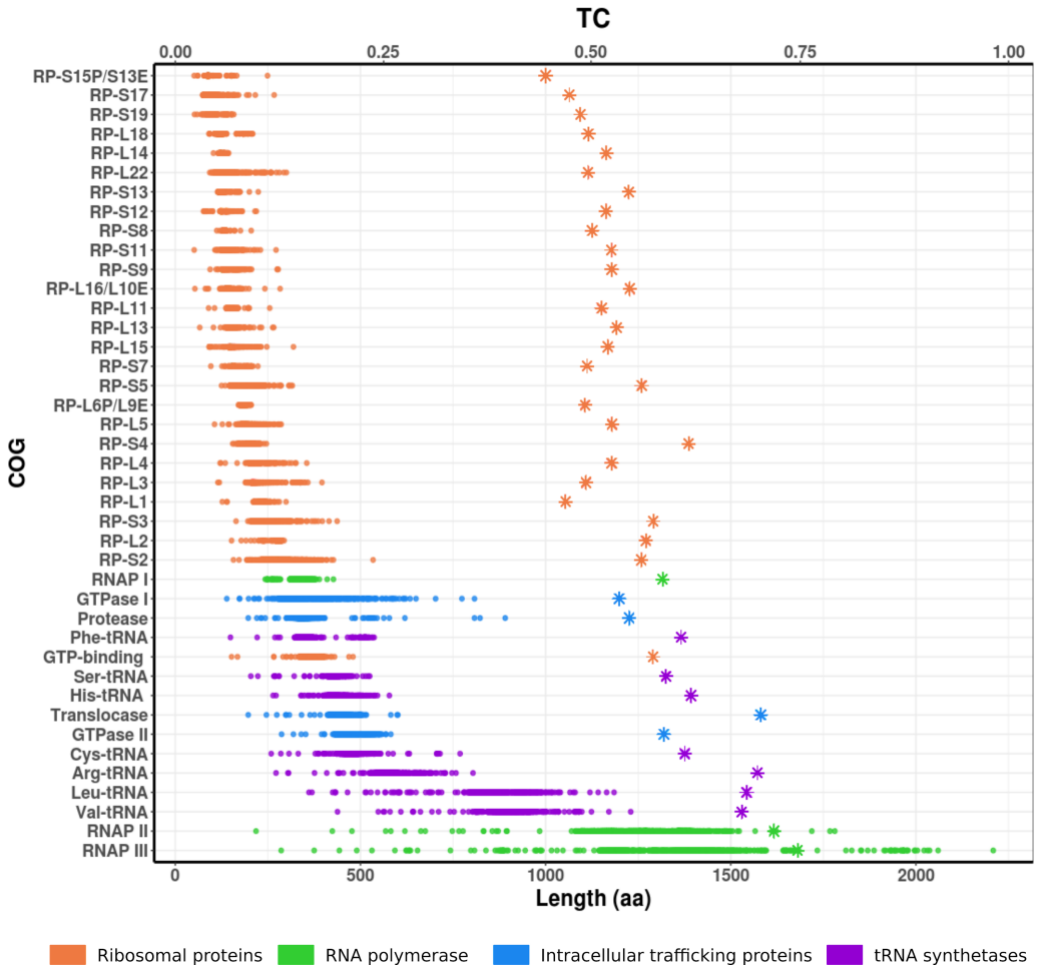

Figure 3.2 Tree certainty and length of marker genes used for the reconstruction of prokaryotic phylogenies. Circles represent the length of each sequence used to reconstruct each COG/protein tree and asterisks TC estimates. Colors: orange, ribosomal proteins; green, RNAP proteins; blue, intracellular trafficking; purple, tRNA proteins. RP = Ribosomal protein.

Because the concatenation of multiple SCMs with disparate evolutionary histories will lead to ambiguous results in the resulting tree, it is also critical to assess the level of phylogenetic congruence between different SCMs before concatenation. To do this we compared the TC values of each SCM against the mean Robinson-Foulds distance (RF) of each SCM's tree against those of all other SCMs. The mean RF distances can be taken as a measure of how consistent the phylogenetic signal of each SCM is compared to all others (Robinson and Foulds 1981). The resulting plot showed a negative correlation between mean RF distance and TC (Fig. 1.3A, Pearson's Rho -0.82, $p<0.001$), consistent with the view that SCMs with high phylogenetic signal tend to provide more consistent topologies because they provide more robust phylogenetic reconstruction. This was most clearly evidenced for the RNAP β and β' subunits, which had the lowest RF distances and highest TC, consistent with their length.

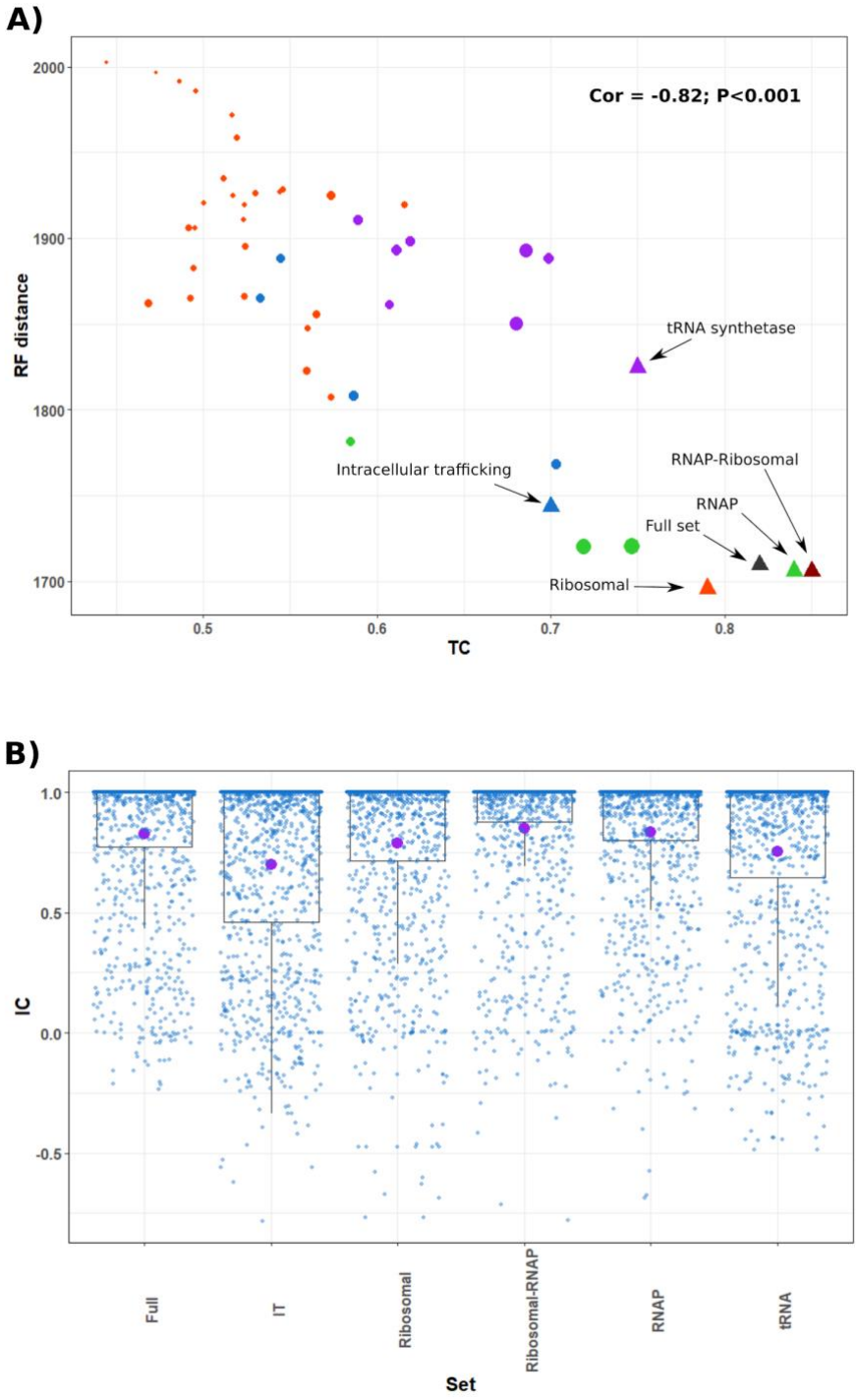

Figure 3.3 Relationship between tree certainty and Robinson-Foulds distance for individual markers and markers sets and IC estimated for marker sets. A) Pearson correlation of mean RF distance vs TC for Individual markers trees (circles) and marker sets trees (triangles). Colors: orange, ribosomal proteins; green, RNAP proteins; blue, intracellular trafficking; purple, tRNA proteins; grey, full set; red, RNAP-ribosomal set. Size of circles is equivalent to the median length of each COG or length of the final alignment (shortest = 98 aa; longest = 1392 aa). RF distance

values represent the mean pairwise distance B) IC (blue) and TC (purple) estimates for the maximum likelihood built from the concatenation of single-copy markers.

High mean RF distances are most likely the result of either orthologous gene displacement (OGD), which will lead to contrasting evolutionary histories in SCMs, or low phylogenetic signal, which will lead to topological differences in SCM trees that are merely the result of inadequate information for tree construction. Distinguishing between these two scenarios is critical because SCMs with low phylogenetic signal can still be used in concatenated alignments, where their phylogenetic signal can be considered additive rather than conflicting. Comparison of mean RF distances and TC values offers a possible way of distinguishing between OGD and low phylogenetic signal. For example, the tRNA-synthetase SCMs exhibited RF distances that are higher than expected given their relatively high TC values (Fig. 3.3A, Supplemental File 2). This pattern is consistent with the higher incidence of both ancient and recent orthologous gene displacement events that have previously been noted in tRNA-synthetases ((Wolf et al. 1999; Creevey et al. 2011); Fournier et al., 2015), which would result in a decoupling of the TC and RF values because they would have an evolutionary history distinct from the other SCMs (Wolf et al. 1999; Creevey et al. 2011). Indeed, inspection of individual SCM phylogenies revealed that tRNA-synthetases have experienced several inter-domain and inter-phylum OGD events (Supplemental File 3). In contrast to tRNA-synthetase genes, the relatively high RF values recovered for ribosomal proteins are likely due to their short length and low individual phylogenetic signal for these SCMs, rather than high incidence of OGD. The shortest ribosomal proteins had the lowest TC and the highest RF distances of all SCMs. In general, our TC and RF results are in agreement with the "Complexity Hypothesis" (Jain et al. 1999), which states that genes of the same structural system that are involved in informational processes (i.e., RNAP and ribosomal proteins) tend to undergo fewer OGD events.

3.3.2 Identifying the best-performing SCM set for interdomain phylogenetic reconstruction

We next evaluated trees constructed using concatenated alignments made from different SCM sets (Table 3.1). We evaluated alignments of all 41 SCMs (Full set) and SCM sets divided according to functional categories: ribosomal proteins (RP), RNAP subunits (RNAP), intracellular trafficking (IF), and tRNA synthetases (tRNA). Moreover, we also evaluated a concatenated set of both, ribosomal proteins and RNAP subunits (RNAP-RP set) because these SCMs had high phylogenetic congruence according to our previous analysis, and both belong to large multimeric complexes where OGD is less likely. All SCMs sets had high median bootstrap support (99-100%, Table 3.1), demonstrating the insufficiency of this metric for assessing differences between concatenated alignments. This finding is consistent with a previous report that bootstrap support provides misleadingly high confidence values for trees based on concatenated alignments (Salichos and Rokas 2013b). TC values provide a more robust metric for evaluating the tree quality (Table 3.1; Fig. 3.3A-3.3B): the tree built using the Full set and the RNAP-RP set had the highest TC values, whereas the most uncertain trees were obtained for the IT and tRNA genes sets. As expected, although individual ribosomal trees showed low certainty values, their concatenation in the RP set showed higher congruence (Table 3.1, 2A-2B), consistent with the view that these SCMs have low phylogenetic signal independently but can be effectively concatenated due to their consistent evolutionary histories. Importantly, the RNAP-RP set outperformed the full set of 41 markers despite having a shorter overall alignment length (Table 3.1, 2A-2B), likely because the IT and tRNA sets incorporate phylogenetic signals incongruent with the other SCMs. This is consistent with studies that have noted that these genes have higher rates of HGT than the other SCMs (Ciccarelli et al. 2006; Creevey et al. 2011). Overall, these results identify the RNAP-RP set best-performing SCM set for multi-domain phylogenetic reconstruction, underscore the importance of evaluating phylogenetic congruence when choosing SCMs for a concatenated

alignment, and demonstrate that the inclusion of additional SCMs does not necessarily improve phylogenetic accuracy.

Table 3.1 Statistics of phylogenetic trees built using the concatenation of SCM.

Marker set	Alignment length (aa)	Number of proteins	Model*	TC**	TC-PMSF***	Median bootstrap
Full set	16141	41	LG+R10	0.82	0.85	100
Intracellular trafficking	1964	4	LG+F+R10	0.70	0.76	99
Ribosomal	5197	27	LG+R10	0.79	0.80	100
tRNA	4993	7	LG+R10	0.75	0.79	100
RNAP	3987	3	LG+F+R10	0.84	-	100
Ribosomal-RNAP	9184	30	LG+R10	0.85	0.86	99

*Best-performing substitution model according to the BIC criterion
**TC = Tree certainty based on best ML tree vs bootstrap replicates
***TC-PMSF = mean TC resulting from five independent tree replicates using a site frequency matrix obtained from the C60 mixture model. This was not calculated for the RNAP marker set because the LG+F+R10 model was a better fit in this case.

We also evaluated the fit of different substitution models to see if this could explain our results. For individual SCMs the Bayesian Information Criterion (BIC) of the best fit substitution model increased linearly with protein length, as expected given that alignment length is used directly in the calculation of BIC. Similarly, BIC and TC were correlated, indicating once again that longer SCMs tend to have a higher phylogenetic signal. Interestingly, for concatenated

alignments the correlation between alignment length and BIC was upheld, but the relationship between BIC and TC was not evident (Fig. 3.4). The RNAP-Ribosomal, Ribosomal, and RNAP alignments all had higher TC than would be expected given their alignment length, indicating that the concatenation of SCMs with congruent phylogenetic signal effectively boosts the accuracy of phylogenetic reconstruction (Fig. 3.4). Analysis of the relationship between BIC, alignment length, and TC may therefore represent a complementary method for identifying sets of marker genes with congruent phylogenetic signals.

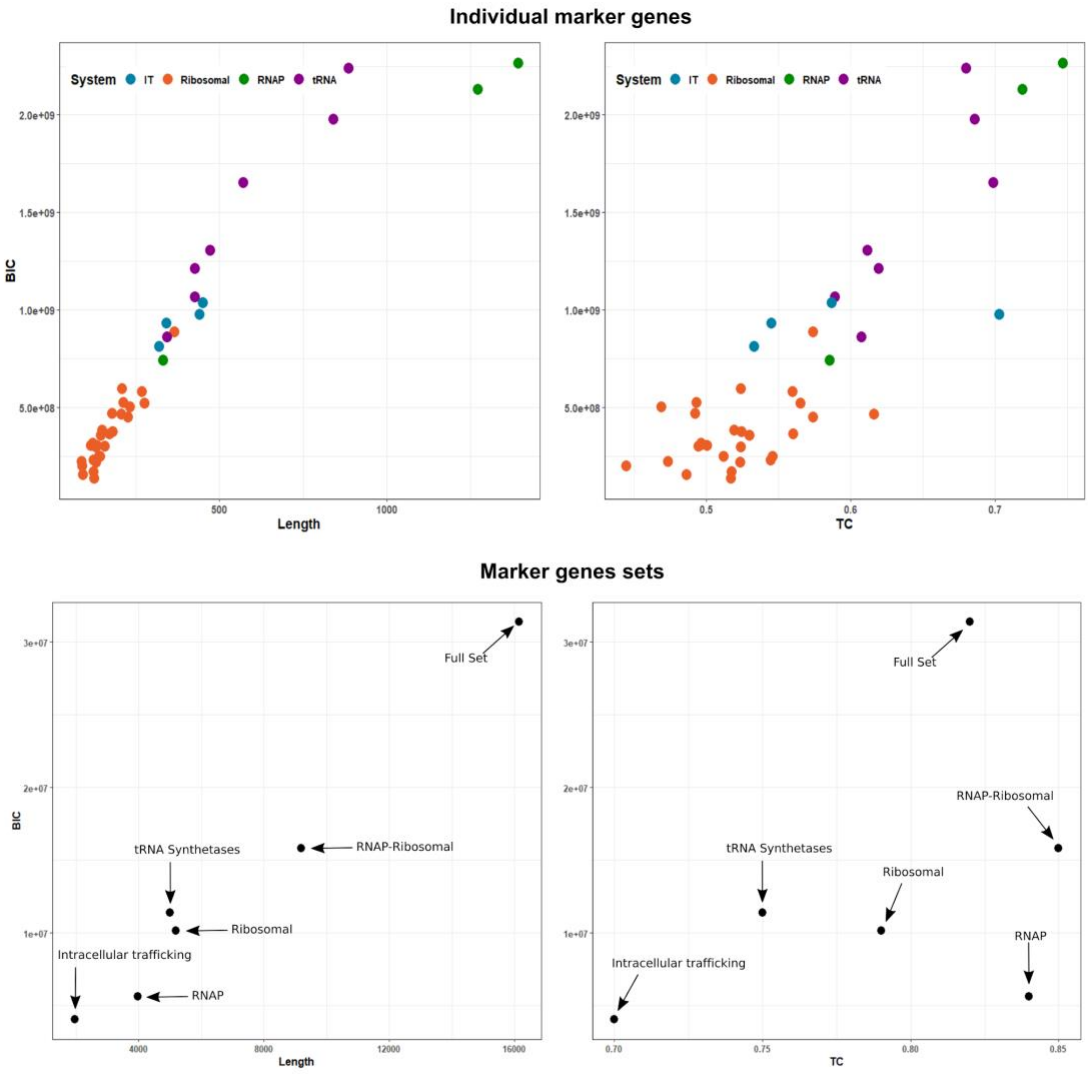

Figure 3.4 Relationship between substitution model fit based on the Bayesian Information Criterion (BIC) and alignment length, and BIC and tree certainty (TC) for individual marker genes and marker genes sets.

3.3.3 Evaluating the effect of taxon sampling in tree topology

Several studies have shown that an increase in taxon sampling can improve phylogenetic accuracy (Pollock et al. 2002; Zwickl and Hillis 2002; Jeffroy et al. 2006), and this strategy is commonly used as a solution to resolve unstable nodes in the tree of life (Young and Gillung 2020). In some cases it has been suggested that conflict among reported trees may result from striking differences

in taxonomic representation, however (Nasir et al. 2016; Cunha et al. 2017), and it remains unclear to what extent the oversampling of some taxa relative to others can deleteriously affect tree reconstruction. To test the effect of relative taxon sampling and taxonomic level selection on the certainty and topology of multi-domain phylogenies, we compared multi-domain trees constructed using different taxon evenness across phyla (Supplemental File 1), with the Gini index used to assess evenness at the phylum level (Supplemental File 6, see Methods). We performed this analysis on three genome sets in which representative genomes were selected at the Order, Family, and Genus level, according to the classifications of the Genome Taxonomy Database (GTDB). For these three genome sets (unbalanced datasets) we employed two strategies to increase taxonomic evenness: 1) we removed poorly represented phyla only (fewer than five genomes present for a given phylum) (partially unbalanced datasets), and 2) we removed both poorly represented phyla and also down-sampled over-represented phyla (balanced datasets) (Supplemental File 6; see Methods for details).

Our results demonstrate that taxon sampling markedly affects both TC values (Table 3.2, Supplemental File 16) and overall tree topology (Fig. 3.6, Supplemental File 7-14). Similar to our benchmarking of different SCM sets, the trees could not be distinguished based on bootstrap support (100% median support for all trees). The TC values of the trees at all the taxonomic levels improved when both poorly represented taxa are removed, and overrepresented groups are downsampled (Table 3.2; Supplemental File 16). Importantly, in all cases we found an increase in TC values when both poorly represented groups were removed and over-represented phyla were down-sampled compared to if only poorly-sampled groups were removed (Table 3.2, Supplemental File 16). This demonstrates that, perhaps counterintuitively, removal of some genomes can actually improve tree quality, and that larger genome sets do not necessarily improve

phylogenetic inference. We surmise that taxon oversampling lowers TC values in part because of complications that arise at the alignment stage; an alignment that is highly over-represented in certain groups would not necessarily be expected to align homologous regions in all taxa equally well, especially if some were highly divergent compared to the over-sampled groups. The incorporation of poorly sampled groups, or the oversampling of some groups relative to others, may therefore lead to long-branch artefacts that can influence the placement of other groups in the tree (Felsenstein 1978; Bergsten 2005). Balanced sampling improved TC values at all levels (Order, Family, and Genus), underscoring the important effect that balanced taxon sampling has when studying deeply divergent nodes. For studies specifically focusing on lineages for which only few genomes are available, we recommend including these genomes in an otherwise balanced tree. This approach would represent a compromise that would both mitigate the deleterious effects of unbalanced taxon sampling while still allowing for phylogenetic placement of the lineage under examination.

Table 3.2 Statistics of phylogenetic trees built using balanced, partially unbalanced, and unbalanced genomes datasets. An RNAP-ribosomal concatenated alignment was used to reconstruct all trees. All trees were built using the LG+R10 substitution model after selection according to the BIC criterion.

Sampling strategy	Genomes included	Alignment length (aa)	TC*	TC10**	Gini index	Median bootstrap
Order even	620	9,203	0.83	0.83	0.44	100
Partially uneven order[#]	722	9,146	0.75	0.71	0.5	100
Order uneven	834	9,298	0.77	0.71	0.70	100
Family even	1,650	9,625	0.86	0.83	0.59	100
Partially uneven family[#]	1,925	9,407	0.81	0.77	0.64	100

Family uneven	2,023	-	0.78	0.72	0.75	100
Genus even	4,340	9,845	0.88	0.79	0.67	100
Partially uneven genus[#]	7,260	9,731	0.8	0.77	0.78	100
Genus uneven	7,325	9,730	0.84	0.8	0.84	100

*TC = Tree certainty based on best ML tree vs bootstrap replicates.
**TC calculated based on the 10% closest nodes to the root.
Partially uneven trees have had low-abundance phyla removed, but over-represented phyla have not been down-sampled.

In addition to evaluating overall TC values for our order-, family-, and genus-level trees, we also sought to examine which contained the highest internode certainty (IC) values specifically for deep-branching nodes and would therefore be most appropriate for examining inter-phyla evolutionary relationships. For this we estimated the TC values of our trees based on only 10% of the nodes closest to the root (TC10 metric, Table 3.2, Supplemental File 16). Our estimates show that although the balanced Genus tree had the highest TC when considering all nodes (TC of 0.88), the balanced Order and Family trees showed the highest certainty in deep nodes (TC10 of 0.83, Table 3.2, Supplemental File 16), indicating the latter two trees are more appropriate for the analysis of inter-phylum relationships.

3.3.4 Analysis of the highest quality multi-domain tree

After benchmarking the best-performing SCM set and the appropriate level of taxon sampling for accurate phylogenetic reconstruction, we sought to examine the evolutionary relationships revealed by our best-performing trees. The balanced Family and Order trees both had the highest TC10 values, we focus our analysis primarily on the former because this phylogeny contains the largest representation of bacterial and archaeal lineages (Fig. 3.6A-3.6B). The salient features we

discuss below were also shared with the balanced Order-level tree, however (Supplemental File 9).

Our phylogeny indicates that the root in Bacteria lies between *Thermotogota* and the rest of the bacterial phyla (Fig. 3.6A-3.6B, Supplemental File 16). Although some analyses state that the basal-branching placement of *Thermotogota* may be an artefact derived from the transfer of archaeal and Actinobacterial genes related with thermophily (Nesbo et al. 2001; Zhaxybayeva et al. 2009; Cavalier-Smith 2010), the topology we report is in agreement with previous studies of the early-branching position of the group (Woese 1987; Bachleitner et al. 1989; Woese et al. 1990), and the potential hyperthermophile of early life (Gaucher et al. 2010). Although early phylogenetic studies showed that *Aquificota* is a deep-branching group (Burggraf et al. 1992; Boussau et al. 2008), analyses based on the presence and absence of Conserved Signature Indels (CSIs) in highly conserved genes have suggested that *Aquificota* is a late-branching group within bacteria (Griffiths and Gupta 2004; Rosenberg et al. 2014), which agrees with their placement in our tree (Fig. 3.6A-3.6B). In addition to the *Thermotogota*, the *Synergistota* were also basal-branching in our tree, and all other bacteria could be divided into two groups corresponding to the superphyla Terrabacteria and Gracilicutes (Battistuzzi et al. 2004; Cavalier-Smith 2006). Terrabacteria include the *Cyanobacteria, Firmicutes, Actinobacteriota, Patescibacteria,* and *Chloroflexota*, among other phyla, while the Gracilicutes include a wide variety of lineages including the *Proteobacteria, Acidobacteria, Bacteriodota, Spirochaetes, Planctomycetota,* and *Verrucomicrobiot*a (Fig. 3.6A-6B).

Our results indicate that the *Patescibacteria* (also called the Candidate Phyla Radiation or CPR) are a derived group that is sister to the *Chloroflexota*, consistent with two recent studies (Coleman et al. 2020; Taib et al. 2020). This is in contrast to other studies that placed this group as either basal-branching or falling outside of the Terrabacteria (Hug et al. 2016; Parks et al. 2017;

Castelle et al. 2018; Méheust et al. 2019). Previous studies have suggested that the inclusion of a distant outgroup like archaea may explain the artifactual basal branching of the *Patescibacteria* (Coleman et al. 2020; Taib et al. 2020), but our result indicate that it is possible to obtain a derived placement of this group even with the inclusion of archaea. Early findings of a basal branching placement of the *Patescibacteria* may have been influenced by unbalanced taxon sampling, indeed, our unbalanced family tree showed that *Patescibacteria* was placed near to the root (Fig. 3.5A, Supplemental File 16) in contrast to our balanced tree which had the same number of *Patescibacteria* genomes (Fig. 3.5B), and our partially unbalanced order tree showed the *Patescibacteria* as basal branching but with a low certainty (Supplemental File 10 and Supplemental File 16). In support of our result, a recent study found similar genomic signatures of monoderm envelope structure in the *Patescibacteria* and *Chloroflexota* and attributed this to the possible transition from diderm to monoderm envelope structures in their common ancestor (Taib et al. 2020).

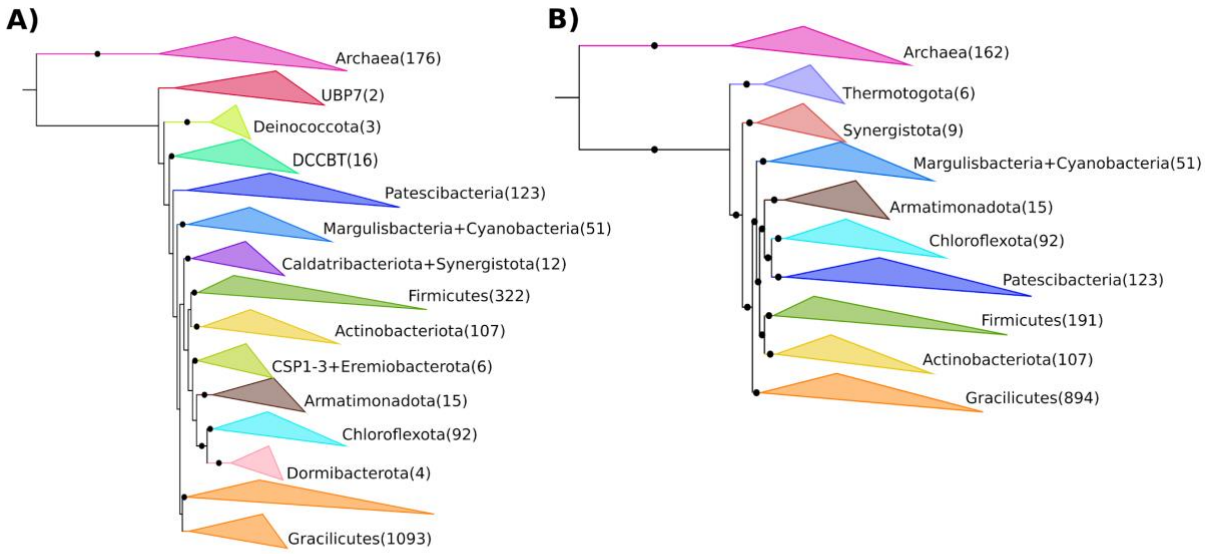

Figure 3.5 Comparison of the phylogenetic placement of Terrabacteria phyla using different sampling strategies. A) Maximum likelihood phylogenetic tree built using an unbalanced taxonomic representation at the family level. B) Maximum likelihood phylogenetic tree built using a balanced taxonomic representation at the family level. Abbreviations: DCCBT; *Dictyoglomota, Coprothermobacterota, Caldisericota, Bipolaricaulota, Thermotogota*. Number of genomes used for each group is indicated in parentheses. Black circles over branches indicate IC>0.5.

Our balanced Family-level tree places the root of Archaea between the DPANN and the rest of Archaea (Fig. 3.6A-3.6B), and this placement has been reported previously in a study based on gene tree-species tree reconciliation (Williams et al., 2017). Since the discovery of the first DPANN representative (Rinke et al. 2013), the placement of this group in the archaeal TOL has been uncertain because of their extreme genome reduction and long branch lengths (Dombrowski et al. 2019). The *Patescibacteria*, another lineage with small genomes and long branches, was also reported to have basal placement in the TOL (Hug et al. 2016), but subsequent work and our findings here suggest that they are a sister lineage to the *Chloroflexota* (Taib et al. 2020; Coleman et al. 2021). The basal placement of DPANN must therefore be treated with some caution. Although the basal branching position and monophyly of the DPANN shows high certainty, the overall low taxon sampling available for archaea relative to bacteria gives us cause for doubt of this result, and it is possible that this is an artifact caused by similar substitution bias and homoplasies that may cause the grouping of unrelated lineages and LBA (Brochier et al. 2005; Philippe and Roure 2011; Petitjean et al. 2014; Gouy et al. 2015; Aouad et al. 2018). This remains a distinct possibility because most of the DPANN described so far share similarities in their host-dependent ectosymbiotic lifestyle and residence in deep subsurface environments (He et al. 2021). Previous studies have shown that the phylogenetic resolution of DPANN is sensitive to the taxa included (Williams et al. 2017; Dombrowski et al. 2019), and we therefore speculate that additional sequencing of archaeal diversity will be necessary to increase the genomic representation of this domain and clarify the placement of the DPANN.

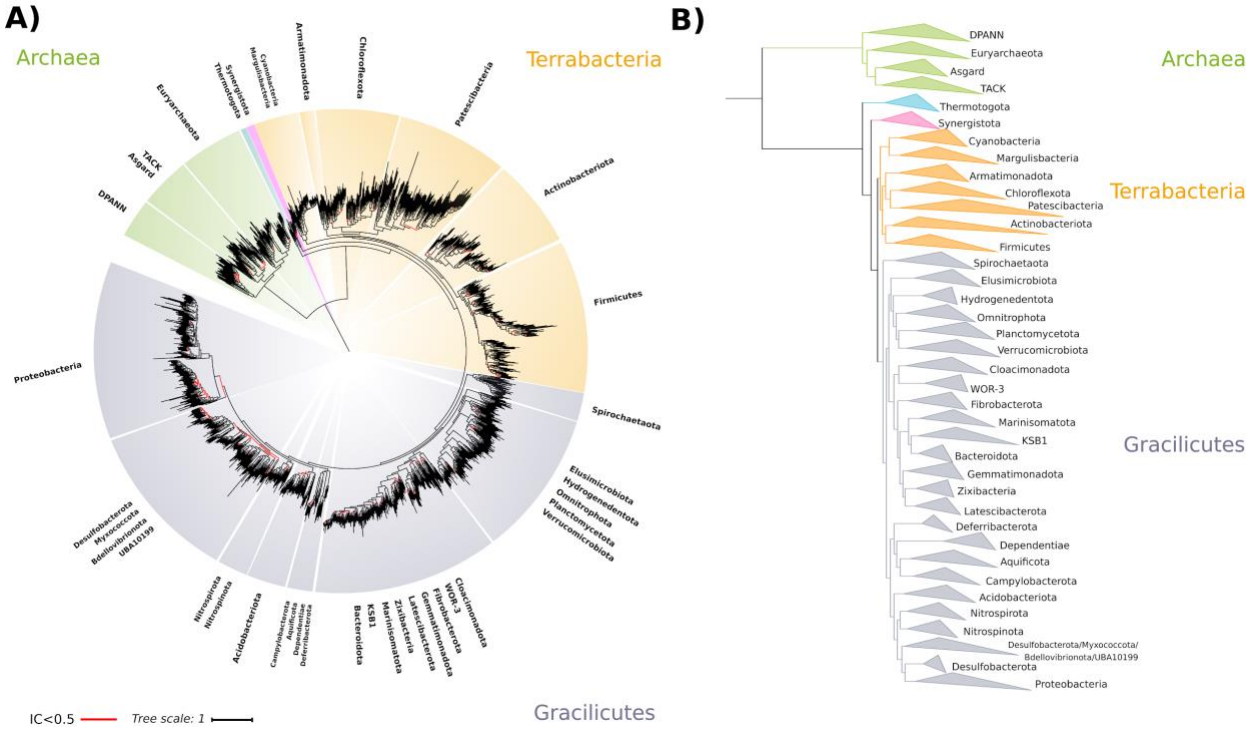

Figure 3.6 Rooted interdomain tree built using a balanced taxonomic representation. A) Maximum likelihood tree built using the concatenation of 30 RNAP subunits and ribosomal protein sequences and the substitution model LG+R10. A balanced sampling strategy was used to select genomes at the family level according to the GTDB (see materials and methods for details). An IC>0.5 indicates that more than 80% of the bootstrap replicate trees support the node shown (Salichos et al. 2014). Euryarchaeota is represented by three different phyla on the GTDB; Methanobacteriota, Thermoplasmatota, and Halobacteriota, TACK Archaea by one phylum; *Thermoproteota*, and the DPANN superphylum by the phyla *Iainarchaeota and Nanoarchaeota* B) Detailed phylogenetic relationships among the phyla studied.

For the rest of the Archaea, our tree recovered the monophyly of *Euryarchaeota* obtained previously (Petitjean et al. 2014; Williams et al. 2017), which contrasts with a study that suggests a paraphyletic origin of the group (Raymann et al. 2015). Interestingly, all our unbalanced trees showed paraphyly in *Euryarchaeota* (Supplemental File 8, Supplemental File 11, Supplemental File 13, Supplemental File 16), raising the possibility that the topology of this group is sensitive to taxon sampling. The placement of Asgard archaea at the base of TACK showed the maximal certainty (Fig. 3.6A; IC =1) and was found in all our trees independently of the sampling strategy

(Fig. 3.6A-3.6B, Supplemental File 7-14). This finding is supported by recent studies (Spang et al. 2015; Adam et al. 2017; Williams et al. 2017; Zaremba-Niedzwiedzka et al. 2017). Some studies have suggested that the placement of Asgard archaea near TACK may be due in part to unbalanced taxon sampling (Nasir et al. 2016; Cunha et al. 2017), but our results are inconsistent with this view.

By using tree certainty metrics, we are also able to identify deep-branching nodes for which the topology remains uncertain, and for which additional analyses will be necessary to resolve evolutionary relationships. In particular we observed high uncertainty in the cluster formed by *Desulfobacterota* (paraphyletic), *Myxococcota, Bdellovibrionota* (which was paraphyletic in our tree), and UBA10199 (Fig. 3.6A-3.6B). These newly established phyla previously belonged to the *Proteobacteria*, therefore it is possible that these groups need revision or reclassification.

3.4 Outlook

Whole-genome phylogenies have become increasingly common in recent years owing to the large volume of genomic data that is now available for diverse bacteria and archaea, and it is now common to use hundreds of concatenated protein sequences to infer the evolutionary relationships of microbial taxa. There are several reasons to doubt that the use of more genes and genomes necessarily improves the quality of the resulting trees, however. For example, some genes may have undergone orthologous gene displacement (OGD), and their evolutionary history will therefore conflict with the other genes in the concatenated alignment, in effect creating phylogenetic noise. Moreover, the common phylogenetic confidence metrics, such as bootstrap support, provide misleadingly high values when applied to long concatenated alignments, and it is

therefore unclear if adding more genes truly increases phylogenetic accuracy or merely artefactually increases support values. Lastly, given the different number of genomes available for different phyla of bacteria and archaea, it is unclear if the relative oversampling of some lineages over others can negatively affect the quality of phylogenetic inference.

Our findings show that both SCM selection and taxon sampling strategies are critical considerations that impact the quality of multi-domain phylogenetic trees constructed from concatenated alignments. We find that selecting SCMs with congruent phylogenetic signals improves the performance of resulting trees generated from concatenated alignments, and that more SCMs do not necessarily improve tree quality. Moreover, we found that taxon sampling can dramatically impact the topology of resulting trees, and that over-sampling of some lineages relative to others can introduce topological inconsistencies and yield nodes with low certainty. Taken together, these results show that more genes and genomes do not necessarily improve phylogenetic inference, and that the use of phylogenetically congruent SCMs on a balanced taxon set is likely to yield the best results. Many of these issues have been previously recognized in phylogenomic analyses of eukaryotes, in particular animals and yeast, suggesting these are common issues that arise in evolutionary analyses of different groups at disparate phylogenetic scales (Rokas and Carroll 2005; Nishihara et al. 2007; Philippe et al. 2011; Salichos and Rokas 2013b).

Our results have several implications for investigations of the TOL. Firstly, our finding that more genes and genomes do not necessarily improve phylogenetic accuracy is important considering that phylogenies constructed from large taxon and SCM sets can hinder the use of complex models and effectively restrict researchers to the use of a small set of tools that are

optimized for speed (i.e., FastTree). Although this issue will only become more pronounced in the future as more genomes continue to be sequenced, our results indicate that down-sampling over-represented groups will both alleviate computational burdens, allow for more complex phylogenetic models to be employed, and ultimately improve tree quality, in particular at deep-branching nodes. Secondly, our results show that phyla for which only few genomes are available will likely have uncertain phylogenetic placement given the inability to include them in balanced trees. This is unavoidable to a large extent and underscores the importance of diversity-based sequencing efforts that expand the genomic representation of poorly-characterized phyla. Thirdly, although it has been proposed that the inclusion of both domains may lead to artifacts due to the evolutionary distance between bacteria and archaea (Coleman et al. 2020), our analysis shows that high fidelity multi-domain trees can be constructed using certain SCM sets and taxon sampling strategies. Lastly, it has recently been suggested that small SCM sets that include many ribosomal proteins are undesirable in multidomain phylogenetic analyses due to their large inter-domain divergence (Zhu et al, 2019), but our results conflict with this view and suggest that the addition of more genes with potentially discordant evolutionary histories will often increase noise and reduce tree quality. Indeed, the long interdomain distance between some SCMs has long been considered to be a signature of their presence in the LUCA (Woese, 1998, Forterre, 2006), which would make them particularly useful markers for analysis of ancient diversification events.

Although our analyses addressed several difficulties that arise in the generation of phylogenetic trees containing both bacteria and archaea (i.e., SCM selection and taxa sampling), other biological factors may still limit the accuracy of phylogenetic inference. For example, substitution models may have difficulty dealing with high evolutionary rates or biased amino acid composition of SCMs, which may in turn lead to long-branch artefacts. We suspect these issues

are at play in the DPANN group, which may lead to their artifactual placement at the base of the archaea in both our trees and those of other studies (Rinke et al. 2013; Hug et al. 2016; Parks et al. 2017; Williams et al. 2017; Dombrowski et al. 2019). Further work will therefore be needed to address these complications, potentially through the developments of additional statistical models that account for these possible biases or through detailed analyses of indels or other phylogenetic markers that are useful for the placement of specific lineages.

3.5 Material and methods

3.5.1 Assessing the congruence of individual marker genes

In order to evaluate the phylogenetic certainty of 41 marker genes commonly used to build prokaryotic phylogenies (Supplemental File 2), we compiled a genomic dataset encompassing a broad diversity of bacteria and archaea. We obtained the best representative genome for each family available on the Genome Taxonomy Database (GTDB) (Release 05-RS95; 17th July 2020) (Chaumeil et al. 2019). The selection criteria included genome completeness, contamination, N50, and the presence of all the marker genes tested, totalling 1119 families (Supplemental File 1). The open reading frames (ORF) obtained from the GTDB were compared to the HMMs of the 41 marker genes using the hmmsearch tool available in HMMER v. 3.2.1 (Eddy 2011) with a specific score cutoff for each marker gene (Supplemental File 2). To generate a reproducible workflow and address the fragmentation of COG0085 and COG0086 orthologs into multiple genes, we developed a custom python program (MarkerFinder) (Supplemental File 16). We previously used an earlier version of this tool to resolve evolutionary relationships in the *Thaumarchaeota* (Aylward and Santoro 2020). Once annotated, marker genes were aligned using Clustal Omega v. 1.2.3 with the default parameters (Sievers and Higgins 2018) and trimmed with trimAl v1.4.rev15 (Capella-Gutiérrez et al. 2009). Maximum likelihood phylogenetic trees were estimated using IQ-

TREE v1.6.12 (Nguyen et al. 2015) with the options -MFP to find the best-fitting substitution model available under the BIC criterion (Kalyaanamoorthy et al. 2017) (the best model for each marker gene is reported in Supplemental File 2), and -bb 1000 to obtain 1000 ultrafast bootstraps. The resulting trees were manually inspected on iTOL (Letunic and Bork 2019) to identify topologies suggestive of LGT (Supplemental File 3). We additionally analyzed the prevalence of these marker genes in 1650 genomes (balanced family dataset, Supplemental File 1) as well as RecA (COG0468), elongation factor G (COG0480), and elongation factor TU (COG0050) (Supplemental File 15).

The congruence of each marker gene tree was assessed by calculating the "Tree Certainty" metric (TC). The TC represents the mean of all the "Internode Certainty" values (IC), an estimate that assesses the degree of conflict of each internal node in a given tree by calculating Shannon's Measure of Entropy (Shannon 1948; Salichos and Rokas 2013b; Kobert et al. 2016b). In contrast to other congruence and support estimates alternative to bootstrap, the IC index reflects to which degree the most favored bipartition is contested (Kobert et al. 2016b). Estimates of IC and TC indices were achieved with the option raxmlHPC implemented on RAxML v8.2.X (Stamatakis 2014). Additionally, we estimated the Robinson and Foulds distance (RF) for each pair of trees (Robinson and Foulds 1981) using IQ-TREE (Nguyen et al. 2015). The RF metric calculates the distance between phylogenetic trees by counting the number of topological changes needed to convert one tree into the other (Robinson and Foulds 1981).

3.5.2 Phylogenetic congruence of concatenated marker genes sets

In addition to the assessment of individual genes congruence, we analyzed the phylogenetic congruence of the maximum likelihood phylogenetic tree built based on the concatenation of all

the marker genes, as well as phylogenetic trees built from the alignment of concatenated marker genes with related functions. Individual trimmed sequences resulting from the previous step were concatenated and maximum likelihood trees and certainty values were estimated as described above. The best model for each phylogenetic is reported in Table 3.2.

3.5.3 Balanced sampling across prokaryotes diversity

To evaluate the effect of taxon sampling on the certainty and topology of the prokaryotic tree of life, we constructed three genomic datasets by selecting representative genomes from the GTDB at the Order, Family, and Genus level. The "best" representative genome at a given taxonomic level was chosen from the GTDB based on its estimated completeness, contamination, and N50. Genomes representatives were filtered based on a completeness cutoff of 70% and the presence of at least 25 out of the 30 marker genes belonging to the RNAP + Ribosomal marker set. In addition, we only used genomes where both COG0086 and COG0085 could be found, because these RNAP subunits are particularly long and have strong phylogenetic signal (Fig. 3.2), and their absence would therefore have a pronounced impact on alignment quality. To assess the evenness of the genome sets, genomes were grouped at the phylum level, and phylum-level distributions were evaluated using the Gini Index (GI). The GI is a widely used metric for equality that varies from 0 (full equality) to 1 (fully unequal) (Gini 1912). Thus, if applied to genome sets, the GI is a systematic metric that describes the level of taxonomic evenness. Once we calculated the GI on the initial Order, Family, and Genus-level genome sets (referred here to as the unbalanced datasets), we increased taxonomic evenness using two methods: 1) we removed phyla that contained < 5 representatives (the partially unbalanced datasets), and 2) we both removed phyla that contained < 5 representatives and down-sampled the most over-represented phyla (referred to

as the balanced datasets). In the second case, we performed phylum-level downsampling using the following equation:

$$S = N(1 - (\frac{N - Quant}{N})^2)$$

Where *N* represents the number of genomes for each phylum and *Quant* the quantile 0.90 for the taxonomic classification of the dataset. We refer to these datasets as "balanced datasets" (Supplemental File 6).

In our analysis, *S* represents the final number of genomes for each downsampled phyla. Genomes for downsampled phyla were selected randomly to ensure an equal representation of all clades within each phylum in the final dataset. Our final data consisted in three unbalanced and three balanced datasets (two for each taxonomic level evaluated) (Supplemental File 1). In addition, we examined the effect of poorly-represented phyla on trees congruence by removing phyla with fewer than five genomes from each unbalanced dataset (Supplemental File 1) (referred here to as "partially unbalanced datasets"). Marker genes belonging to RNAP subunits and ribosomal proteins were identified and concatenated as described previously, and maximum likelihood trees built using the option -m MFP and -b 1000. Additionally, tree certainty values were estimated and each tree topology explored manually using iTOL.

3.5.4 Assessment of Model Fit on Tree Certainty

In order to assess if tree certainty values could be inflated due to substitution model misspecifications, we explored the relationship between tree certainty, alignment length, and

substitution model fit (BIC) for both individual trees and marker gene sets. We plotted TC vs BIC and BIC vs median length of individual marker genes and alignment length for marker sets (Fig. 3.4), which suggest that tree certainty is not directly affected by substitution model fit. In addition, we assessed the impact of model complexity on the TC of our SCM sets by repeating the model selection analysis but including the C10-C60 mixture models. The BIC indicated that the C60 mixture model was the best fitting model for all the SCM sets except the RNAP set, for which the LG+R10 model was the best fit. We obtained a site-frequency matrix according to the PMSF method (Wang et al. 2018) using the C60 mixture model and then ran five independent maximum likelihood trees on each SCM alignment. Results of this analysis are presented in Supplemental File 4. Our results showed that the RP set once again had the highest TC, supporting our previous findings that the RP is the best SCM set tested in our study (Supplemental File 4). Similarly, we tested whether the results observed in our balanced sampling analysis were caused by model misspecifications by using the C60 mixture model on our balanced, partially unbalanced, and unbalanced sets at the order level (best fitting model according to the BIC criterion). Although the TC of unbalanced trees improved when using a more complex model, the balanced tree still showed the highest tree certainty, suggesting that our results are consistent and independent of the substitution model used (Supplemental File 5). Lastly, we reran our balanced family tree using the C60 model to confirm that substitution model complexity did not adversely affect tree topology, and we did not observe substantial topological changes or improvements in the TC value (Supplemental File 17).

3.6 Data availability

The supplemental files of this chapter are available of the Figshare collection: https://doi.org/10.6084/m9.figshare.c.6240942.v1

3.7 Acknowledgments

We acknowledge the use of the Virginia Tech Advanced Research Computing Center for bioinformatic analyses performed in this study. This work was supported by grants from the Institute for Critical Technology and Applied Science and the NSF (IIBR-1918271) and a Simons Early Career Award in Marine Microbial Ecology and Evolution to F.O.A. We thank members of the Aylward lab for helpful comments on an earlier version of this manuscript.

3.8 References

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

Altermann W, Kazmierczak J. 2003. Archean microfossils: a reappraisal of early life on Earth. Res. Microbiol. 154:611–617.

Anon. 2011. Microbial phylogeny and diversity: Small subunit ribosomal RNA sequence analysis and beyond. Microbiol. Res. 166:99–110.

Aouad M, Taib N, Oudart A, Lecocq M, Gouy M, Brochier-Armanet C. 2018. Extreme halophilic archaea derive from two distinct methanogen Class II lineages. Mol. Phylogenet. Evol. 127:46–54.

Aylward FO, Santoro AE. 2020. Heterotrophic Thaumarchaea with Small Genomes Are Widespread in the Dark Ocean. mSystems [Internet] 5. Available from: http://dx.doi.org/10.1128/mSystems.00415-20

Bachleitner M, Ludwig W, Stetter KO, Schleifer KH. 1989. Nucleotide sequence of the gene coding for the elongation factor Tu from the extremely thermophilic eubacterium Thermotoga maritima. FEMS Microbiol. Lett. 48:115–120.

Battistuzzi FU, Feijao A, Hedges SB. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evol. Biol. 4:44.

Bergsten J. 2005. A review of long-branch attraction. Cladistics [Internet] 21:163–193. Available from: http://dx.doi.org/10.1111/j.1096-0031.2005.00059.x

Berkemer SJ, McGlynn SE. 2020. A New Analysis of Archaea–Bacteria Domain Separation: Variable Phylogenetic Distance and the Tempo of Early Evolution. Mol. Biol. Evol. 37:2332–2340.

Bleidorn C. 2017. Sources of Error and Incongruence in Phylogenomic Analyses. In: Bleidorn C, editor. Phylogenomics: An Introduction. Cham: Springer International Publishing. p. 173–193.

Boussau B, Guéguen L, Gouy M. 2008. Accounting for horizontal gene transfers explains

conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. BMC Evol. Biol. 8:272.

Brochier C, Gribaldo S, Zivanovic Y, Confalonieri F, Forterre P. 2005. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? Genome Biol. 6:R42.

Burggraf S, Olsen GJ, Stetter KO, Woese CR. 1992. A phylogenetic analysis of Aquifex pyrophilus. Syst. Appl. Microbiol. 15:352–356.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Castelle CJ, Brown CT, Anantharaman K, Probst AJ, Huang RH, Banfield JF. 2018. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat. Rev. Microbiol. 16:629–645.

Cavalier-Smith T. 2006. Rooting the tree of life by transition analyses. Biol. Direct 1:19.

Cavalier-Smith T. 2010. Deep phylogeny, ancestral groups and the four ages of life. Philos. Trans. R. Soc. Lond. B Biol. Sci. 365:111–132.

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics [Internet]. Available from: http://dx.doi.org/10.1093/bioinformatics/btz848

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287.

Coleman GA, Davín AA, Mahendrarajah TA, Szánthó LL, Spang A, Hugenholtz P, Szöllősi GJ, Williams TA. 2021. A rooted phylogeny resolves early bacterial evolution. Science [Internet] 372. Available from: http://dx.doi.org/10.1126/science.abe0511

Coleman GA, Davín AA, Mahendrarajah T, Spang A, Hugenholtz P, Szöllősi GJ, Williams TA. 2020. A rooted phylogeny resolves early bacterial evolution. Available from: http://dx.doi.org/10.1101/2020.07.15.205187

Creevey CJ, Doerks T, Fitzpatrick DA, Raes J, Bork P. 2011. Universally distributed single-copy genes indicate a constant rate of horizontal transfer. PLoS One 6:e22099.

Cunha VD, Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P. 2017. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLOS Genetics [Internet] 13:e1006810. Available from: http://dx.doi.org/10.1371/journal.pgen.1006810

Da Cunha V, Gaia M, Nasir A, Forterre P. 2018. Asgard archaea do not close the debate about the universal tree of life topology. PLoS Genet. 14:e1007215.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics [Internet] 6:361–375. Available from: http://dx.doi.org/10.1038/nrg1603

Dombrowski N, Lee J-H, Williams TA, Offre P, Spang A. 2019. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. FEMS Microbiol. Lett. [Internet] 366. Available from: http://dx.doi.org/10.1093/femsle/fnz008

Doolittle WF. 1999. Phylogenetic Classification and the Universal Tree. Science [Internet] 284:2124–2128. Available from: http://dx.doi.org/10.1126/science.284.5423.2124

Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput. Biol. 7:e1002195.

Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Systematic Biology [Internet] 27:401–410. Available from: http://dx.doi.org/10.1093/sysbio/27.4.401

Gadagkar SR, Rosenberg MS, Kumar S. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J. Exp. Zool. B Mol. Dev. Evol. 304:64–74.

Gaucher EA, Kratzer JT, Randall RN. 2010. Deep phylogeny--how a tree can help characterize early life on Earth. Cold Spring Harb. Perspect. Biol. 2:a002238.

Gini C. 1912. Variabilita e mutabilita. Studi economicoaguridici delle facoltta di giurizprudenza dell. Universite di Cagliari III Parte II.

Gouy R, Baurain D, Philippe H. 2015. Rooting the tree of life: the phylogenetic jury is still out. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370:20140329.

Gribaldo S, Philippe H. 2002. Ancient Phylogenetic Relationships. Theoretical Population Biology [Internet] 61:391–408. Available from: http://dx.doi.org/10.1006/tpbi.2002.1593

Griffiths E, Gupta RS. 2004. Signature sequences in diverse proteins provide evidence for the late divergence of the Order Aquificales. Int. Microbiol. 7:41–52.

He C, Keren R, Whittaker ML, Farag IF, Doudna JA, Cate JHD, Banfield JF. 2021. Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. Nature Microbiology [Internet]. Available from: http://dx.doi.org/10.1038/s41564-020-00840-5

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, et al. 2016. A new view of the tree of life. Nat Microbiol 1:16048.

Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. U. S. A. 96:3801–3806.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14:587–589.

Klenk H-P, Göker M. 2010. En route to a genome-based classification of Archaea and Bacteria?

Syst. Appl. Microbiol. 33:175–182.

Kobert K, Salichos L, Rokas A, Stamatakis A. 2016a. Computing the Internode Certainty and Related Measures from Partial Gene Trees. Mol. Biol. Evol. 33:1606–1617.

Kobert K, Salichos L, Rokas A, Stamatakis A. 2016b. Computing the Internode Certainty and Related Measures from Partial Gene Trees. Mol. Biol. Evol. 33:1606–1617.

Konstantinidis KT, Tiedje JM. 2007. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Curr. Opin. Microbiol. 10:504–509.

Lerat E, Daubin V, Moran NA. 2003. From Gene Trees to Organismal Phylogeny in Prokaryotes:The Case of the γ-Proteobacteria. PLoS Biology [Internet] 1:e19. Available from: http://dx.doi.org/10.1371/journal.pbio.0000019

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47:W256–W259.

Méheust R, Burstein D, Castelle CJ, Banfield JF. 2019. The distinction of CPR bacteria from other bacteria based on protein family content. Nat. Commun. 10:4173.

Nasir A, Kim KM, Da Cunha V, Caetano-Anollés G. 2016. Arguments Reinforcing the Three-Domain View of Diversified Cellular Life. Archaea 2016:1851865.

Nesbo CL, L'Haridon S, Stetter KO, Ford Doolittle W. 2001. Phylogenetic Analyses of Two "Archaeal" Genes in Thermotoga maritima Reveal Multiple Transfers Between Archaea and Bacteria. Molecular Biology and Evolution [Internet] 18:362–375. Available from: http://dx.doi.org/10.1093/oxfordjournals.molbev.a003812

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Nishihara H, Okada N, Hasegawa M. 2007. Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biol. 8:R199.

Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542.

Petitjean C, Deschamps P, López-García P, 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 DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Philippe H, Laurent J. 1998. How good are deep phylogenetic trees? Current Opinion in Genetics & Development [Internet] 8:616–623. Available from: http://dx.doi.org/10.1016/s0959-437x(98)80028-2

Philippe H, Roure B. 2011. Difficult phylogenetic questions: more data, maybe; better methods, certainly. BMC Biol. 9:91.

Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51:664–671.

de Queiroz A, Gatesy J. 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22:34–41.

Raymann K, Brochier-Armanet C, 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.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature [Internet] 499:431–437. Available from: http://dx.doi.org/10.1038/nature12352

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Mathematical Biosciences [Internet] 53:131–147. Available from: http://dx.doi.org/10.1016/0025-5564(81)90043-2

Rokas A, Carroll SB. 2005. More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy. Mol. Biol. Evol. 22:1337–1344.

Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F eds. 2014. The Prokaryotes: Other Major Lineages of Bacteria and The Archaea. Springer, Berlin, Heidelberg

Salichos L, Rokas A. 2013a. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Salichos L, Rokas A. 2013b. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31:1261–1271.

Segata N, Börnigen D, Morgan XC, Huttenhower C. 2013. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat. Commun. 4:2304.

Shannon CE. 1948. A Mathematical Theory of Communication. Bell System Technical Journal [Internet] 27:379–423. Available from: http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Science [Internet] 27:135–145. Available from: http://dx.doi.org/10.1002/pro.3290

Simmons MP, Gatesy J. 2016. Biases of tree-independent-character-subsampling methods. Mol. Phylogenet. Evol. 100:424–443.

Simon C. 2020. An Evolving View of Phylogenetic Support. Syst. Biol. [Internet]. Available from: http://dx.doi.org/10.1093/sysbio/syaa068

Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJG. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521:173–179.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stott CM, Bobay L-M. 2020. Impact of homologous recombination on core genome phylogenies. BMC Genomics 21:829.

Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, et al. 2013. Metagenomic species profiling using universal phylogenetic marker genes. Nat. Methods 10:1196–1199.

Taib N, Megrian D, Witwinowski J, Adam P, Poppleton D, Borrel G, Beloin C, Gribaldo S. 2020. Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. Nat Ecol Evol 4:1661–1672.

Wang H-C, Minh BQ, Susko E, Roger AJ. 2018. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Syst. Biol. 67:216–235.

Werner F, Grohmann D. 2011. Evolution of multisubunit RNA polymerases in the three domains of life. Nat. Rev. Microbiol. 9:85–98.

Williams TA, Cox CJ, Foster PG, Szöllősi GJ, Embley TM. 2020. Phylogenomics provides robust support for a two-domains tree of life. Nat Ecol Evol 4:138–147.

Williams TA, Foster PG, Nye TM, Cox CJ, Embley TM. 2012. A congruent phylogenomic signal places eukaryotes within the Archaea. Proc. Biol. Sci. [Internet] 279. Available from: https://pubmed.ncbi.nlm.nih.gov/23097517/

Williams TA, Szöllősi GJ, Spang A, Foster PG, Heaps SE, Boussau B, Ettema TJG, Embley TM. 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. Proc. Natl. Acad. Sci. U. S. A. 114:E4602–E4611.

Woese CR. 1987. Bacterial evolution. Microbiological Reviews [Internet] 51:221–271. Available from: http://dx.doi.org/10.1128/mmbr.51.2.221-271.1987

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U. S. A. 74:5088–5090.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U. S. A. 87:4576–4579.

Wolf YI, Aravind L, Grishin NV, Koonin EV. 1999. Evolution of aminoacyl-tRNA synthetases--analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res. 9:689–710.

Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 9:R151.

Young AD, Gillung JP. 2020. Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. Systematic Entomology [Internet] 45:225–247. Available from: http://dx.doi.org/10.1111/syen.12406

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, et al. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353–358.

Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, DeBoy RT, Nelson KE, Nesbø CL, Doolittle WF, Gogarten JP, et al. 2009. On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proc. Natl. Acad. Sci. U. S. A. 106:5865–5870.

Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.

Chapter 4. Research Project

Genome Size Distributions in Bacteria and Archaea are Strongly Linked to Evolutionary History at Broad Phylogenetic Scales

Previously published: Martinez-Gutierrez CA and Aylward FO. 2022. Genome Size Distributions in Bacteria and Archaea are Strongly Linked to Evolutionary History at Broad Phylogenetic Scales. PLoS Genetics, 18(5), e1010220.

Co-authors contributed in the following ways: Conceived and designed this work: CAMG and FOA. Wrote the paper: CAMG and FOA.

4.1 Abstract

The evolutionary forces that determine genome size in bacteria and archaea have been the subject of intense debate over the last few decades. Although the preferential loss of genes observed in prokaryotes is explained through the deletional bias, factors promoting and preventing the fixation of such gene losses often remain unclear. Importantly, statistical analyses on this topic typically do not consider the potential bias introduced by the shared ancestry of many lineages, which is critical when using species as data points because of the potential dependence on residuals. In this study, we investigated the genome size distributions across a broad diversity of bacteria and archaea to evaluate if this trait is phylogenetically conserved at broad phylogenetic scales. After model fit, Pagel's lambda indicated a strong phylogenetic signal in genome size data, suggesting that the diversification of this trait is influenced by shared evolutionary histories. We used a phylogenetic generalized least-squares (PGLS) analysis to test whether phylogeny influences the predictability of genome size from dN/dS ratios and 16S copy number, two variables postulated to play a role in shaping genome size. These results confirm that failure to account for evolutionary history can lead to biased interpretations of genome size predictors. Overall, our results indicate that although bacteria and archaea can rapidly gain and lose genetic material through gene transfer

and deletions, respectively, phylogenetic signal for genome size distributions can still be recovered at broad phylogenetic scales that should be considered when inferring the drivers of genome size evolution.

4.2 Author Summary

The evolutionary forces driving genome size in bacteria and archaea have been subject to debate during the last decades. Typically, independent comparative analyses have suggested that unique variables, such as the strength of selection, environmental complexity, and mutation rate, are the main drivers of this trait, without considering for potential biases derived from shared ancestry. Here, we applied a phylogeny-based statistical approach to assess how tightly genome size in bacteria and archaea is linked to evolutionary history. Moreover, we also evaluated the predictability of genome size from the strength of purifying selection and ecological strategy on a broad diversity of bacteria and archaea genomes under a phylogenetic comparative framework. Our approach indicates that despite the ability of bacteria and archaea to rapidly exchange genes, a strong phylogenetic signal to genome size distributions can be recovered at broad phylogenetic scales.

4.3 Introduction

Bacterial and archaeal genomes are densely packed with genes and contain relatively little non-coding DNA, therefore an increase in genome size is directly translated into more genes (Mira et al. 2001; Lynch 2006; Koonin 2009). In contrast, multicellular eukaryotes generally show genome expansion due to the proliferation of noncoding-DNA because of high genetic drift (Lynch 2006). The absence of non-functional elements in prokaryotes is explained through the bias towards more deletions than insertions; newly acquired or existing genes are removed if selection on those genes is insufficient for their maintenance in the population (Wolf et al. 1999; Lawrence et al. 2001;

Bobay and Ochman 2017). Although narrowly constrained when compared with eukaryotes, prokaryotic genome sizes still vary by over one order of magnitude. Assuming an intrinsic deletion bias across all prokaryotes, it remains unclear what evolutionary forces determine which genes are maintained and which are lost, and what determines the variability of genome sizes across the broad diversity of bacteria and archaea.

Multiple individual factors have been hypothesized to be primary drivers of genome size in bacteria and archaea. Early studies suggested that effective population size (*Ne*) may be the primary force that determines genome size and fluidity in prokaryotes (Sela et al. 2016; Andreani et al. 2017). For example, genome reduction has been observed in host-dependent bacteria that have small *Ne* and correspondingly high levels of genetic drift due to population contractions. Under such evolutionary constraints, slightly deleterious deletions accumulate and cause overall genome reduction (Moran and Mira 2001; van Ham et al. 2003; Woolfit and Bromham 2003; Batut et al. 2014; Chong et al. 2019). Paradoxically, later studies focusing on abundant free-living planktonic lineages in the ocean suggested that genome reduction can also be observed in bacteria with larger *Ne* that experience strong purifying selection (Grote et al. 2012; Giovannoni et al. 2014; Kashtan et al. 2014; Biller et al. 2015). In this case selection favors genomic economization, such as the removal of paralogs and intergenic sequences. Factors other than *Ne* and the strength of purifying selection have also been postulated to play a role in determining prokaryotic genome size. Recently, one study suggested that environmental stress leads to genome streamlining in soil bacteria (Simonsen 2021), and other genomics studies have suggested that habitat complexity and ecological strategy (Rodríguez-Gijón et al. 2021), as well as the capability to use oxygen (Nielsen et al. 2021) may also play major roles in determining genome size in bacteria and archaea (Rodríguez-Gijón et al. 2021). Mutation rate has also been proposed to be a major factor

determining genome size (Marais et al. 2008; Bourguignon et al. 2020). In particular, it was suggested that a high mutation rate would be the primary cause of genome reduction in both streamlined and host-dependent bacteria due to the erosion of genes, loss of function, and subsequent deletion (Marais et al. 2008; Bourguignon et al. 2020; Marais et al. 2020). However, other studies analyzing the mutation rate of the abundant picocyanobacteria *Prochlorococcus* show estimates similar to *Escherichia coli*, casting doubt on the view that high mutation rates drive genome reduction in all cases (Osburne et al. 2011; Chen et al. 2021). Given the large number of forces that have been proposed to be primary determinants of genome size, it remains largely unknown whether genome size in prokaryotes is driven by unique variables, their interaction, or variables that have specific influence depending on the lineage.

Importantly, most statistical analyses exploring the association between genome size and other traits have typically not used phylogenetic comparative methods that are necessary when using species as data points. Shared evolutionary history may obscure the relationship between traits because the phylogenetic dependence between lineages leads to the violation of the statistical assumption of independence in residuals. Thus, conventional statistical methods can lead to overestimation of the strength of the association between traits (Felsenstein 1985; Garland et al. 1999). In this study, we estimated the phylogenetic signal of genome size across a broad diversity of bacterial and archaeal genomes available on the Genome Taxonomy Database (GTDB) (Chaumeil et al. 2019). Although genome size has been shown to change rapidly in prokaryotes due to HGT and gene loss, we sought to test if this trait still bore a phylogenetic signal across broad phylogenetic scales. Moreover, because previous studies have suggested that effective population size or ecological niche are potential drivers of genome size (Koonin 2009; Sela et al. 2016), we evaluated whether correlations with these factors would change if evolutionary history

was taken into account. Our work provides important insights into the complex mechanisms that shape genome size in bacteria and archaea, and the importance of considering shared evolutionary relationships when studying its evolution to avoid bias in the association between traits.

4.4 Results and discussion

4.4.1 Genome size distribution across major phyla of bacteria and archaea

To explore the distribution of genome size across the Tree of Life of bacteria and archaea and to measure phylogenetic signal across broad phylogenetic scales, we built a phylogenetic tree using one representative genome of 836 genera belonging to 33 phyla available on the GTDB (Supplemental File 2). For the reconstruction of this phylogeny, we used a set of ribosomal proteins and RNA polymerase subunits that we have previously benchmarked (Martinez-Gutierrez and Aylward 2021). The size of genomes in our analysis and across the phylogeny varied by over two orders of magnitude (0.6-14.3 Mbp, Fig. 4.1A and 4.2). The minimum and maximum corresponded to two bacterial lineages with contrasting lifestyles: the endosymbiont *Buchnera aphidicola* of the phylum Proteobacteria and the free-living Actinobacteria *Nonomuraea sp.* (Fig. 4.1A). The greatest within-phylum variation of genome size was observed for the phyla Actinobacteria and Cyanobacteria, whereas the phylum with shortest genomes belonged to symbiotic bacteria of the phylum Patescibacteria (Fig. 4.1A). We also evaluated the difference in genome size found within the genera used in our study, which we report here as the variance (Fig. 1B) and the difference between the largest and smallest genomes within each genus (Supplemental File 3). Most of the genera used in our analysis (571 out of 863) showed a difference smaller than 1 Mbp, but some genera exhibited a wide range of genome sizes; for example, the genera *Streptomyces* and *Nonomuraea* showed a difference of 6.29 and 6.06 Mpb between the smaller and the larger genomes, respectively (Fig. 4.1B, Supplemental File 3). The large difference found between the

largest and smallest genome of some of the genera in our dataset is consistent with previous observations of considerable differences in the genome size and genome content of many closely related taxa (Casjens 1998; Lan and Reeves 2000; Dobrindt and Hacker 2001; Lawrence and Hendrickson 2005).

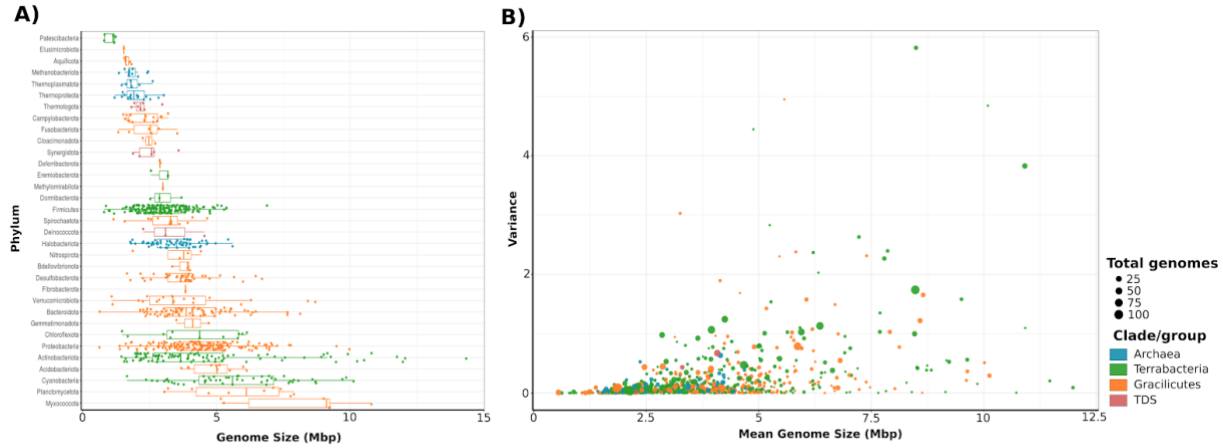

Figure 4.1 Distribution of genome size in bacteria and archaea taxonomic groups. A) Distribution at the phylum level. First, third quantile, and median are shown for each phylum distribution. B) Relationship between mean genome size and genome size variance for each genus cluster. Abbreviations: TDS = Thermotogota, Deinococcota, and Synergistota. Raw data for genome size can be found in Supplemental File 3.

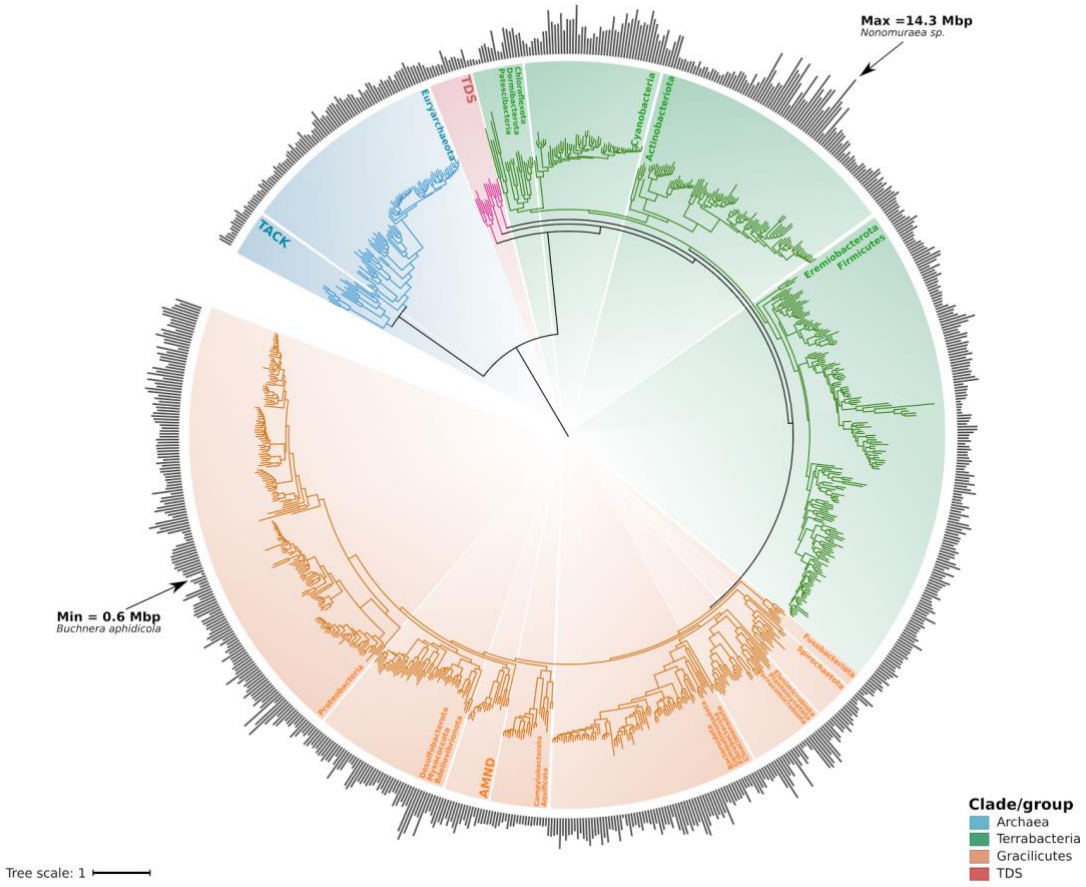

Figure 4.2 Genome size distribution across the Tree of Life of bacteria and archaea using one representative genome for each genus. Phylogenetic tree was built using a concatenated alignment of ribosomal and RNA polymerase sequences through a maximum likelihood approach and the substitution model LG+R10. Abbreviations: TACK = Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota; TDS = Thermotogota, Deinococcota, and Synergistota; AMND = Acidobacteriota, Methylomirabilota, Nitrospirota, Deferribacterota. Raw data for genome size can be found in Supplemental File 3.

4.4.2 Genome size in bacteria and archaea is strongly dependent on phylogenetic history at broad evolutionary scales

Although it is well known that genome size can vary markedly between closely-related bacteria and archaea (Casjens 1998; Lan and Reeves 2000; Dobrindt and Hacker 2001; Lawrence and Hendrickson 2005), it is still possible that overall genome size distributions are linked to evolutionary history at broader phylogenetic scales, which we define here as anything broader than the genus level according the GTDB classification (Fig. 4.2). Due to the shared evolutionary history of some lineages, traits of related groups often resemble each other more than when compared with randomly-selected species in the same phylogenetic tree (phylogenetic signal) (Blomberg et al. 2003; Revell et al. 2008; Münkemüller et al. 2012). We therefore sought to investigate the phylogenetic signal of genome size distributions in our genome dataset (Fig. 4.2). Phylogenetic methods are needed to analyze these associations because any study involving statistical analyses and species as data points potentially violates the assumption of independence of residuals (Felsenstein 1985; Freckleton et al. 2002)**.**

When studying phylogenetic signal, it is recommended to measure it at two different levels: 1) in traits' raw data and 2) in the residuals resulting from statistical models (e.g., regressions) (Revell 2010). As a first approximation, we assessed whether genome size distribution data show phylogenetic signal by estimating Blomberg's K (Blomberg et al. 2003) on the genome size of the GTDB genomes dataset (Fig. 4.2). Values of Blomberg's K between 0 and 1 indicate that the sizes of closely related genomes resemble each other but less than expected under the Brownian Motion model (BM) of trait evolution, where trait variation is proportional to phylogenetic distance (Felsenstein 1985). Conversely, a K of 1 is evidence of genome size variation according to the Brownian Motion expectation (Blomberg et al. 2003). We observed phylogenetic signal in genome

size data that is strong but different to what would be expected under the Brownian Motion model (BM) (K=0.51, P=0.001), suggesting that although genome size shows phylogenetic signal, variation is not fully explained through phylogenetic distance in our tree (Felsenstein 1973).

In addition, we tested the fit of different models of trait evolution for genome size, including Brownian Motion (Felsenstein 1973), Ornstein-Uhlenbeck (Butler and King 2004), Early-Burst (Harmon et al. 2010), a diffusion model, Pagel's model (Pagel 1999), a drift model, and a white-noise model (non-phylogenetic signal) (Table 4.1). According to a likelihood ratio test performed ($P<0.001$ when compared with the next-best likelihood), Pagel's model showed the best fit (Table 4.1) with a lambda value of 0.90 ($P<0.001$). The Pagel's lambda (λ) represents how strongly phylogenetic relationships predict the observed pattern of variation of a trait at the tips of a phylogeny, and varies from 0 (no phylogenetic signal) to 1 (phylogenetic signal under BM) (Pagel 1999). Although we obtained different estimates for Blomberg's K and Pagel's λ, we considered that λ is more reliable because this metric is more robust than Blomberg's K in situations of erroneous branch lengths (Molina-Venegas and Rodríguez 2017). Our λ estimate supports our conclusion that genome size data in bacteria and archaea show phylogenetic signal. These findings indicate that genome size in bacteria and archaea does not evolve independently of broad evolutionary relationships. To confirm that our phylogenetic signal estimates are not unduly influenced by the phylogenetic scale that we examined, we repeated our analyses using a larger set of genomes consisting of multiple representatives for each genus (Supplemental File 2) and we observed a similar trend (Supplemental File 6), suggesting that the phylogenetic signal trend observed in genome size data is not the result of a biased taxonomic representation. Moreover, for our genus-level tree we estimated kappa (k) and delta (δ) on genome size data, two parameters that describe the mode of evolution of a trait (punctuated vs gradual) and the rate change across the

phylogeny (acceleration vs deceleration), respectively (Hernández et al. 2013). Our estimates ($k=0.24$ and $\delta=3$) are consistent with a gradual and late diversification of genome size in bacteria and archaea, which might indicate lineage-specific adaptations (Pagel 1999; Hernández et al. 2013).

Table 4.1 Summary of model fitting for genome size data. We highlighted the model that showed the highest likelihood and the lowest AIC.

Model	Loglik	Parameters	AIC
Brownian motion	-1463.3	Sigma = 12.3 Root state = 2.7	2930.7
Ornstein-Uhlenbeck	-1420.7	alpha = 2.7 Sigma = 17.5 Root state = 3.1	2847.4
Early-Burst	-1463.3	a = 0 Sigma = 12.3 Root state = 2.7	2932.7
Pagel's model*	-1415.6	Lambda = 0.9 Sigma = 6.2 Root state = 2.7	2837.2
Trend diffusion	-1447.7	Slope = 100 Sigma = 0.1 Root state = 2.9	2901.3
Drift	-1463.3	Drift = -99.9 Sigma = 12.2 Root state = 102.7	2932.7
White-noise	-1695.6	Sigma = 3.4 Root state = 3.9	3395.2

*Significantly higher likelihood when compared with the rest of the models tested according to the chisq test ($P<0.001$)

Because phylogenetic signal estimates can be biased due to sample size (Kamilar and Cooper 2013), we measured phylogenetic signal within each phylum (Fig. 4.1A). Our results indicate that most of the phyla with a small sample size (<25 genomes) showed remarkably large or small K and λ values (Supplemental File 5), consistent with previous findings that small sample sizes lead to biased estimates (Freckleton et al. 2002; Kamilar and Cooper 2013). We did not observe a linear increase in λ values with the number of genomes tested, however, suggesting that the large lambda estimate found in our overall genome size data is not associated with our large sample size (Supplemental File 5).

4.4.3 Non-phylogenetic regression overestimates the effect of dN/dS on genome size

We next explored whether the residuals resulting from the statistical association between genome size and other traits show phylogenetic signal. Previous studies have suggested that high levels of genetic drift are related with a decrease in genome size in bacteria (Kuo et al. 2009; Sela et al. 2016). However, such studies were based on a limited set of genomes available at the time and did not include a broad repertoire of streamlined genomes, which are notable for their small genomes and large effective population sizes (Batut et al. 2014; Martinez-Gutierrez and Aylward 2019). We first investigated whether this trend is maintained when including a broader diversity of taxa by calculating pairwise dN/dS values for each genus in the GTDB genomes dataset (Sela et al. 2016). Our non-phylogenetic generalized least squares (GLS) showed a positive and significant but low correlation between genome size and dN/dS ($P<0.001$, Pseudo-$R^2=0.04$, Table 4.2, Fig. 4.3A). This result contrasts with earlier studies reporting a strong relationship between genome content and dN/dS (Kuo et al. 2009; Sela et al. 2016); we attribute this large discrepancy to the broad taxonomic representation in our dataset, which includes small genomes under both, strong purifying selection and genetic drift (Batut et al. 2014). Interestingly, when considering phylogeny

through the better-fitting Pagel's model, our phylogenetic generalized least squares model (PGLS) showed poorer predictability and a non-significant relationship between both variables (P=0.5, Pseudo-R^2= 0.0006, Table 4.2, Fig. 4.3A). Similar results were found in a study that analyzed the phylogenetic signal associated with genome size across prokaryotes and eukaryotes (Whitney and Garland 2010; Whitney et al. 2011). In this previous study, authors showed that the phylogenetic signal found in genome size data caused a biased association between Ne.μ (approximated through nucleotide diversity) and other genetic traits, including genome size (Whitney and Garland 2010; Whitney et al. 2011). Our PGLS analysis indicates that not only does genome size data show phylogenetic signal, but that the residuals of our regression models also bear this signal (Table 4.2), confirming the need of assessment of phylogenetic-based methods when studying the evolution of genome size (Revell and Collar 2009; Kamilar and Cooper 2013). We also calculated the lambda parameter on our dN/dS data, and the value found (λ=0.68; 95% CI= 0.56-0.77) indicates a relatively high phylogenetic signal for this variable, suggesting that phylogenetically related microorganisms tend to experience similar levels of selection. Altogether, these results suggest that correlations between dN/dS and genome size found previously are largely driven by poor sampling and artifacts that arise by not specifically accounting for the recent shared evolutionary history of many lineages (Felsenstein 1985).

Table 4.2 Statistics of the regression models relating genome size and dN/dS and 16S rRNA as predictor variables using Generalized Least Square and Phylogenetic Least Square analyses. We highlighted the models that were statistically significant ($\alpha = 0.05$).

Model	Predictor variable	Kappa (95% CI)	Lambda (95% CI)	Delta (95% CI)	Slope	Intercept	P-val	AIC	R^2
Generalized Least Square									
Genome Size ~ Median dN/dS	dN/dS	-	-	-	13.57	2.97	<0.001	3366.2	0.04*
Genome Size ~ 16S rRNA copies	16S rRNA copies	-	-	-	0.12	3.65	0.002	3387.7	0.01*
Genome Size ~ Median dN/dS + 16S rRNA copies	dN/dS + 16S rRNA copies	-	-	-	14.11/ 0.13	2.7	<0.001	3355.9	0.05*
Phylogenetic Generalized Least Square									
Genome Size ~ Median dN/dS	dN/dS	0.48 (0.39-0.58)	0.98[a,b] (0.96-0.99)	2.44 (2.01-2.85)	1.26	2.46	0.5	2748.8	0.0006**
Genome Size ~ 16S rRNA copies	16S rRNA copies	0.49 (0.34-0.59)	0.98[a] (0.96-0.99)	2.49 (2.06-2.9)	0.08	2.42	0.003	2740.268	0.01**
Genome Size ~ Median	dN/dS + 16S rRNA	0.49 (0.40-0.59)	0.98[a,b] (0.96-0.99)	2.51 (2.08-2.93)	1.29/0.08	2.35	0.009	2741.79	0.01**

dN/dS + 16S rRNA copies* **	copies								

*Nagelkerke's R^2
** Multiple R square; percentage of variance explained between a null model and the actual model given that precise model of trait change
***Anova did not show significant differences between models Genome Size ~ 16S rRNA copies and Genome Size ~ Median dN/dS + 16S rRNA copies (P=0.48)
[a]Significantly different than 0 (no phylogenetic signal)
[b]Significantly different than 1 (Brownian Motion expectation)

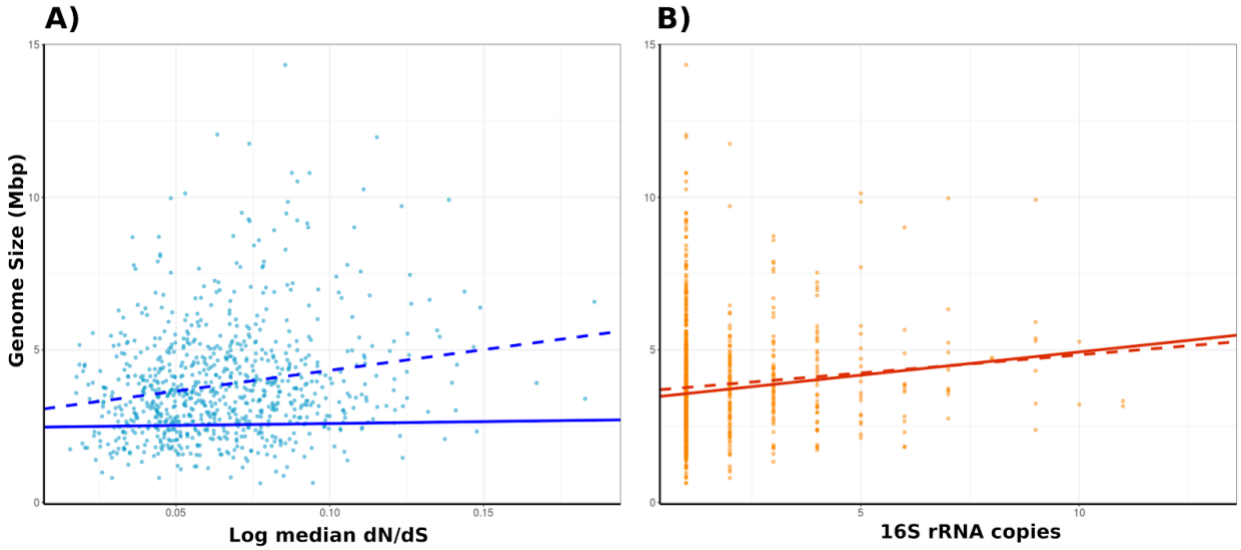

Figure 4.3 Relationship between genome size and genomic traits for bacteria and archaea using one representative genome for each genus. A) Regression line of the relationship between genome size and dN/dS ratio before (dashed line) and after (solid line) taking phylogenetic relationships into account through the Pagel's model. B) Regression line of the relationship between genome size and 16S rRNA copies before (dashed line) and after (solid line) taking phylogenetic relationships into account through the Brownian Motion model. Parameters of the regression equation for both relationships can be found in Table 4.2. Raw data can be found in Supplemental File 3.

Although our results indicate that dN/dS is a poor predictor of genome size in bacteria and archaea (Fig. 3A), it is worth mentioning that dN/dS only reflects recent evolutionary constraints due to saturation of substitutions at synonymous sites (Rocha et al. 2006; Luo et al. 2017). Therefore, we do not discount that genome reduction may be driven in part by processes such as

population bottlenecks and periods of relaxed selection that happened in the past but are not reflected in dN/dS estimations. This scenario has been suggested for *Prochlorococcus,* in which the genome simplification observed in this clade could be the result of periods of relaxed selection experienced in the past (Luo et al. 2017).

4.4.4 Ecological strategy plays a role on genome size in bacteria and archaea

In addition to testing the effect of the strength of selection on genome size, we also assessed the predictability of genome size from 16S rRNA copies as an approximation to ecological strategy using both, GLS and PGLS. Previous studies have shown that copies of the *rrn* operon can be a predictor of the number of ribosomes that a cell can produce simultaneously, and that this reflects the ecological strategy in microorganisms (Klappenbach et al. 2000; Niederdorfer et al. 2017). A large number of *rrn* copies is associated with the ability to adapt quickly to fluctuating environmental conditions (i.e., "boom and bust" strategies) (Condon et al. 1995), while multiple *rrn* copies would confer a metabolic burden to slow-growing microorganisms living in stable or low-nutrients environments because of ribosome overproduction (Klappenbach et al. 2000). Similarly, to what we observed for dN/dS, we found a weak, positive, and significant relationship between genome size and 16S rRNA copies when using GLS ($P<0.001$, Pseudo-$R^2= 0.01$, Table 4.2, Fig. 4.3B). Interestingly, we still observed a significant relationship when accounting for the phylogenetic signal in the residuals through a PGLS analysis ($P=0.003$, $R^2=0.01$, Table 4.2, Fig. 4.3B). However, the Pagel's lambda of this model was not significantly different from 1 (Table 4.2), indicating that the residuals of this model show a distribution closer to the BM expectation. After fitting under the BM, we still observed a positive and significant relationship between genome size and 16S rRNA copies ($P<0.001$, Pseudo-$R^2=0.02$). Although the predictability of 16S rRNA is weak under both BM and Pagel's model, our findings suggest that environment

complexity plays a role on genome size independently of phylogenetic relationships. This is consistent with the observation that larger genomes tend to inhabit environments with temporal variability and diversity of resources (Guieysse and Wuertz 2012; Chuckran et al. 2022). In addition to fitting our model using dN/dS and 16S rRNA copies individually as predictors, we fitted an additive model with both variables (Table 4.2). An ANOVA test showed that a model including both variables does not significantly improve the fit when compared with the model based on 16S rRNA copies as a unique predictor variable ($P = 0.48$).

4.4.5 A hypothesis for the evolutionary processes that shape genome size in bacteria and archaea

According to our phylogenetic comparative framework (Table 4.1-4.2, Fig. 4.3), lineages with recent shared evolutionary history tend to maintain similar sizes since the divergence from their common ancestor. Nevertheless, the pattern of variation in genome size data ($\lambda=0.90$) differs from what would be expected under the Brownian Motion model. This finding suggests that besides evolutionary relationships, there are other variables defining genome size in prokaryotes. Our results are consistent with the view that genome size in prokaryotes is the result of a complex interplay of multiple variables, including evolutionary history, past events such as population bottlenecks, and environmental complexity (substrates available, variability in environmental factors, biotic pressure, etc.), but it can remain relatively stable at broad phylogenetic scales. Although several factors have been proposed to be singular drivers of genome size in prokaryotes, such as effective population size (Lynch and Conery 2003), ecological strategy (Konstantinidis and Tiedje 2004; Rodríguez-Gijón et al. 2021), and mutation rate (Marais et al. 2008; Bourguignon et al. 2020; Marais et al. 2020), our findings strongly suggest that genome size is a complex trait

determined by the interaction of multiple variables, and that the relative importance of these factors may vary across lineages.

Phylogenetic signal estimates can vary across phylogenetic scales (Graham et al.), and it is therefore possible that the strong phylogenetic signal found in our analyses is weaker or not observed at finer scales. This is particularly expected in clades that experience rapid genome turnover due to the acquisition and loss of genes through horizontal gene transfer events (HGT) and deletions, respectively. For example, genome contraction events are expected in endosymbionts like *Buchnera* and *Blattabacterium,* which are thought to derive from a large-genome ancestor (Moran and Mira 2001), and are frequently undergoing bottlenecks and periods of diversity loss (Moran and Mira 2001; Tamas et al. 2002; van Ham et al. 2003). Such exacerbated loss of genes and diversity is enhanced by the nearly absent homologous recombination found in vertically transmitted endosymbionts (McCutcheon and Moran 2011). These observations are consistent with the relatively high dN/dS value and small genome size that we observed for *Buchnera* and *Blattabacterium* (Fig. 4.4). In contrast, some abundant marine clades inhabiting the open ocean such as *Prochloroccocus* and *Pelagibacter* have undergone long periods of adaptation and specialization to their stable environments (Giovannoni 2017; López-Pérez et al. 2020). The open ocean is characterized by chronically-oligotrophic nutrient conditions that are stable throughout the year (Partensky and Garczarek 2010), and genes that are under relaxed selection are therefore pseudogenized and lost (Batut et al. 2014). The latter is supported by the unusual growth requirements and low number of transcriptional regulators found in *Pelagibacter*, which is expected to limit its response to changing environmental conditions (Carini et al. 2013; Cottrell and Kirchman 2016). Consistent with these observations, we observed low dN/dS values, small genome size, and fewer 16S rRNA for these streamlined bacteria (Fig. 4.4). The small genomes

observed in both endosymbionts and free-living planktonic lineages are therefore likely the result of distinct evolutionary processes, as previously proposed (Giovannoni et al. 2014).

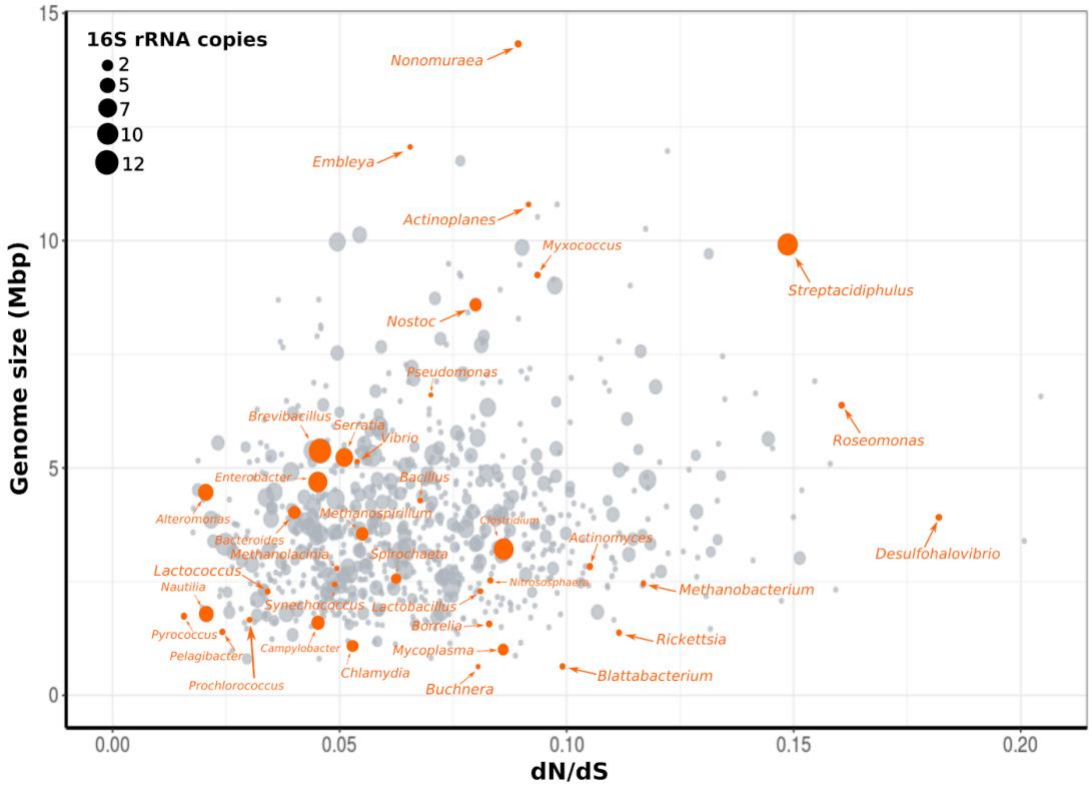

Figure 4.4 Relationship between genome size and dN/dS in bacteria and archaea. dN/dS values represent the median estimate for each genus cluster. Dots represent a representative genome for each genus and size is equivalent to the number of 16S rRNA gene copies. Raw data can be found in Supplemental File 3.

In contrast to the genome simplification observed in host-dependent and streamlined prokaryotes, genome expansion is expected in free-living lineages that inhabit complex environments like soils or sediments, where microenvironments with strikingly different abiotic conditions can be found millimeters apart (Fierer 2017). Although temporal diversity declines and

sweeps for specific gene variants are likely to occur in soil prokaryotes due to rapidly changing environmental conditions (Takeuchi et al. 2015; Fierer 2017), larger genomes may be positively selected in these environmental realms due to variable abiotic and biotic constraints. Indeed, a study exploring the genes enriched in larger genomes of soil prokaryotes found a larger proportion of genes involved in regulation and secondary metabolism, and were depleted in genes related with translation, replication, cell division, and nucleotides metabolism when compared with smaller genomes (Konstantinidis and Tiedje 2004). These environmental and genomic findings are consistent with the large genome sizes, high dN/dS, and multiple 16S rRNA copies estimated in our study for soil microorganisms of the genera *Streptacidiphilus, Actinomyces, Conexibacter, Actinoplanes,* and *Myxococcus* (Fig. 4.4), the latter showing complex fruiting body development (Goldman et al. 2007). It is interesting to note that the largest genomes analyzed in our study (>6 Mpb) tend to experience intermediate levels of purifying selection (dN/dS), suggesting that either extremely high or low purifying selection are not conducive to genomic expansion events.

4.5 Outlook

Despite the increase of genomes available on publicly available databases, the evolutionary processes and factors driving genome size and content in bacteria and archaea are continuously debated. Several studies have proposed ecological strategies, the strength of purifying selection, and mutation rate as prominent forces that individually determine prokaryotic genome size. Our statistical approach shows that at broad phylogenetic scales evolutionary history plays a large role in structuring genome size distributions across bacteria and archaea. Genome size is therefore not independent of phylogeny, and a failure to account for this can lead to misleading associations between traits. In some ways our finding of a strong phylogenetic signal to genome size in prokaryotes across broad evolutionary timescales is paradoxical given the well-known variability

of prokaryotic genome size within species and between closely-related lineages (Casjens 1998; Lan and Reeves 2000; Dobrindt and Hacker 2001; Lawrence and Hendrickson 2005). These two realities need not conflict, however; for example, it is possible that genome size fluctuates rapidly at short evolutionary timescales but remains relatively constant over long periods due to an overall balancing of gene gain and loss over long periods of time. The significant but poor relationship between genome size and 16S rRNA copies suggest that besides phylogenetic history, ecological strategy plays a role in shaping genome size in bacteria and archaea, although this single trait is insufficient to completely represent ecological strategies. Future studies will be necessary to evaluate the evolution of genome size on a lineage-by-lineage basis, however the strong phylogenetic signal observed in genome size data indicates that analyses involving this trait cannot consider species as phylogenetically independent, therefore phylogenetic relatedness should be assessed and considered to avoid simplified models and biased associations between traits.

4.6 Material and methods

4.6.1 Genomes compilation and phylogenetic reconstruction

To estimate the phylogenetic signal in genome size data at a broad phylogenetic scale, we compiled a genomes dataset that included a broad diversity of bacteria and archaea. All the representative genomes available on the Genome Taxonomy Database (GTDB) (Release 05-RS95; 17th July 2020) (Chaumeil et al. 2019) were filtered based on completeness (>=95%) and contamination (<=5%) and then classified at the class levels. Genomes belonging to the phylum Patescibacteria (also known as Candidate Phyla Radiation or CPR) were filtered using the parameters completeness>=80% and contamination<=5%. After filtering and classification, classes with more than 500 genomes were randomly down sample to 500 genomes. The resulting genomes were

clustered based on their taxonomic identity at the genus level and genera with fewer than two genomes were discarded from further analyses. Our final dataset consisted of 4380 genomes classified in 836 genera (Supplemental File 1). For phylogenetic reconstruction, we randomly selected one genome from each genus (referred hereafter as GTDB genomes dataset) and used the MarkerFinder pipeline reported previously (Martinez-Gutierrez and Aylward 2021). This pipeline consisted in the identification of 27 ribosomal proteins and three RNA polymerase genes (Ribosomal-RNAP set) (Sunagawa et al. 2013) using HMMER3. The resulting individual sequences were aligned with ClustalOmega and concatenated. We trimmed the concatenated alignment with trimAl (Capella-Gutiérrez et al. 2009) using the option -gt 0.1. The Ribosomal-RNAP alignment was then used to build the phylogenetic tree with IQ-TREE 1.6.12 (Nguyen et al. 2015) with the substitutions model LG+R10 and the options -wbt, -bb 1000, and --runs 10 (Le and Gascuel 2008; Soubrier et al. 2012; Minh et al. 2013). The resulting phylogeny was manually inspected on iTOL (Letunic and Bork 2019) (Fig. 4.2). Raw phylogenetic tree is included in Supplemental File 7.

4.6.2 dN/dS estimation and *rrn* genes identification

To investigate whether the phylogenetic signal in genome size data leads to biased associations with other variables like the strength of selection and ecological strategy, we estimated the ratio of synonymous and nonsynonymous substitutions (dN/dS) within each genus cluster of our GTDB genomes dataset using two sets of conserved marker genes, checkm_bact and checkm_arch for bacteria and archaea, respectively (Parks et al. 2015). Genomes used to calculate the dN/dS for each genus cluster are reported in Supplemental File 2. The open reading frames (ORFs) retrieved from the GTDB were compared to the HMMs of the checkm_bact (120 marker genes) and checkm_arch marker (122 marker genes) sets using the hmmsearch tool available in HMMER v.

3.2.1 with the reported model-specific cutoffs (Eddy 2011). We aligned the amino acid sequences for each marker gene and each genus cluster individually using ClustalOmega (Sievers and Higgins 2018), and then converted amino acid alignments into codon alignments using PAL2NAL with the parameter --nogap (Suyama et al. 2006). We used the resulting codon alignments to estimate the pairwise ratio of synonymous and nonsynonymous substitutions for each pair of genomes using the maximum likelihood approximation (codeML) available on PAML 4.9h (runmode=-2) (Yang 2007). In order to avoid bias associated with divergence, dN/dS estimates with dS>1.5 were removed due to potential saturation. We also discarded pairwise comparisons with dS<0.1 because these might represent dN/dS values calculated from genomes of the same population. Moreover, dN/dS values >10 were considered artifactual (Martinez-Gutierrez and Aylward 2019). Genomes with fewer than 25 dN/dS estimates remaining after filtering were discarded. We used the resulting median dN/dS of our representative genomes for further analysis. In order to examine the effect of genes' selection on final dN/dS estimations, we randomly selected 40 genera and identified their core genes using CoreCruncher (Harris et al. 2021) using usearch (Edgar 2010) and the default parameters except for -score 80. We estimated the pairwise dN/dS for each core gene using the approach described previously and estimated the median dN/dS for our genus-representative genomes. A linear regression between the dN/dS values resulting from core genes and the 120 marker genes set for the 40 genera showed similar results (Supplemental File 4), therefore we used the latter for further statistical analyses. In addition, we predicted ribosomal RNA genes in our representative genomes as an approximation to ecological strategy using Barrnap with the default parameters (barrnap 0.9: rapid ribosomal RNA prediction; https://github.com/tseemann/barrnap). Genome size, 16S rRNA copies, and dN/dS values for the GTDB representative genomes dataset are reported in Supplemental File 3.

4.6.3 Statistical analyses

Due to the tendency of related species to resemble each other because of their shared phylogenetic ancestry, we assessed the suitability of a phylogeny-based method for our regression analyses by first estimating Blomberg's K on genome size data (Blomberg et al. 2003) using the phylosignal function on R (Kembel et al. 2010). This parameter represents the phylogenetic signal in a continuous trait, and goes from 0 (no phylogenetic signal) to ∞ (phylogenetic signal) with the null hypothesis (K=1) meaning that the trait analyzed evolves under Brownian Motion (Felsenstein 1973; O'Meara et al. 2006). In addition, we also tested the fit of different trait evolution models, including Brownian Motion (Felsenstein 1973), Ornstein-Uhlenbeck (Butler and King 2004), Early-Burst (Harmon et al. 2010), a diffusion model, Pagel's model (Pagel 1999), a drift model, and a white-noise model (non-phylogenetic). We tested the predictability of genome size from dN/dS and 16S rRNA copies as predictor variables using the "glm" function available on R. Since we detected phylogenetic signal in genome size data, we additionally accounted for potential phylogenetic nonindependence in the residuals using the PGLS method with the function pgls on the R package Caper (Anon) and the Pagel's model (Pagel 1999), as well as the function gls available on the package ape (Paradis and Schliep 2019). We additionally tested the effect of sample size on the calculation of Blomberg's K and Pagel's lambda by estimating these parameters within each phylum (Fig. 4.1A and Supplemental File 5). The trait data and the phylogeny used in these analyses can be found in Supplemental File 3 and Supplemental File 7, respectively. In addition to testing phylogenetic signal in a broad-scale phylogeny (Supplemental File 3, Fig. 4.2), we built a phylogenetic tree with multiple representative genomes for each genus and the IQ-TREE workflow used for the rest of our analyses. The genome size data and the phylogenetic tree used for this analysis can be found in Supplemental File 2 and Supplemental File 8, respectively.

4.7 Data availability

The supplemental files of this chapter are available of the Figshare collection: https://doi.org/10.6084/m9.figshare.c.6240945.v1

4.8 Acknowledgments

We acknowledge the use of the Virginia Tech Advanced Research Computing Center for bioinformatic analyses performed in this study. This investigation was supported by grants from the Institute for Critical Technology and Applied Science and the National Science Foundation (IIBR-1918271), and a Simons Early Career Award in Marine Microbial Ecology and Evolution to F.O.A. We kindly thank members of the Aylward Lab for their insightful comments on an earlier version of this manuscript and Prof. Josef Uyeda for advice on phylogeny-based statistical methods.

4.9 References

Andreani NA, Hesse E, Vos M. 2017. Prokaryote genome fluidity is dependent on effective population size. ISME J. 11:1719–1721.

Anon. Website. Available from: Orme D, Freckleton R, Thomas G, Petzold T, Fritz S, Isaac N, Pears W. 2012. Caper: Comparative Analyses of Phylogenetics and Evolution in R. Version 0.5. [WWW document] URL http://cran.r-project.org/web/packages/caper/caper. pdf [accessed 25 April 2013].

Batut B, Knibbe C, Marais G, Daubin V. 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. Nat. Rev. Microbiol. 12:841–850.

Biller SJ, Berube PM, Lindell D, Chisholm SW. 2015. Prochlorococcus: the structure and function of collective diversity. Nat. Rev. Microbiol. 13:13–27.

Blomberg SP, Garland T Jr, Ives AR. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745.

Bobay L-M, Ochman H. 2017. The Evolution of Bacterial Genome Architecture. Front. Genet. 8:72.

Bourguignon T, Kinjo Y, Villa-Martín P, Coleman NV, Tang Q, Arab DA, Wang Z, Tokuda G, Hongoh Y, Ohkuma M, et al. 2020. Increased Mutation Rate Is Linked to Genome Reduction in Prokaryotes. Curr. Biol. 30:3848–3855.e4.

Butler MA, King AA. 2004. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. Am. Nat. 164:683–695.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Carini P, Steindler L, Beszteri S, Giovannoni SJ. 2013. Nutrient requirements for growth of the extreme oligotroph "Candidatus Pelagibacter ubique" HTCC1062 on a defined medium. The ISME Journal [Internet] 7:592–602. Available from: http://dx.doi.org/10.1038/ismej.2012.122

Casjens S. 1998. The diverse and dynamic structure of bacterial genomes. Annu. Rev. Genet. 32:339–377.

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics [Internet]. Available from: http://dx.doi.org/10.1093/bioinformatics/btz848

Chen Z, Wang X, Song Y, Zeng Q, Zhang Y, Luo H. 2021. Prochlorococcus have low global mutation rate and small effective population size. Nature Ecology & Evolution [Internet]. Available from: http://dx.doi.org/10.1038/s41559-021-01591-0

Chong RA, Park H, Moran NA. 2019. Genome Evolution of the Obligate Endosymbiont Buchnera aphidicola. Mol. Biol. Evol. 36:1481–1489.

Chuckran PF, Hungate BA, Schwartz E, Dijkstra P. 2022. Variation in genomic traits of microbial communities among ecosystems. FEMS Microbes [Internet] 2. Available from: http://dx.doi.org/10.1093/femsmc/xtab020

Condon C, Liveris D, Squires C, Schwartz I, Squires CL. 1995. rRNA operon multiplicity in Escherichia coli and the physiological implications of rrn inactivation. J. Bacteriol. 177:4152–4156.

Cottrell MT, Kirchman DL. 2016. Transcriptional Control in Marine Copiotrophic and Oligotrophic Bacteria with Streamlined Genomes. Appl. Environ. Microbiol. 82:6010–6018.

Dobrindt U, Hacker J. 2001. Whole Genome Plasticity in Pathogenic Bacteria.

Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput. Biol. 7:e1002195.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics [Internet] 26:2460–2461. Available from: http://dx.doi.org/10.1093/bioinformatics/btq461

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Hum. Genet. 25:471–492.

Felsenstein J. 1985. Phylogenies and the Comparative Method. The American Naturalist [Internet] 125:1–15. Available from: http://dx.doi.org/10.1086/284325

Fierer N. 2017. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat. Rev. Microbiol. 15:579–590.

Freckleton, Freckleton, Harvey, Pagel. 2002. Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence. The American Naturalist [Internet] 160:712. Available from: http://dx.doi.org/10.2307/3078855

Garland T, Midford PE, Ives AR. 1999. An Introduction to Phylogenetically Based Statistical Methods, with a New Method for Confidence Intervals on Ancestral Values. American Zoologist [Internet] 39:374–388. Available from: http://dx.doi.org/10.1093/icb/39.2.374

Giovannoni SJ. 2017. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. Ann. Rev. Mar. Sci. 9:231–255.

Giovannoni SJ, Cameron Thrash J, Temperton B. 2014. Implications of streamlining theory for microbial ecology. ISME J. 8:1553–1565.

Goldman B, Bhat S, Shimkets LJ. 2007. Genome evolution and the emergence of fruiting body development in Myxococcus xanthus. PLoS One 2:e1329.

Graham CH, Storch D, Machac A. Phylogenetic scale in ecology and evolution. Available from: http://dx.doi.org/10.1101/063560

Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, Rappé MS. 2012. Streamlining and core genome conservation among highly divergent members of the SAR11 clade. MBio [Internet] 3. Available from: http://dx.doi.org/10.1128/mBio.00252-12

Guieysse B, Wuertz S. 2012. Metabolically versatile large-genome prokaryotes. Curr. Opin. Biotechnol. 23:467–473.

van Ham RCHJ, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernández JM, Jiménez L, Postigo M, Silva FJ, et al. 2003. Reductive genome evolution in Buchnera aphidicola. Proc. Natl. Acad. Sci. U. S. A. 100:581–586.

Harmon LJ, Losos JB, Jonathan Davies T, Gillespie RG, Gittleman JL, Bryan Jennings W, Kozak KH, McPeek MA, Moreno-Roark F, Near TJ, et al. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396.

Harris CD, Torrance EL, Raymann K, Bobay L-M. 2021. CoreCruncher: Fast and Robust Construction of Core Genomes in Large Prokaryotic Data Sets. Mol. Biol. Evol. 38:727–734.

Hernández CE, Rodríguez-Serrano E, Avaria-Llautureo J, Inostroza-Michael O, Morales-Pallero B, Boric-Bargetto D, Canales-Aguirre CB, Marquet PA, Meade A. 2013. Using phylogenetic information and the comparative method to evaluate hypotheses in macroecology. Methods in Ecology and Evolution [Internet] 4:401–415. Available from: http://dx.doi.org/10.1111/2041-210x.12033

Kamilar JM, Cooper N. 2013. Phylogenetic signal in primate behaviour, ecology and life history. Philosophical Transactions of the Royal Society B: Biological Sciences [Internet] 368. Available from: http://dx.doi.org/10.1098/rstb.2012.0341

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, et al. 2014. Single-cell genomics reveals hundreds of coexisting

subpopulations in wild Prochlorococcus. Science 344:416–420.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463–1464.

Klappenbach JA, Dunbar JM, Schmidt TM. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl. Environ. Microbiol. 66:1328–1333.

Konstantinidis KT, Tiedje JM. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc. Natl. Acad. Sci. U. S. A. 101:3160–3165.

Koonin EV. 2009. Evolution of genome architecture. Int. J. Biochem. Cell Biol. 41:298–306.

Kuo C-H, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res. 19:1450–1454.

Lan R, Reeves PR. 2000. Intraspecies variation in bacterial genomes: the need for a species genome concept. Trends Microbiol. 8:396–401.

Lawrence JG, Hendrickson H. 2005. Genome evolution in bacteria: order beneath chaos. Curr. Opin. Microbiol. 8:572–578.

Lawrence JG, Hendrix RW, Casjens S. 2001. Where are the pseudogenes in bacterial genomes? Trends Microbiol. 9:535–540.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47:W256–W259.

López-Pérez M, Haro-Moreno JM, Coutinho FH, Martinez-Garcia M, Rodriguez-Valera F. 2020. The Evolutionary Success of the Marine Bacterium SAR11 Analyzed through a Metagenomic Perspective. mSystems [Internet] 5. Available from: http://dx.doi.org/10.1128/mSystems.00605-20

Luo H, Huang Y, Stepanauskas R, Tang J. 2017. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. Nat Microbiol 2:17091.

Lynch M. 2006. Streamlining and simplification of microbial genome architecture. Annu. Rev. Microbiol. 60:327–349.

Lynch M, Conery JS. 2003. The Origins of Genome Complexity. Science [Internet] 302:1401–1404. Available from: http://dx.doi.org/10.1126/science.1089370

Marais GAB, Batut B, Daubin V. 2020. Genome Evolution: Mutation Is the Main Driver of Genome Size in Prokaryotes. Curr. Biol. 30:R1083–R1085.

Marais GAB, Calteau A, Tenaillon O. 2008. Mutation rate and genome reduction in endosymbiotic and free-living bacteria. Genetica 134:205–210.

Martinez-Gutierrez CA, Aylward FO. 2019. Strong Purifying Selection Is Associated with Genome Streamlining in Epipelagic Marinimicrobia. Genome Biol. Evol. 11:2887–2894.

Martinez-Gutierrez CA, Aylward FO. 2021. Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. Mol. Biol. Evol. [Internet]. Available from: http://dx.doi.org/10.1093/molbev/msab254

McCutcheon JP, Moran NA. 2011. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 10:13–26.

Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30:1188–1195.

Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17:589–596.

Molina-Venegas R, Rodríguez MÁ. 2017. Revisiting phylogenetic signal; strong or negligible impacts of polytomies and branch length information? BMC Evol. Biol. 17:53.

Moran NA, Mira A. 2001. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol. 2:RESEARCH0054.

Münkemüller T, Lavergne S, Bzeznik B, Dray S, Jombart T, Schiffers K, Thuiller W. 2012. How to measure and test phylogenetic signal. Methods in Ecology and Evolution [Internet] 3:743–756. Available from: http://dx.doi.org/10.1111/j.2041-210x.2012.00196.x

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Niederdorfer R, Besemer K, Battin TJ, Peter H. 2017. Ecological strategies and metabolic trade-offs of complex environmental biofilms. NPJ Biofilms Microbiomes 3:21.

Nielsen DA, Fierer N, Geoghegan JL, Gillings MR, Gumerov V, Madin JS, Moore L, Paulsen IT, Reddy TBK, Tetu SG, et al. 2021. Aerobic bacteria and archaea tend to have larger and more versatile genomes. Oikos [Internet] 130:501–511. Available from: http://dx.doi.org/10.1111/oik.07912

O'Meara BC, Ané C, Sanderson MJ, Wainwright PC. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Osburne MS, Holmbeck BM, Coe A, Chisholm SW. 2011. The spontaneous mutation frequencies of Prochlorococcus strains are commensurate with those of other bacteria. Environ. Microbiol. Rep. 3:744–749.

Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature [Internet] 401:877–884. Available from: http://dx.doi.org/10.1038/44766

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics [Internet] 35:526–528. Available from: http://dx.doi.org/10.1093/bioinformatics/bty633

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25:1043–1055.

Partensky F, Garczarek L. 2010. Prochlorococcus: advantages and limits of minimalism. Ann. Rev. Mar. Sci. 2:305–331.

Revell LJ. 2010. Phylogenetic signal and linear regression on species data. Methods in Ecology and Evolution [Internet] 1:319–329. Available from: http://dx.doi.org/10.1111/j.2041-210x.2010.00044.x

Revell LJ, Collar DC. 2009. PHYLOGENETIC ANALYSIS OF THE EVOLUTIONARY CORRELATION USING LIKELIHOOD. Evolution [Internet] 63:1090–1100. Available from: http://dx.doi.org/10.1111/j.1558-5646.2009.00616.x

Revell LJ, Harmon LJ, Collar DC. 2008. Phylogenetic signal, evolutionary process, and rate. Syst. Biol. 57:591–601.

Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J. Theor. Biol. 239:226–235.

Rodríguez-Gijón A, Nuy JK, Mehrshad M, Buck M, Schulz F, Woyke T, Garcia SL. A genomic perspective across Earth's microbiomes reveals that genome size in Archaea and Bacteria is linked to ecosystem type and trophic strategy. Available from: http://dx.doi.org/10.1101/2021.01.18.427069

Sela I, Wolf YI, Koonin EV. 2016. Theory of prokaryotic genome evolution. Proc. Natl. Acad. Sci. U. S. A. 113:11399–11407.

Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Science [Internet] 27:135–145. Available from: http://dx.doi.org/10.1002/pro.3290

Simonsen AK. 2021. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. The ISME Journal [Internet]. Available from: http://dx.doi.org/10.1038/s41396-021-01082-x

Soubrier J, Steel M, Lee MSY, Der Sarkissian C, Guindon S, Ho SYW, Cooper A. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. Mol. Biol. Evol. 29:3345–3358.

Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, et al. 2013. Metagenomic species profiling using universal phylogenetic marker genes. Nat. Methods 10:1196–1199.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research [Internet] 34:W609–W612. Available from: http://dx.doi.org/10.1093/nar/gkl315

Takeuchi N, Cordero OX, Koonin EV, Kaneko K. 2015. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. BMC Biol. 13:20.

Tamas I, Klasson L, Canbäck B, Näslund AK, Eriksson A-S, Wernegreen JJ, Sandström JP, Moran NA, Andersson SGE. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science 296:2376–2379.

Whitney KD, Boussau B, Baack EJ, Garland T Jr. 2011. Drift and genome complexity revisited. PLoS Genet. 7:e1002092.

Whitney KD, Garland T Jr. 2010. Did genetic drift drive increases in genome complexity? PLoS Genet. [Internet] 6. Available from: http://dx.doi.org/10.1371/journal.pgen.1001080

Wolf YI, Aravind L, Grishin NV, Koonin EV. 1999. Evolution of aminoacyl-tRNA synthetases--analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res. 9:689–710.

Woolfit M, Bromham L. 2003. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. Mol. Biol. Evol. 20:1545–1555.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Chapter 5. Research Project

A Phylogenomic Timeline of Bacterial and Archaeal Diversification in the Ocean

Previously submitted: Martinez-Gutierrez CA, Uyeda JC and Aylward FO. 2022. A Phylogenomic Timeline of Bacterial and Archaeal Diversification in the Ocean. Under Review.

Co-authors contributed in the following ways: Conceived and designed this work: CAMG, UJC, and FOA. Wrote the paper: CAMG, UJC, and FOA.

5.1 Abstract

Microbial plankton play a central role in marine biogeochemical cycles, but the timing in which abundant lineages colonized contemporary ocean environments remains unclear. Here, we reconstructed the geological dates in which major clades of bacteria and archaea diversified into the ocean using a high-resolution benchmarked phylogenetic tree, which allows for simultaneous and direct comparison of the diversification of multiple divergent lineages. Our findings suggest that extant prokaryotic diversity is the product of three main waves of colonization that followed major oxygenation events. The first took place after the initial oxygenation of the ocean 2.3 Ga, after which several lineages that thrive in microaerophilic conditions colonized marine niches and continue to proliferate in oxygen minimum zones today. The second two events took place around 0.8 and 0.4 Ga, respectively, and were followed by the colonization of lineages that currently drive key biogeochemical cycles in contemporary oxygenated niches. Our work clarifies the timing at which abundant lineages of bacteria and archaea diversified into the ocean, allow us to link their diversification with key geological events throughout Earth's history, and demonstrates that the redox state of the ocean was likely the primary factor that drove major diversification events.

5.2 Main Text

The ocean plays a central role in the functioning and stability of Earth's biogeochemistry (Falkowski et al. 1998; Field et al. 1998; Falkowski et al. 2008). Due to their abundance, diversity, and physiological versatility, microbes mediate the vast majority of the metabolic activities that lead to the organic matter transformations that sustain higher trophic levels (Falkowski et al. 1998; Falkowski et al. 2008; Mason et al. 2009; Brown et al. 2014). For instance, marine microorganisms regulate a large fraction of the organic carbon pool (Ducklow and Doney 2013), drive elemental cycling of nitrogen, phosphorus, sulfur, and iron, (Zehr and Kudela 2011), and participate in the ocean-atmosphere exchange of climatically important gasses (Vila-Costa et al. 2006). Starting in the 1980s, analysis of small-subunit ribosomal RNA genes began to reveal the identity of dominant clades of bacteria and archaea that were notable for their ubiquity and high abundance, and subsequent analyses highlighted their diverse physiological activities in the ocean (Giovannoni and Stingl 2005). Phylogenetic studies showed that these clades are broadly distributed across the Tree of Life (TOL) and encompass a range of phylogenetic breadths (Giovannoni and Stingl 2005). Cultivation-based studies and the large-scale generation of genomes from metagenomes have continued to make major progress in examining the genomic diversity and metabolism of these major lineages, but we still lack a comprehensive understanding of the evolutionary events leading to their origin and diversification in the ocean.

Several independent studies have used molecular phylogenetic methods to date the diversification of some marine microbial lineages, such as the Ammonia Oxidizing Archaea of the order *Nitrososphaerales* (Marine Group I, MGI) (Ren et al. 2019; Yang et al. 2021), picocyanobacteria of the genera *Synechococcus* and *Prochlorococcus* (Sánchez-Baracaldo 2015; Sánchez-Baracaldo et al. 2019; Zhang et al. 2021), and marine alphaproteobacterial groups that

included the SAR11 and Roseobacter clades (Luo et al. 2013). Discrepancies in the methodological frameworks used in these studies often hinder comparisons between clades, however, and results for individual clades often conflict (Sánchez-Baracaldo 2015; Ren et al. 2019; Yang et al. 2021; Zhang et al. 2021). Moreover, it has been difficult to directly evaluate bacterial and archaeal clades due to the vast evolutionary distances between these domains. For these reasons it has remained challenging to compare the diversification of different marine lineages and develop a comprehensive understanding of microbial diversification in the ocean and its relationship with major geological events throughout Earth's history.

To address these challenges, we constructed a multi-domain phylogenetic tree that allowed us to directly compare the diversification dates of 13 planktonic marine bacterial and archaeal clades that are known for their abundance and major roles in marine biogeochemical cycles in the modern ocean (Fig. 5.1). Tree reconstruction was based on a benchmarked set of marker genes that we have previously shown to be congruent for inter-domain phylogenetic reconstruction ((Martinez-Gutierrez and Aylward 2021), details in Methods, Supplemental File 2). Our phylogenetic framework included non-marine clades for phylogenetic context, and overall recapitulates known relationships across the Tree of Life (ToL), including the clear demarcation of the Gracilicutes and Terrabacteria superphyla in Bacteria). To gain insight into the geological landscape in which these major marine clades first diversified, we performed a Bayesian relaxed molecular dating analysis on our benchmarked Tee of Life (ToL) using several calibrated nodes (Fig. 5.1 and Table 5.1). Due to the limited representation of microorganisms in the fossil record and the difficulties to associate fossils to taxonomic groups, we employed geochemical evidence as calibration points (Table 1). Moreover, because of the uncertainty in the length of the branch linking bacteria and archaea, the crown node for each domain was calibrated independently. We

used the age of the presence of liquid water as approximated through the dating of zircons (Valley et al. 2014), as well as the most ancient record of biogenic methane (Ueno et al. 2006; Valley et al. 2014) as maximum and minimum prior ages for bacteria and archaea (4400 and 3460 My, respectively, Fig. 5.1 and Table 5.1). For internal calibration, we used the recent identification of non-oxygenic Cyanobacteria to constrain the diversification node of oxygenic Cyanobacteria with a minimum age of 2320 My, the estimated age for the Great Oxidation Event (GOE) (Holland 2002; Bekker et al. 2004; Holland 2006). Similarly, we applied this reasoning for the calibration of the crown group of aerobic Ammonia Oxidizing Archaea, aerobic Ca. *Marinimicrobia*, and the Nitrite oxidizing bacteria, using their strict aerobic metabolism as evidence for a maximum age of 2320 Ga (Fig. 5.1 and Table 5.1).

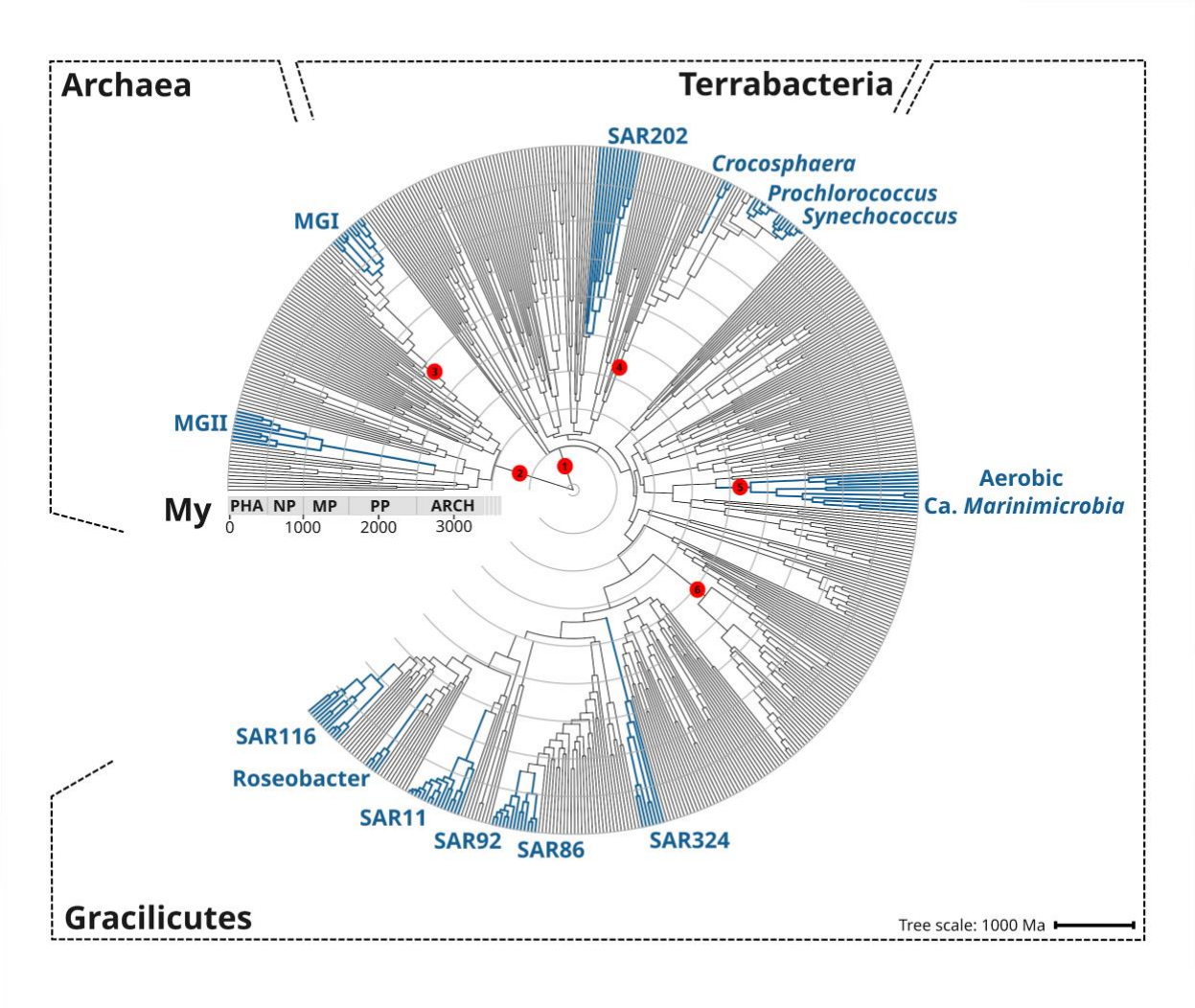

Figure 5.1 Rooted inter-domain Tree of Life used for molecular dating analyses. Maximum likelihood tree built using the concatenation of 30 RNAP subunits and ribosomal protein sequences and the substitution model LG+R10. Colored branches represent the marine clades dated in our study. Red dots show the temporal calibration points used in our molecular dating analyses (Table 1).

Table 5.1 Temporal calibrations used as priors for the molecular dating of the main marine clades. Node on Tree column refers to the calibrated nodes shown in Fig. 5.1.

Node on Tree	Calibration group	Minimum (MY)	Maximum (MY)	Evidence	Reference
1,2	Bacteria-Archaea Root	-	4400	Identification of the most ancient zircons showing evidence of liquid water.	Valley et al., 2014
1,2	Bacteria-Archaea Root	3460	-	Identification of the most ancient traces of methane.	Ueno et al., 2006
3	Aerobic *Nitrososphaerales*	-	2320	Strict aerobic metabolism.	Holland, 2006
4	Oxygenic Cyanobacteria	2320	-	Oxygenation of the atmosphere. The Great Oxidation Event has been associated with oxygenic Cyanobacteria.	Holland, 2006
5	Aerobic Marinimicrobia	-	2320	Strict aerobic metabolism.	Holland, 2006
6	Nitrite oxidizing bacteria	-	2320	Strict aerobic metabolism.	Holland, 2006

Our Bayesian estimates demonstrate that the diversification of the major clades of marine bacteria and archaea can be broadly divided into three waves that were coincident with the major oxygenation events of the atmosphere and the ocean (Fig. 5.2). The first wave occurred near the time of the Great Oxidation Event (GOE), and included the clades SAR202, aerobic Ca. *Marinimicrobia*, SAR324, and the Marine Group II of the phylum Euryarchaeota (MGII). Within this first wave, the most ancient clade was the SAR202 (2479 My, 95% CI = 2465-2492 My), whose diversification took place near before the GOE (Fig. 5.2). The diversification SAR202

before the first oxygenation event suggests that this group evolved under the oxygen oasis proposed to have existed in pre-GOE Earth (Anbar et al. 2007; Ossa Ossa et al. 2019; Reinhard and Planavsky 2022). Moreover, a recent study proposed that SAR202 played a role in the shift of the redox state of the atmosphere during the GOE by partially metabolizing organic matter through a flavin dependent Baeyer–Villiger monooxygenase, thereby enhancing the burial of organic matter and contributing to the net accumulation of oxygen in the atmosphere (Landry et al. 2017; Shang et al. 2022). After the GOE, we detected the diversification of aerobic Ca. *Marinimicrobia* (2196 My, 95% CI = 2173-2219 My), the SAR324 clade (1686 My, 95% CI = 1658-1715 My), and the MGII clade (1184 My, 95% CI = 1166-1202 My) (Fig. 5.2). Although these ancient clades may have first diversified in an oxic environment, the abrupt first increase of oxygen during the GOE was followed by a relatively rapid drop in ocean and atmosphere oxygen levels (Alcott et al. 2019; Hodgskiss et al. 2019; Reinhard and Planavsky 2022). It is therefore likely that these clades diversified in the microaerophilic and variable oxygen conditions that prevailed during this period (Holland 2002; Bekker et al. 2004; Holland 2006). Indeed, the oxygen landscape in which these marine clades first diversified is consistent with their current physiology. Although these groups are capable of using oxygen as terminal electron acceptor, they are prevalent in marine oxygen minimum zones (OMZs) today, where they use a wide range of alternative electron acceptors (e.g., nitrate) (Sheik et al. 2014; Pajares et al. 2020; Thrash et al.). The facultative aerobic or microaerophilic metabolism in these clades is likely a vestige of the low oxygen environment of most of the Proterozoic Eon, and in this way OMZs can be considered to be modern-day refugia of these ancient ocean conditions.

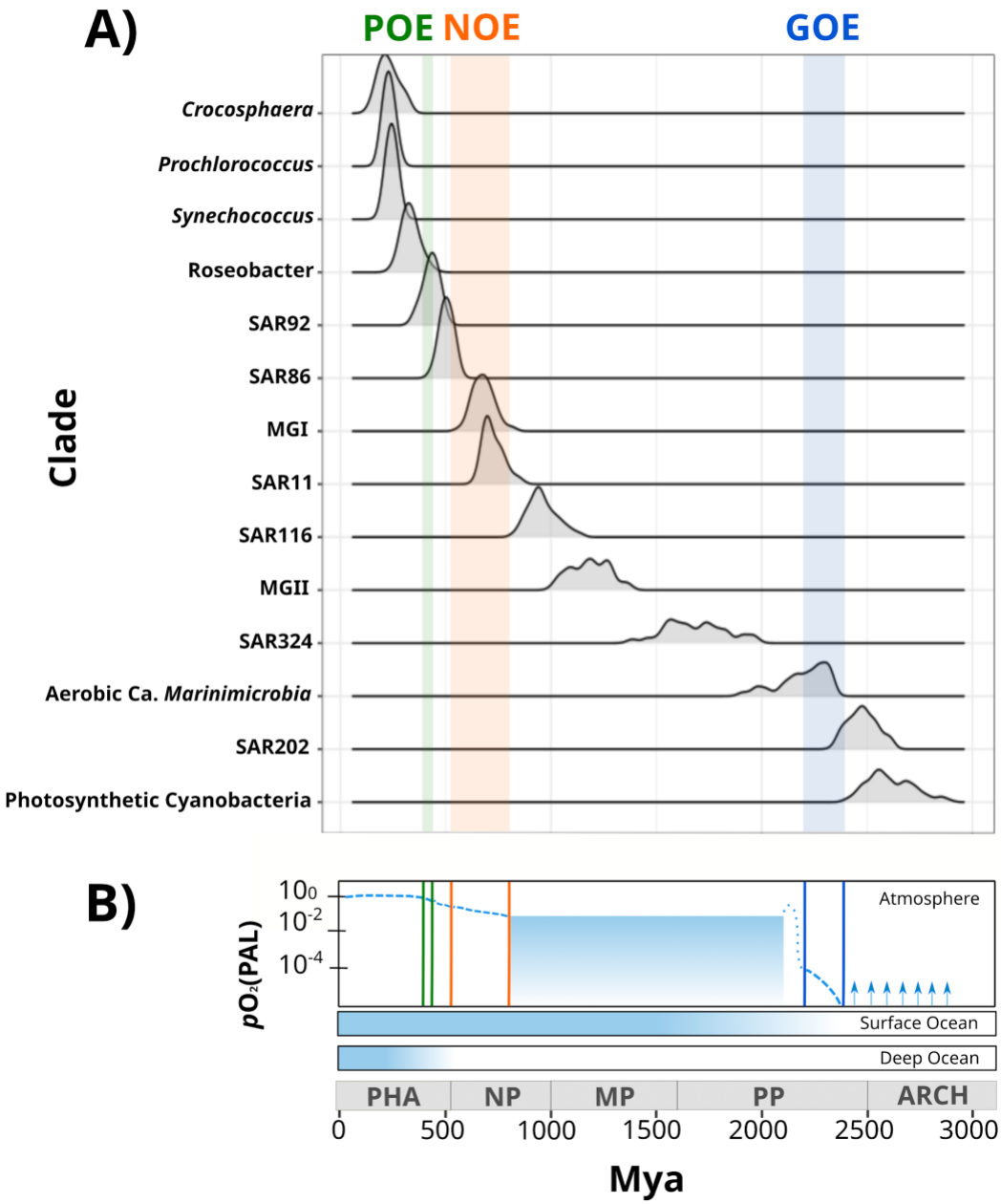

Figure 5.2 Age distribution of marine microbial clades and its relationship with the main Earth oxygenation events. Ridges represent 100 age estimates from one of the Bayesian chains (see methods). Abbreviations: POE: Paleozoic Oxidation Event; NOE: Neoproterozoic Oxidation Event; GOE: Great Oxidation Event; Pha: Paleozoic; NP: Neoproterozoic; MP: Mesoproterozoic; PP: Paleoproterozoic; Arch; Archaean. Low panel adapted from previous work (Alcott et al., 2019).

The second wave of diversification can be traced back to the time around the Neoproterozoic Oxygenation Event (POE) that occurred 800-540 My (Och and Shields-Zhou

2012; Alcott et al. 2019), and included the clades SAR116 (959 My, 95% CI = 945-973 My), SAR11 (725 My, 95% CI = 715-734 My), SAR86 (503 My, 95% CI = 497-509 My), SAR92 (430 My, 95% CI = 423-437 My), and Roseobacter (332 My, 95% CI = 323-340 My) (Fig. 5.2). The relative late appearance of these heterotrophic lineages abundant in the open ocean today is consistent with the low productivity and low oxygen concentrations in both shallow and deep waters that likely prevailed during the Mid-Proterozoic (1800-800 My), a period known as the "boring billion" (Anbar and Knoll 2002; Holland 2006; Planavsky et al. 2014; Tang et al. 2016; Crockford et al. 2018; Hodgskiss et al. 2019; Reinhard and Planavsky 2022). The diversification of these clades may be indirectly associated with the tectonic activity and a Snowball event before the NOE (Hoffman et al. 1998; Anbar and Knoll 2002; Shields-Zhou and Och 2011), which increased the availability of nutrients (Anbar and Knoll 2002) and is also coincident with the widespread diversification of large eukaryotic algae during the Late-Proterozoic (Vidal and Moczydłowska-Vidal 1997; Butterfield 2001; Anbar and Knoll 2002; Porter 2004; Shields-Zhou and Och 2011). It is therefore plausible that an increase in nutrients as well as the broad diversification of eukaryotic plankton enhanced the mobility of organic and inorganic nutrients beyond the coastal areas, and increased the burial of organic matter that consequently led to an increment in atmospheric oxygen concentrations (Knoll et al. 2006; Shields-Zhou and Och 2011). The scenario in which heterotrophic marine clades diversified in part as a consequence of the new niches built by marine eukaryotes has been previously proposed to have driven the diversification of the Roseobacter clade (Luo et al. 2013; Luo and Moran 2014). Our results suggest this phenomenon broadly influenced the diversification of several other lineages.

We also registered the diversification of the chemolithoautotroph archaeal MGI into the ocean during this second wave (678 My, 95% CI = 668-688 My) (Fig 2 and 3), which is

comparable with the age reported by another independent study (Yang et al. 2021). This is consistent with an increase in the oxygen concentrations of the ocean during this period (Reinhard and Planavsky 2022), a necessary requisite for ammonia oxidation. Moreover, the widespread sulfidic conditions that have been proposed to have prevailed in the Mid-Proterozoic ocean likely limited the availability of redox-sensitive metals like Cu, necessary for ammonia monooxygenases (Anbar and Knoll 2002; Hatzenpichler 2012). It is therefore possible that a low concentration of oxygen and limited inorganic nutrients registered before the NOE were limiting factors that delayed the colonization of AOA into the ocean.

The most recent and last wave of microbial diversification led to the appearance of late-branching phototrophs of the genera *Synechococcus* (243 My, 95% CI = 238-247 My), *Prochlorococcus* (230 My, 95% CI = 225-234 My), and the nitrogen-fixer *Crocosphaera* (228 My, 95% CI = 218-237 My). These dates agree with an independent study that points to a relatively late evolution of the marine cyanobacterial clades *Prochlorococcus* and *Synechococcus* (Sánchez-Baracaldo 2015). Picocyanobacteria and *Crocosphaera* are known for being essential components of phytoplanktonic communities in the modern open ocean due to their large contribution to carbon and nitrogen fixation, respectively (Montoya et al. 2004; Scanlan et al. 2009; Flombaum et al. 2013). For example, the nitrogen fixation activities of *C. watsonii* in the open ocean today support the demands of nitrogen-starved microbial food webs found in oligotrophic waters (Hewson et al. 2009). The late diversification of these lineages suggests that the oligotrophic open ocean is a relatively new ecosystem (Fig. 5.3). The oligotrophic ocean today is characterized by the rapid turnover of nutrients that depends on the efficient mobilization of essential elements through the ocean (Karl 2002). Due to its distance from terrestrial nutrient inputs, productivity in the open ocean is therefore dependent on local nitrogen fixation, which was probably enhanced after the

widespread oxygenation of the ocean that made Mo widely available due to its high solubility in oxic seawater (Canfield et al. 2007; Scott et al. 2008; Wei et al. 2021). Such widespread oxygenation was registered 430-390 My in an event referred to here as the Paleozoic Oxidation Event (Berner and Raiswell 1983; Sperling et al. 2015; Lenton et al. 2016; Tostevin and Mills 2020) (POE, Fig. 5.2-5.3). The definitive oxygenation of the atmosphere and the ocean was the result of an increment of the burial of organic carbon in sedimentary rocks due to the diversification of the earliest land plants (Lenton et al. 2016; Planavsky et al. 2021; Reinhard and Planavsky 2022), which has been also associated with increased phosphorus weathering rates (Bergman et al. 2004; Lenton et al. 2016), global impacts on the global element cycles (Dahl and Arens 2020), and an increase in overall ocean productivity (Planavsky et al. 2021). The late diversification of oligotrophic-specialized clades after the POE suggests that the establishment of the oligotrophic open ocean as we know it today would only have been possible once modern oxygen concentrations and biogeochemical dynamics were reached (Karl 2002; Reinhard and Planavsky 2022).

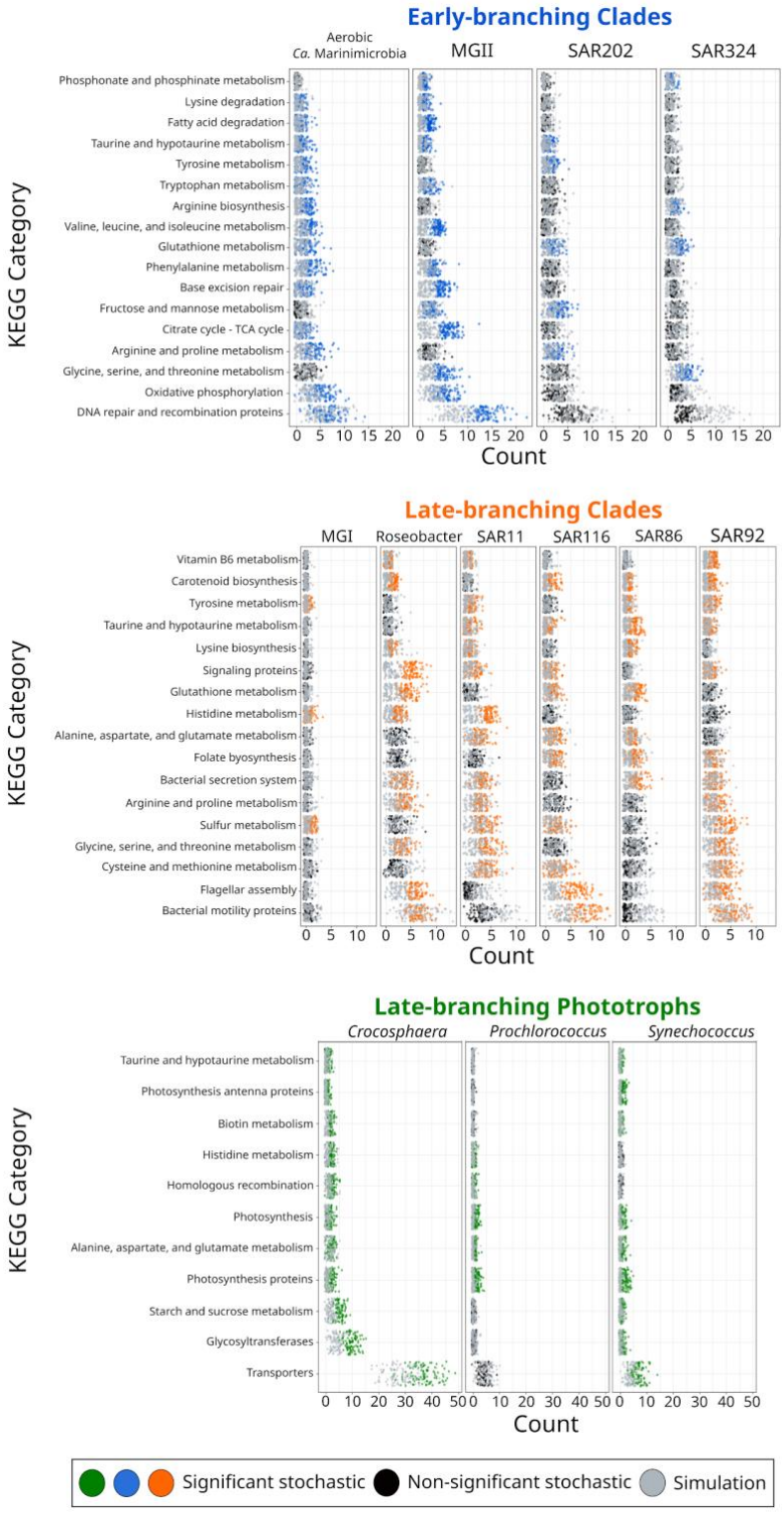

Figure 5.3 KEGG categories enriched in each diversification wave. KEGG categories enriched at the crown node of each marine microbial clade. Clades were classified based on the diversification timing shown in Fig. 2. Enriched categories were identified by statistically

comparing a stochastic mapping distribution with an all-rates-different vs a null distribution with a constant rate model without conditioning on the presence/absence data at the tips. Each dot represents one replicate (See Methods). X-axis represents the number of KOs gained at each crown node for each KEGG category. Stochastic mapping and null distributions were sorted for visualization purposes.

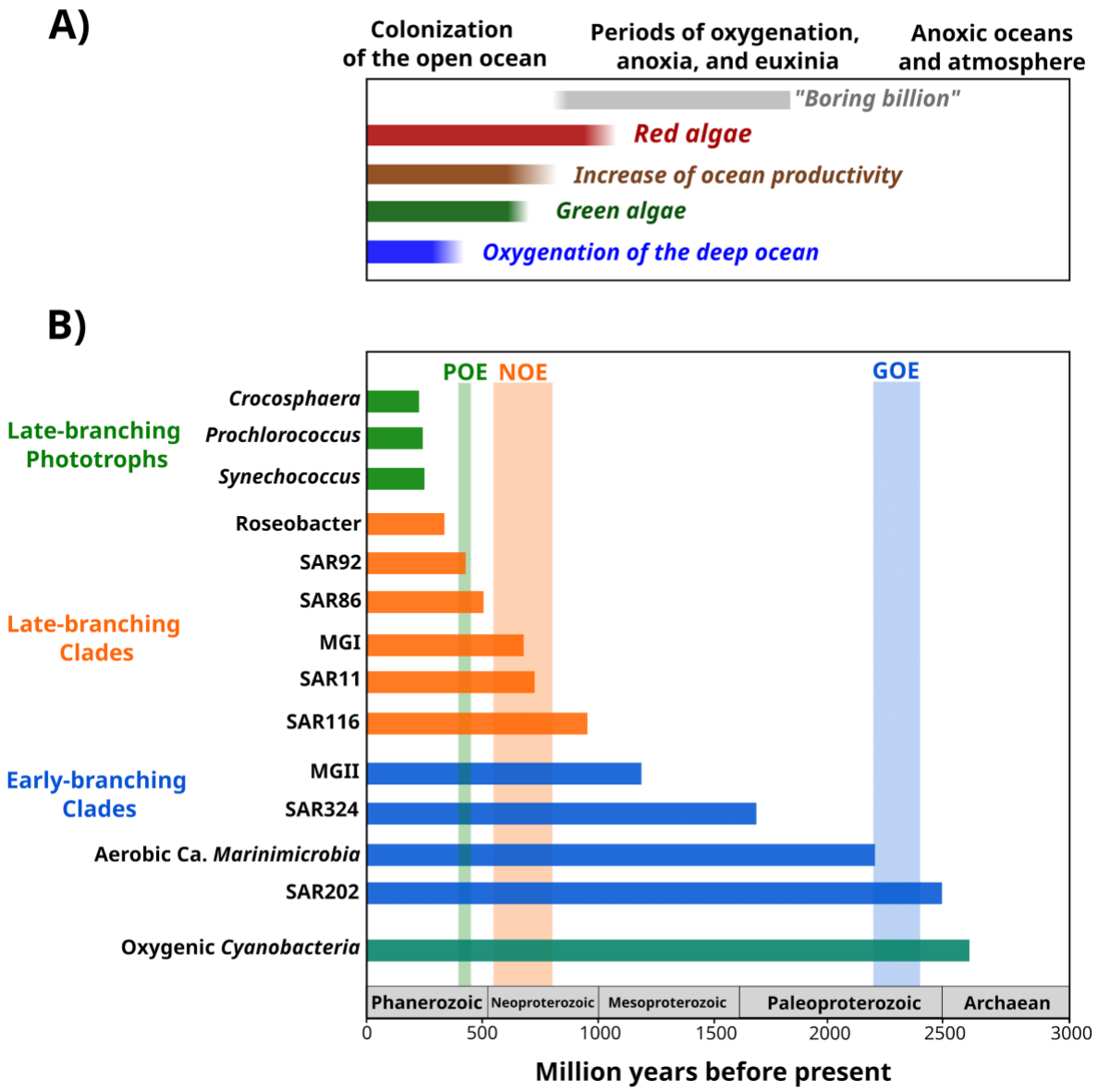

Figure 5.4 Summary of the relationship between the timing of the diversification of the marine microbial clades and the main geological and biological events. A) Geological and biological events potentially involved in the diversification of marine clades. Ages show are based on previously published data: "Boring billion" (Brasier and Lindsay 1998), red algae fossils (Butterfield 2000), increased of ocean productivity (Och and Shields-Zhou 2012), green algae

fossils (Moczydłowska 1996), oxygenation of the deep ocean (Lenton et al. 2016). B) Hypothetical association between the diversification of marine clades and oxidation events throughout Earth's history. Length of each bar represents the estimated age for marine clades based on a Bayesian approach. The timing of the oxygenation events is based on previous work (Alcott et al. 2019).

In order to shed further light on the drivers of the colonization of the ocean, we investigated whether the diversification of marine microbial clades was linked to the acquisition of novel metabolic capabilities. Due to the remarkable timing of diversification during the waves described previously, we classified the marine clades as late-branching phototrophs, late-branching clades, and early branching clades based on their diversification timing (Fig. 5.3 and 4). To identify the enrichment of KEGG categories at the nodes of marine clades diversification (Fig. 5.1), we performed 100 cycles of stochastic mapping analysis on each of the 112,248 protein families encoded in our genome's dataset. We compared our results with a null hypothesis distribution in which a constant rate model was implemented unconditionally of observed data (see Methods). Statistical comparisons of the stochastic and the null distribution show that each diversification wave was associated with the enrichment of specific KEGGs categories that were consistent with the geochemical context of their diversification (Fig. 5.3). For example, Early-branching Clades (EBC; Fig. 5.3 and 5.4) diversifying near the time of the GOE showed an overrepresentation of proteins involved in DNA repair, Recombination, Base excision repair, and Glutathione metabolism, probably as a result of the increase of oxygen during the GOE and the rise of reactive oxygen species (Masip et al. 2006; Burrows 2009; Khademian and Imlay 2021). Moreover, the EBC were enriched in proteins involved in ancient aerobic pathways, including oxidative phosphorylation and TCA cycle (Fig. 5.3), as well as genes implicated in the degradation of fatty acids under aerobic conditions, for example the enzyme alkane 1-monooxygenase in MGII (Supplemental File 6). We also identified the enrichment of proteins for the metabolism of Phosphonate, a common phosphate source in the ocean and Taurine (Acker et al. 2022), an

osmoprotectant found in bacteria (McParland et al. 2021). The enrichment of these KEGG categories likely represents specific adaptations to the marine environment. Our findings suggest that the diversification of EBC in the ocean was linked to the evolution of aerobic metabolism, the acquisition of metabolic capabilities to exploit the newly created niches due to the increase of oxygen, and the expansion of genes to deal with oxidative stress.

The emergence of Late-branching Clades (LBC; Fig. 5.3 and 5.4), whose diversification occurred around the time of the NOE (Shields-Zhou and Och 2011) and the proposed first diversification of eukaryotic algae (Parfrey et al. 2011), was characterized by the enrichment of substantially different gene repertories compared to EBC (Fig. 5.3). For instance, the heterotrophic LBC Roseobacter, SAR116, and SAR92 show an enrichment of flagellar assembly and motility genes (Fig. 5.3), including genes for flagellar biosynthesis, flagellin, and the flagellar basal-body rod protein (Supplemental File 7). Motile marine heterotrophs like Roseobacter species have been associated with the marine phycosphere, a region surrounding individual phytoplankton cells releasing carbon-rich nutrients (Seymour et al. 2017; Mühlenbruch et al. 2018). Although the phycosphere can also be found in prokaryotic phytoplankton (Seymour et al. 2017), given the late diversification of abundant marine prokaryotic phytoplankton (Fig. 5.3), it is plausible that the emergence of these clades was closely related to the establishment of ecological proximity with eukaryotic algae. The potential diversification of heterotrophic LBC due to their ecological proximity with eukaryotic algae was supported by the enrichment of Vitamin B6 metabolism and Folate Biosynthesis proteins, which are key nutrients involved in phytoplankton-bacteria interactions (Croft et al. 2005; Seymour et al. 2017). Furthermore, we identified the enrichment of Carotenoid Biosynthesis proteins, for example a spheroidene monooxygenase, carotenoid 1,2-hydratase, beta-carotene hydroxylase, and lycopene beta-cyclase. These genes might be associated

with the presence of proteorhodopsin, light-driven proton pumps broadly distributed in marine microorganisms inhabiting energy-depleted oligotrophic areas (de la Torre et al. 2003). We also identified the enrichment of genes for the metabolism of taurine, an osmoprotectant found in bacteria (McParland et al. 2021), but also in marine metazoans and algae, and commonly used as source carbon by marine prokaryotes (Clifford et al. 2019), strengthening our findings that the diversification of these clades was linked to the use of organic nutrients produced by eukaryotes. Although the diversification of MGI during this period, we only identified amino acids metabolism and sulfur metabolism genes. The enrichment of genes potentially involved in bacteria-phytoplankton interactions suggest that the diversification of marine clades during the NOE was intimately linked to the establishment of ecological relationships with eukaryotes to exchange nutrients.

Late-branching phototrophs that diversified around the time of the POE (LBP; Fig. 5.3 and 5.4), showed enrichment of transporters in *Synechococcus* and *Crocosphaera* (Fig. 5.3). In particular, the diversification of *Crocosphaera* was characterized by the acquisition of transporters for inorganic nutrients like cobalt, nickel, iron, phosphonate, phosphate, ammonium, and magnesium, along with organic nutrients including amino acids and polysaccharides (Supplemental File 6). The acquisition of a wide diversity of transporters by the *Crocosphaera* is consistent with their blooming lifestyle seen in the open ocean today (Hewson et al. 2009; Wilson et al. 2017), which requires a rapid and efficient use of the scarce nutrients available. We also identified genes for osmotic pressure tolerance, for example a Ca-activated chloride channel homolog, a magnesium exporter, and a fluoride exporter (Supplemental File 6). The acquisition of salt-tolerance genes suggests that *Crocosphaera* might have diversified from a non-marine group into the ocean. Contrastingly, our results show that *Synechococcus* only acquired transporters for

inorganic nutrients (e.g., iron and sulfate), whereas *Prochlorococcus* did not show an enrichment of transporters (Supplemental File 6). The absence of salt-tolerance related genes suggests that the ancestor of these Picocyanobacterial clades inhabited a low-nutrient marine habitat. Similar to LBC, we identified the enrichment of taurine metabolism KOs in *Crocosphaera* and *Synechococcus*, suggesting that its use as osmoprotectant and potential substrate is widespread among planktonic microorganisms (Clifford et al. 2019). Prochlorococcus show enrichment in fewer categories than the rest of phototrophic clades diversifying during the same period, however we still observed the enrichment of photosynthesis-associated genes (Fig. 5.3). This is supported by previous findings that suggest that the diversification of this clade was accompanied by changes in the photosynthetic apparatus compared with *Synechococcus*, its sister group (Biller et al. 2015). Overall, the diversification of LBP was marked by the capacity to thrive in the oligotrophic ocean by exploiting organic and inorganic nutrients and modifying the photosynthetic apparatus as observed in *Crocosphaera* and *Synechococcus*, and *Prochlorococcus*, respectively.

5. 3 Outlook

Throughout Earth's history, major diversification events are typically accompanied by large-scale geological events due to the close interaction between biotic processes and abiotic conditions. The specific evolutionary and geological events that have driven the diversification of microbial lineages in the ocean today have remained poorly understood, however, due to a combination of the inherent difficulties of studying biological events that occurred in deep time and the lack of a fossil record for microbial life. Here we present a comprehensive timeline for the colonization of abundant marine clades into the ocean and reveal that major oxygenation events in Earth's history played critical roles in creating new niches for microbial diversification. These colonization events

subsequently led to the establishment of the biogeochemical cycles that govern the environmental stability of our planet today.

5.4 Material and methods

5.4.1 Genomes sampling and phylogenetic reconstruction

In order to obtain a comprehensive understanding of the diversification of the main marine planktonic clades, we built a multi-domain phylogenetic tree that included a broad diversity of bacterial and archaeal genomes. We compiled a balanced genome dataset subsampled from the Genome Taxonomy Database (GTDB, v95, (Chaumeil et al. 2019)), including marine representatives. In addition, we improved the representation of marine genomes by subsampling genomes from the GORG database (Pachiadaki et al. 2019), and adding several Thermoarchaeota genomes available on the JGI (Nordberg et al. 2014). We discarded genomes belonging to the DPANN superphylum due to the uncertainty of their placement within the Archaea (Martinez-Gutierrez and Aylward 2021). The list of genomes used is reported in Supplemental File 1.

We reconstructed a phylogenetic tree through the MarkerFinder pipeline developed previously (Martinez-Gutierrez and Aylward 2021), which resulted in an alignment of 27 ribosomal genes and three RNA polymerase genes (RNAP) (Martinez-Gutierrez and Aylward 2021). The MarkerFinder pipeline consists of 1) the identification of ribosomal and RNAP genes using HMMER v 3.2.1 with the reported model-specific cutoffs (Eddy 2011), 2) alignment with ClustalOmega (Eddy 2011; Sievers and Higgins 2018), and 3) concatenation of individual alignments. The resulting concatenated alignment was trimmed using trimAl (Capella-Gutiérrez et al. 2009) with the option -gt 0.1. Phylogenetic tree inference was carried out with IQ-TREE v1.6.12 (Nguyen et al. 2015) with the options -wbt, -bb 1000 (Minh et al. 2013), -m LG+R10 (substitution model previously selected with the option -m MFP according to the Bayesian

Information Criterion (Kalyaanamoorthy et al. 2017)), and --runs 5 to select the tree with the highest likelihood. The tree with the highest likelihood was manually inspected to discard the presence of topological inconsistencies and artifacts on iTOL (Letunic and Bork 2019) (Fig .1). Raw phylogenetic tree is presented in Supplemental File 2.

5.4.2 Assessment of Quality Tree

Due to the key importance of tree quality for the tree-dependent analysis performed in our study, we assessed the congruence of our prokaryotic ToL through the Tree Certainty metric (TC) (Salichos et al. 2014; Martinez-Gutierrez and Aylward 2021). Our estimate (TC = 0.91) indicates high congruence in our phylogeny. We also evaluated whether the topology of our ToL is consistent with a high-quality prokaryotic ToL reported previously (Martinez-Gutierrez and Aylward 2021). In general, we observed consistency in the placement of the all the phyla, as well as the bacterial superphyla (Terrabacteria and Gracilicutes) between both trees, except for the sisterhood of Actinobacteriota and Armatimonadota, which differs from the sisterhood of Actinobacteriota and Firmicutes in the reference tree (Martinez-Gutierrez and Aylward 2021). Despite the discrepancy observed, we do not expect a substantial change in our conclusions regarding the diversification of the marine clades derived from the molecular dating and stochastic mapping analyses.

5.4.3 Estimating the divergence of bacterial and archaeal marine clades through molecular dating

In order to investigate the divergence time of the marine planktonic clades of interest, the phylogenetic tree obtained above was used to perform a molecular dating analysis using a Bayesian framework. Our analysis was performed through Phylobayes v4.1c (Lartillot et al. 2009) with the program pb on four independent chains. For each chain, the input consisted in the phylogenetic

tree, the amino acids alignment, the calibrations, and an autocorrelated relaxed log normal model (-ln) (Thorne et al. 1998) with the default molecular evolution model. Convergence was tested every 5000 cycles using the program tracecomp with a burn-in of 250 cycles and sampling every 2 cycles. We also ran an independent chain on the priors using the option -root to assess the suitability of our calibrations (Supplemental File 3 and 4). After 100,000 cycles, our chains reached convergence in 8 out of 12 parameters (Supplemental File 5). To assess the uncertainty derived from the parameters that did not reach convergence, we estimated the divergence ages for each of our four chains using the last 1000 cycles and a range of 10 cycles to have a sample of 100 age estimates using the program readdiv (Supplemental File 5). We report the confidence intervals of the 100 replicates throughout the manuscript.

In order to determine the impact of our priors (Fig. 5.1 and Table 5.1) on the age estimates of the calibrated nodes in our tree and assess the suitability of the ages used as priors for our analyses, we ran an independent MCMC chain without the amino acid alignment using the option -root on Phylobayes. Our prior-only analysis yielded a posterior age falling within the maximum and minimum priors used for the crown group of archaea and bacteria. For the internal calibrated nodes, we observed posterior estimates consistent with the priors used for each case except for aerobic ammonia oxidizing archaea (Supplemental File 5 and 6). Overall, this result suggests that the calibrations used as priors were adequate for our analyses.

To evaluate the reproducibility of our Bayesian molecular dating analysis, we applied an independent second approach based on Penalized Likelihood (PL) through TreePL (Smith and O'Meara 2012) on 1000 replicate bootstrap trees that had fixed topology but varying branch lengths. Replicate trees were generated with the program bsBranchLenghts available on RAxML

v8.2.12 (Stamatakis 2014). For each replicate run, we initially used the option "prime" to identify the optimization parameters and applied the parameters "through" to continue iterations until convergence in the parameters of each of the 1000 runs. Moreover, we estimated the optional smoothing value for each replicate tree and ran cross-validation with the options "cv" and "randomcv" (Smith and O'Meara 2012). The divergence times resulting from the 1000 bootstrap trees were used to assess the age variation for each node of interest (Supplemental File 5).

5.4.4 Comparison among PhyloBayes chains and between TreePL age estimates

Although Bayesian parameters did not reach convergence after 100,000 cycles (Supplemental File 5), the estimated ages resulting from our independent four chains were similar when compared to each other (Supplemental File 4). Moreover, our Bayesian and Penalized likelihood approaches showed similar divergence times, strengthening the conclusions of our study. We only observed remarkable discrepancies in Photosynthetic Cyanobacteria (PL showing more recent divergence during the GOE), and the marine Picocyanobacteria *Synechococcus* and *Prochlorococcus* (PL showing more ancient divergence during the POE). Despite these discrepancies, the differences observed between both approaches do not alter the main conclusions of or study

5.4.5 Comparing Bayesian diversification estimates with previous studies

Several diversification estimates shown in our study disagree with previously published analyses. For example, a recent molecular dating estimate suggested that the transition of AOA-Archaea from terrestrial environments into marine reals occurred before the NOE (Ren et al. 2019) during a period known as the "boring million" characterized by low productivity and minimum oxygen concentrations in the atmosphere (0.1% the present levels) (Anbar and Knoll 2002; Holland 2006; Hodgskiss et al. 2019; Reinhard and Planavsky 2022). Our estimates point to a later diversification

of this lineage during or after the NOE (678 Mya, 95% CI = 668-688 Mya) (Fig. 5.2), which is comparable with the age reported by another independent study (Yang et al. 2021). Another study disagreeing with our diversification estimates reported the origin of the Picocyanobacterial clade *Prochlorococcus* to be 0.8 Ga, before the Snowball Earth period registered during the Cryogen (Zhang et al. 2021). However, our ToL sampling agrees with another independent study that points to a relatively late evolution of *Prochlorococcus* (Sánchez-Baracaldo 2015).

5.4.6 Orthologous groups detection, stochastic mapping, and functional annotation

To investigate the genomic novelties associated with the diversification of the marine microbial lineages considered in our study, we identified enriched KEGG categories in the crown node of each clade. First, we predicted protein orthologous groups with ProteinOrtho v6 (Lechner et al. 2011) using the option "lastp". Furthermore, we performed a functional annotation using the KEGG database (Kanehisa and Goto 2000; Kanehisa 2019; Kanehisa et al. 2021) through HMMER3 with an e-value of 10^{-5} on all proteins. Proteins with multiple annotations were filtered to keep the best-scored annotation, and we used the Majority Rule Principle for those clusters that had protein sequences with differing annotated KOs. The presence/absence matrix resulting from the identification of orthologous groups was used together with the phylogenetic tree utilized for molecular dating to perform 100 replicate stochastic mapping analyses on each orthologous group with the SIMMAP program implemented on Phytools (Bollback 2006; Revell 2012). Moreover, we created a null distribution consisting in simulating data for each protein cluster from the Q-matrix predicted from our stochastic mapping analysis but using a constant rate and without presence/absence data in the tips. Since some of the protein clusters show a low exchange rate (identified because one of the rows in the Q-matrix was equal to zero), we manually changed the exchange rate from zero to 0.00001. For each distribution, we estimated the number of genes

gained at the crowd node of the marine clades and classified the individual clusters according to their KEGG category. Clusters without a known annotation on the KEGG database were discarded. The resulting KEGG categories distributions for our stochastic mapping and null analyses were statistically compared using a wilcox test (N= 100 for each distribution). KEGG categories showing statistically more gains in our stochastic mapping distribution were considered enriched.

5.5 Data availability

The supplemental files of this chapter are available of the Figshare collection: https://doi.org/10.6084/m9.figshare.c.6242646.v1

5.6 Acknowledgements

We acknowledge the use of the Virginia Tech Advanced Research Computing Center for bioinformatic analyses performed in this study. This investigation was supported by grants from the Institute for Critical Technology and Applied Science and the National Science Foundation (IIBR-1918271), and a Simons Early Career Award in Marine Microbial Ecology and Evolution to F.O.A. We kindly thank members of the Aylward Lab for their insightful comments on an earlier version of this manuscript.

5.7 References

Acker M, Hogle SL, Berube PM, Hackl T, Coe A, Stepanauskas R, Chisholm SW, Repeta DJ. 2022. Phosphonate production by marine microbes: Exploring new sources and potential function. Proc. Natl. Acad. Sci. U. S. A. 119:e2113386119.

Alcott LJ, Mills BJW, Poulton SW. 2019. Stepwise Earth oxygenation is an inherent property of global biogeochemical cycling. Science 366:1333–1337.

Anbar AD, Duan Y, Lyons TW, Arnold GL, Kendall B, Creaser RA, Kaufman AJ, Gordon GW, Scott C, Garvin J, et al. 2007. A whiff of oxygen before the great oxidation event? Science 317:1903–1906.

Anbar AD, Knoll AH. 2002. Proterozoic ocean chemistry and evolution: a bioinorganic bridge? Science 297:1137–1142.

Anon. 2009. Localized Plasticity in the Streamlined Genomes of Vinyl Chloride Respiring Dehalococcoides.

Bekker A, Holland HD, Wang P-L, Rumble D, Stein HJ, Hannah JL, Coetzee LL, Beukes NJ. 2004. Dating the rise of atmospheric oxygen. Nature [Internet] 427:117–120. Available from: http://dx.doi.org/10.1038/nature02260

Bergman NM, (Tim) Lenton TM, Watson AJ, Dynamics B, Biogeochemistry. 2004. COPSE: A new model of biogeochemical cycling over Phanerozoic time.

Berner RA, Raiswell R. 1983. Burial of organic carbon and pyrite sulfur in sediments over phanerozoic time: a new theory. Geochimica et Cosmochimica Acta [Internet] 47:855–862. Available from: http://dx.doi.org/10.1016/0016-7037(83)90151-5

Biller SJ, Berube PM, Lindell D, Chisholm SW. 2015. Prochlorococcus: the structure and function of collective diversity. Nat. Rev. Microbiol. 13:13–27.

Bollback JP. 2006. SIMMAP: stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics 7:88.

Brasier MD, Lindsay JF. 1998. A billion years of environmental stability and the emergence of eukaryotes: new data from northern Australia. Geology 26:555–558.

Brown MV, Ostrowski M, Grzymski JJ, Lauro FM. 2014. A trait based perspective on the biogeography of common and abundant marine bacterioplankton clades. Mar. Genomics 15:17–28.

Burrows CJ. 2009. Surviving an Oxygen Atmosphere: DNA Damage and Repair. ACS Symp. Ser. Am. Chem. Soc. 2009:147–156.

Butterfield NJ. 2000. Bangiomorpha pubescensn. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiology [Internet] 26:386–404. Available from: http://dx.doi.org/10.1666/0094-8373(2000)026<0386:bpngns>2.0.co;2

Butterfield NJ. 2001. Paleobiology of the late Mesoproterozoic (ca. 1200 Ma) Hunting Formation, Somerset Island, arctic Canada. Precambrian Research [Internet] 111:235–256. Available from: http://dx.doi.org/10.1016/s0301-9268(01)00162-0

Canfield DE, Poulton SW, Narbonne GM. 2007. Late-Neoproterozoic Deep-Ocean Oxygenation and the Rise of Animal Life. Science [Internet] 315:92–95. Available from: http://dx.doi.org/10.1126/science.1135013

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics [Internet]. Available from: http://dx.doi.org/10.1093/bioinformatics/btz848

Clifford EL, Varela MM, De Corte D, Bode A, Ortiz V, Herndl GJ, Sintes E. 2019. Taurine Is a

Major Carbon and Energy Source for Marine Prokaryotes in the North Atlantic Ocean off the Iberian Peninsula. Microb. Ecol. 78:299–312.

Crockford PW, Hayles JA, Bao H, Planavsky NJ, Bekker A, Fralick PW, Halverson GP, Bui TH, Peng Y, Wing BA. 2018. Triple oxygen isotope evidence for limited mid-Proterozoic primary productivity. Nature [Internet] 559:613–616. Available from: http://dx.doi.org/10.1038/s41586-018-0349-y

Croft MT, Lawrence AD, Raux-Deery E, Warren MJ, Smith AG. 2005. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. Nature 438:90–93.

Dahl TW, Arens SKM. 2020. The impacts of land plant evolution on Earth's climate and oxygenation state – An interdisciplinary review. Chemical Geology [Internet] 547:119665. Available from: http://dx.doi.org/10.1016/j.chemgeo.2020.119665

Ducklow HW, Doney SC. 2013. What Is the Metabolic State of the Oligotrophic Ocean? A Debate. Annual Review of Marine Science [Internet] 5:525–533. Available from: http://dx.doi.org/10.1146/annurev-marine-121211-172331

Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput. Biol. 7:e1002195.

Falkowski PG, Barber RT, Smetacek V V. 1998. Biogeochemical Controls and Feedbacks on Ocean Primary Production. Science 281:200–207.

Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science 320:1034–1039.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281:237–240.

Flombaum P, Gallegos JL, Gordillo RA, Rincón J, Zabala LL, Jiao N, Karl DM, Li WKW, Lomas MW, Veneziano D, et al. 2013. Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. Proceedings of the National Academy of Sciences [Internet] 110:9824–9829. Available from: http://dx.doi.org/10.1073/pnas.1307701110

Giovannoni SJ, Stingl U. 2005. Molecular diversity and ecology of microbial plankton. Nature [Internet] 437:343–348. Available from: http://dx.doi.org/10.1038/nature04158

Hatzenpichler R. 2012. Diversity, physiology, and niche differentiation of ammonia-oxidizing archaea. Appl. Environ. Microbiol. 78:7501–7510.

Hewson I, Poretsky RS, Beinart RA, White AE, Shi T, Bench SR, Moisander PH, Paerl RW, Tripp HJ, Montoya JP, et al. 2009. In situ transcriptomic analysis of the globally important keystone N2-fixing taxon Crocosphaera watsonii. ISME J. 3:618–631.

Hodgskiss MSW, Crockford PW, Peng Y, Wing BA, Horner TJ. 2019. A productivity collapse to end Earth's Great Oxidation. Proceedings of the National Academy of Sciences [Internet] 116:17207–17212. Available from: http://dx.doi.org/10.1073/pnas.1900325116

Hoffman PF, Kaufman AJ, Halverson GP, Schrag DP. 1998. A Neoproterozoic Snowball Earth.

Science [Internet] 281:1342–1346. Available from: http://dx.doi.org/10.1126/science.281.5381.1342

Holland HD. 2002. Volcanic gases, black smokers, and the great oxidation event. Geochimica et Cosmochimica Acta [Internet] 66:3811–3826. Available from: http://dx.doi.org/10.1016/s0016-7037(02)00950-x

Holland HD. 2006. The oxygenation of the atmosphere and oceans. Philosophical Transactions of the Royal Society B: Biological Sciences [Internet] 361:903–915. Available from: http://dx.doi.org/10.1098/rstb.2006.1838

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14:587–589.

Kanehisa M. 2019. Toward understanding the origin and evolution of cellular organisms. Protein Sci. 28:1947–1951.

Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. 2021. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 49:D545–D551.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27–30.

Karl DM. 2002. Nutrient dynamics in the deep blue sea. Trends Microbiol. 10:410–418.

Khademian M, Imlay JA. 2021. How Microbes Evolved to Tolerate Oxygen. Trends Microbiol. 29:428–440.

Knoll AH, Javaux EJ, Hewitt D, Cohen P. 2006. Eukaryotic organisms in Proterozoic oceans. Philos. Trans. R. Soc. Lond. B Biol. Sci. 361:1023–1038.

Landry Z, Swan BK, Herndl GJ, Stepanauskas R, Giovannoni SJ. 2017. SAR202 Genomes from the Dark Ocean Predict Pathways for the Oxidation of Recalcitrant Dissolved Organic Matter. MBio [Internet] 8. Available from: http://dx.doi.org/10.1128/mBio.00413-17

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics 12:124.

Lenton TM, Dahl TW, Daines SJ, Mills BJW, Ozaki K, Saltzman MR, Porada P. 2016. Earliest land plants created modern levels of atmospheric oxygen. Proc. Natl. Acad. Sci. U. S. A. 113:9704–9709.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47:W256–W259.

Luo H, Csuros M, Hughes AL, Moran MA. 2013. Evolution of divergent life history strategies in marine alphaproteobacteria. MBio [Internet] 4. Available from: http://dx.doi.org/10.1128/mBio.00373-13

Luo H, Moran MA. 2014. Evolutionary ecology of the marine Roseobacter clade. Microbiol. Mol. Biol. Rev. 78:573–587.

Martinez-Gutierrez CA, Aylward FO. 2021. Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. Mol. Biol. Evol. 38:5514–5527.

Masip L, Veeravalli K, Georgiou G. 2006. The many faces of glutathione in bacteria. Antioxid. Redox Signal. 8:753–762.

Mason OU, Di Meo-Savoie CA, Van Nostrand JD, Zhou J, Fisk MR, Giovannoni SJ. 2009. Prokaryotic diversity, distribution, and insights into their role in biogeochemical cycling in marine basalts. ISME J. 3:231–242.

McParland EL, Alexander H, Johnson WM. 2021. The Osmolyte Ties That Bind: Genomic Insights Into Synthesis and Breakdown of Organic Osmolytes in Marine Microbes. Frontiers in Marine Science [Internet] 8. Available from: http://dx.doi.org/10.3389/fmars.2021.689306

Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30:1188–1195.

Moczydłowska M. 1996. Paleobiology of the neoproterozoic svanbergfjellet formation, spitsbergen. Palaeogeography, Palaeoclimatology, Palaeoecology [Internet] 122:247–248. Available from: http://dx.doi.org/10.1016/0031-0182(96)85042-5

Montoya JP, Holl CM, Zehr JP, Hansen A, Villareal TA, Capone DG. 2004. High rates of N2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. Nature 430:1027–1032.

Mühlenbruch M, Grossart H-P, Eigemann F, Voss M. 2018. Mini-review: Phytoplankton-derived polysaccharides in the marine environment and their interactions with heterotrophic bacteria. Environ. Microbiol. 20:2671–2685.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I. 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 42:D26–D31.

Och LM, Shields-Zhou GA. 2012. The Neoproterozoic oxygenation event: Environmental perturbations and biogeochemical cycling. Earth-Science Reviews [Internet] 110:26–57. Available from: http://dx.doi.org/10.1016/j.earscirev.2011.09.004

Ossa Ossa F, Hofmann A, Spangenberg JE, Poulton SW, Stüeken EE, Schoenberg R, Eickmann B, Wille M, Butler M, Bekker A. 2019. Limited oxygen production in the Mesoarchean ocean. Proc. Natl. Acad. Sci. U. S. A. 116:6647–6652.

Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, Poulton NJ, Burkart MD, La Clair JJ, Chisholm SW, et al. 2019. Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. Cell 179:1623–1635.e11.

Pajares S, Varona-Cordero F, Hernández-Becerril DU. 2020. Spatial Distribution Patterns of Bacterioplankton in the Oxygen Minimum Zone of the Tropical Mexican Pacific. Microb. Ecol. 80:519–536.

Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc. Natl. Acad. Sci. U. S. A. 108:13624–13629.

Planavsky NJ, Crowe SA, Fakhraee M, Beaty B, Reinhard CT, Mills BJW, Holstege C, Konhauser KO. 2021. Evolution of the structure and impact of Earth's biosphere. Nature Reviews Earth & Environment [Internet] 2:123–139. Available from: http://dx.doi.org/10.1038/s43017-020-00116-w

Planavsky NJ, Reinhard CT, Wang X, Thomson D, McGoldrick P, Rainbird RH, Johnson T, Fischer WW, Lyons TW. 2014. Earth history. Low mid-Proterozoic atmospheric oxygen levels and the delayed rise of animals. Science 346:635–638.

Porter SM. 2004. The fossil record of early eukaryotic diversification. The Paleontological Society Papers [Internet] 10:35–50. Available from: http://dx.doi.org/10.1017/s1089332600002321

Reinhard CT, Planavsky NJ. 2022. The History of Ocean Oxygenation. Ann. Rev. Mar. Sci. 14:331–353.

Ren M, Feng X, Huang Y, Wang H, Hu Z, Clingenpeel S, Swan BK, Fonseca MM, Posada D, Stepanauskas R, et al. 2019. Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. ISME J. 13:2150–2161.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution [Internet] 3:217–223. Available from: http://dx.doi.org/10.1111/j.2041-210x.2011.00169.x

Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31:1261–1271.

Sánchez-Baracaldo P. 2015. Origin of marine planktonic cyanobacteria. Sci. Rep. 5:17418.

Sánchez-Baracaldo P, Bianchini G, Di Cesare A, Callieri C, Chrismas NAM. 2019. Insights Into the Evolution of Picocyanobacteria and Phycoerythrin Genes (mpeBA and cpeBA). Frontiers in Microbiology [Internet] 10. Available from: http://dx.doi.org/10.3389/fmicb.2019.00045

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR, Post AF, Hagemann M, Paulsen I, Partensky F. 2009. Ecological Genomics of Marine Picocyanobacteria. Microbiology and Molecular Biology Reviews [Internet] 73:249–299. Available from: http://dx.doi.org/10.1128/mmbr.00035-08

Scott C, Lyons TW, Bekker A, Shen Y, Poulton SW, Chu X, Anbar AD. 2008. Tracing the stepwise oxygenation of the Proterozoic ocean. Nature 452:456–459.

Seymour JR, Amin SA, Raina J-B, Stocker R. 2017. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. Nature Microbiology [Internet]

2. Available from: http://dx.doi.org/10.1038/nmicrobiol.2017.65

Shang H, Rothman DH, Fournier GP. 2022. Oxidative metabolisms catalyzed Earth's oxygenation. Nature Communications [Internet] 13. Available from: http://dx.doi.org/10.1038/s41467-022-28996-0

Sheik CS, Jain S, Dick GJ. 2014. Metabolic flexibility of enigmatic SAR324 revealed through metagenomics and metatranscriptomics. Environmental Microbiology [Internet] 16:304–317. Available from: http://dx.doi.org/10.1111/1462-2920.12165

Shields-Zhou G, Och L. 2011. The case for a Neoproterozoic Oxygenation Event: Geochemical evidence and biological consequences. GSA Today [Internet] 21:4–11. Available from: http://dx.doi.org/10.1130/gsatg102a.1

Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 27:135–145.

Smith SA, O'Meara BC. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics 28:2689–2690.

Sperling EA, Wolock CJ, Morgan AS, Gill BC, Kunzmann M, Halverson GP, Macdonald FA, Knoll AH, Johnston DT. 2015. Statistical analysis of iron geochemical data suggests limited late Proterozoic oxygenation. Nature [Internet] 523:451–454. Available from: http://dx.doi.org/10.1038/nature14589

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Tang D, Shi X, Wang X, Jiang G. 2016. Extremely low oxygen concentration in mid-Proterozoic shallow seawaters. Precambrian Research [Internet] 276:145–157. Available from: http://dx.doi.org/10.1016/j.precamres.2016.02.005

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.

Thrash JC, Cameron Thrash J, Seitz KW, Baker BJ, Temperton B, Gillies LE, Rabalais NN, Henrissat B, Mason OU. Metabolic roles of uncultivated bacterioplankton lineages in the northern Gulf of Mexico "Dead Zone." Available from: http://dx.doi.org/10.1101/095471

de la Torre JR, Christianson LM, Béjà O, Suzuki MT, Karl DM, Heidelberg J, DeLong EF. 2003. Proteorhodopsin genes are distributed among divergent marine bacterial taxa. Proc. Natl. Acad. Sci. U. S. A. 100:12830–12835.

Tostevin R, Mills BJW. 2020. Reconciling proxy records and models of Earth's oxygenation during the Neoproterozoic and Palaeozoic. Interface Focus [Internet] 10:20190137. Available from: http://dx.doi.org/10.1098/rsfs.2019.0137

Ueno Y, Yamada K, Yoshida N, Maruyama S, Isozaki Y. 2006. Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. Nature 440:516–519.

Valley JW, Cavosie AJ, Ushikubo T, Reinhard DA, Lawrence DF, Larson DJ, Clifton PH, Kelly

TF, Wilde SA, Moser DE, et al. 2014. Hadean age for a post-magma-ocean zircon confirmed by atom-probe tomography. Nature Geoscience [Internet] 7:219–223. Available from: http://dx.doi.org/10.1038/ngeo2075

Vidal G, Moczydłowska-Vidal M. 1997. Biodiversity, speciation, and extinction trends of Proterozoic and Cambrian phytoplankton. Paleobiology [Internet] 23:230–246. Available from: http://dx.doi.org/10.1017/s0094837300016808

Vila-Costa M, Simó R, Harada H, Gasol JM, Slezak D, Kiene RP. 2006. Dimethylsulfoniopropionate uptake by marine phytoplankton. Science 314:652–654.

Wei G-Y, Planavsky NJ, He T, Zhang F, Stockey RG, Cole DB, Lin Y-B, Ling H-F. 2021. Global marine redox evolution from the late Neoproterozoic to the early Paleozoic constrained by the integration of Mo and U isotope records. Earth-Science Reviews [Internet] 214:103506. Available from: http://dx.doi.org/10.1016/j.earscirev.2021.103506

Wilson ST, Aylward FO, Ribalet F, Barone B, Casey JR, Connell PE, Eppley JM, Ferrón S, Fitzsimmons JN, Hayes CT, et al. 2017. Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium Crocosphaera. Nat Microbiol 2:17118.

Yang Y, Zhang C, Lenton TM, Yan X, Zhu M, Zhou M, Tao J, Phelps TJ, Cao Z. 2021. The Evolution Pathway of Ammonia-Oxidizing Archaea Shaped by Major Geological Events. Mol. Biol. Evol. 38:3637–3648.

Zehr JP, Kudela RM. 2011. Nitrogen cycle of the open ocean: from genes to ecosystems. Ann. Rev. Mar. Sci. 3:197–225.

Zhang H, Sun Y, Zeng Q, Crowe SA, Luo H. 2021. Snowball Earth, population bottleneck and Prochlorococcus evolution. Proceedings of the Royal Society B: Biological Sciences [Internet] 288. Available from: http://dx.doi.org/10.1098/rspb.2021.1956