OPEN ACCESS

University of BRISTOL

This electronic thesis or dissertation has been downloaded from Explore Bristol Research, http://research-information.bristol.ac.uk

Author:
Coleman, Gareth

Title:
A phylogenomic exploration of early bacterial evolution

General rights
Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

Take down policy
Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact collections-metadata@bristol.ac.uk and include the following information in your message:

• Your contact details
• Bibliographic details for the item, including a URL
• An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.

A phylogenomic exploration of early bacterial evolution

Gareth Andrew Coleman

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Life Sciences

University of Bristol School of Biological Sciences
Life Sciences Building
24 Tyndall Avenue BS8 1TQ

September 2020

Word count: 46,642

Abstract

There are many challenges associated with the reconstruction of early evolutionary history. This is particularly true in the case of Bacteria. Despite being one of the two primary domains of life, and therefore crucial to our understanding of the early history of life, there is little consensus regarding the deepest evolutionary relationships within the bacterial tree. Due to the large spans of time that have elapsed since the origin of the domain, there are many difficulties in modelling their evolution, with bacterial phylogenies frequently affected by artefacts in the analyses. There are therefore a number of questions still unresolved regarding the relationships between major phyla, the root of the tree, and indeed whether the abundant horizontal gene transfer known to characterise prokaryotic evolution has not obscured vertical signal to the point of rendering a tree analogy moot. Recent discoveries of a huge diversity of new uncultured phyla provide new data, but are often difficult to resolve within the bacterial tree, with the relationships between the major bacterial lineages still showing little resolution. Bacteria also represent the most genetically and metabolically diverse organisms on the planet, and as such there are many questions pertaining to the evolution of diverse physiologies and metabolism through time. In this thesis, we attempt to address these issues by using innovative genomic approaches while incorporating much of the previously unknown bacterial diversity. We produce a rooted tree of Bacteria, demonstrate the inadequacies of outgroup rooting, and quantify the contributions of both vertical and horizontal signal to bacterial evolution. We additionally infer the order of events in early bacterial evolution, and reconstruct ancestral metabolisms for the earliest bacterial lineages. Taken together, these results can be integrated to produce a model of early bacterial evolution which contributes to our understanding of the earliest phase of life on Earth.

Author Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Gareth Coleman

11th September 2020

Statement of collaboration

Chapters 2, 3 and 4 are derived from a paper currently under revision in collaboration with Adrián A. Davín, Tara Mahendrarajah, Anja Spang, Philip Hugenholtz, Gergely J. Szöllősi, and Tom A. Williams. Gareth A. Coleman is the first author of this paper. Details of contributions are given on the title page of each chapter.

Coleman, G.A., Davín, A.A., Mahendrarajah, T., Spang, A.A., Hugenholtz, P., Szöllősi, G.J. and Williams, T.A., 2020. A rooted phylogeny resolves early bacterial evolution. *bioRxiv*.

https://www.biorxiv.org/content/10.1101/2020.07.15.205187v1

Chapter 5 has been published as a Coleman *et al.* (2019) in Genome Biology and Evolution in collaboration with Richard D. Pancost and Tom A. Williams. Gareth A. Coleman is the first author of this paper.

Coleman, G.A., Pancost, R.D. and Williams, T.A., 2019. Investigating the origins of membrane phospholipid biosynthesis genes using outgroup-free rooting. *Genome biology and evolution*, *11*(3), pp.883-898.

https://academic.oup.com/gbe/article/11/3/883/5310093

Chapters 1 and 6 are entirely the work of Gareth A. Coleman

Gareth Coleman

11th September 2020

Acknowledgements

At the near completion of my PhD, there are many people to thank for their various contributions, great and small, that ultimately helped me get to this point. First and foremost, I would like to thank my primary supervisor Tom Williams for giving me this position, for being a fantastic supervisor, and for always being with me every step of the way. I would also like to thank my secondary supervisor Rich Pancost for the pastoral support and advice he gave. Additionally, I would like to thank the Royal Society for funding this PhD, and the University of Bristol, including all the admin staff, for hosting and giving access to the resources that made it possible. Many thanks go to my collaborators, Adrián Davín, Tara Mahendrarajah, Anja Spang, Philip Hugenholtz, Gergely Szöllősi, who helped make this research so much better, and allowed a real contribution to science to be made.

I would also like to thank the University of Bristol Palaeobiology Research group, and all its members who helped me in various ways, both scientifically, and as friends. Particular thanks go to my fellow lab group members, Céline Petitjean and Edmund Moody. Additionally, I must thank my various friends, both within and without the university, for giving me all their love and support throughout this time. I would therefore like to thank Emma Landon, Anna Williams, Tom Kettlety, Ellen MacDonald, Nuria Melisa Morales García, Suresh Singh, Rhys Charles, Antonio Ballell Mayoral, Alessia Tasca, Pierre-Aurélian Gilliot, Tom Smith, Ben Griffin, George Atkinson, Rebekah Locke, Alasdair Price, and Chris Weaver, amongst many others who have offered their friendship and support at various times. I also give my thanks and love to my family for their continued support, including my brother Luke Coleman, my sister Amy Cheal, and of course my parents Mick and Chris Coleman, to whom I owe so much.

Lastly, but most importantly, I give my eternal love and gratitude to my husband, Raphaël, without whom I could not have finished this PhD. I wish to thank him for loving and supporting me, as well as shouldering the burden of having to putting up with me, especially during such a fraught time. I could not ask for a better life partner.

« On ne voit bien qu’avec le cœur. L’essentiel est invisible pour les yeux. »

- Antoine de Saint-Exupéry, Le Petit Prince

Écrivain et poète 1900-1944

« Rien dans la vie n'est à craindre. Ce n'est qu'à être compris. »

- Marie Skłodowska-Curie

Physicienne et scientifique 1867-1934

Table of Contents

Title page **1**
Abstract **3**
Author Declaration **5**
Statement of collaboration **7**
Acknowledgements **9**
Table of contents **13**
List of figures **16**
List of tables **17**

Chapter 1
Bacteria and the challenges of reconstructing early evolution **20**
1.1 The challenges of deep-time phylogenetics 22
1.2 Approaching deep-time evolution using whole genomes 25
1.3 The case of Bacteria 27
1.4 A rooted tree of Bacteria 28
1.5 Evolution of core metabolism in Bacteria 31
1.6 One membrane or two? The evolution of the cell envelope 35
1.7 Timing of bacterial evolution 38
1.8 The Lipid divide 40
1.9 A model for the evolution of early life 45

Chapter 2
Phylogenomics produces a rooted tree of Bacteria **47**
2.1 Introduction 49
2.2 Methods 51
2.3 Results and discussion 59
2.4 Conclusion 89

Chapter 3
Ancestral reconstruction of the last bacterial common ancestor **91**
3.1 Introduction 93

3.2 Methods 95
3.3 Results and discussion 99
3.4 Conclusion 115

Chapter 4

Genomic evolution of Bacteria through time 119

4.1 Introduction 121
4.2 Methods 123
4.3 Results and discussion 124
4.4 Conclusion 143

Chapter 5

Investigating the Origins of Membrane Phospholipid Biosynthesis Genes Using Outgroup-Free Rooting 146

5.1 Introduction 148
5.2 Methods 151
5.3 Results and discussion 154
5.4 Conclusion 175

Chapter 6

Toward an integrated model of early bacterial evolution 178

6.1 Addressing the challenges of deep-time phylogenetics 180
6.2 The root of the bacterial tree between two large and diverse clades 182
6.3 An acetogenic origin of Bacteria? 183
6.4 The early origin of motile diderm cells 185
6.5 Diversification of Bacteria through time 186
6.6 The Lipid divide 186
6.7 What can we say about the last universal common ancestor? 188
6.8 Future directions 189

References **192**

Appendices

A. Species tables for Chapter 2 221

B. Full heatmap 237

C. Supplementary tree figures for Chapter 5 (phospholipids) 238

D. Papers derived from work included in this thesis 295

List of figures

1.1 Schematic bacterial tree with various proposed roots indicated.
1.2 Evolutionary hypothesis for the origin of the outer membrane.
1.3 Phospholipid biosynthesis pathways in Archaea and Bacteria.
1.4 Different models of the origin and early evolution of phospholipid biosynthesis in Archaea and Bacteria.

2.1 Unrooted bacterial phylogenies inferred from the GTDB dataset.
2.2 Unrooted bacterial phylogenies inferred from the GTDB-independent dataset.
2.3 Outgroup-rooted bacterial phylogeny inferred from the GTDB dataset.
2.4 Outgroup-rooted bacterial phylogenies inferred from the GTDB-independent dataset.
2.5 RMC and MAD rooted bacterial phylogenies derived from the GTDB-independent dataset.
2.6 Accuracy of gene rooting methods.
2.7 Root positions determined by ALE for both the GTDB and the GTDB-independent datasets.
2.8 Verticality of bacterial evolution.
2.9 Relationship between verticality and gene family size.
2.10 Rooted phylogeny of Archaea and Bacteria

3.1 Components of the flagellum and chemotaxis inferred in LBCA.
3.2 Distribution of COG families from key metabolic pathways inferred in LBCA.
3.3 Metabolic map of central metabolic pathways inferred in LBCA.
3.4 Ancestral reconstruction of LBCA.

4.1 Nodes for which ancestral gene content were inferred.
4.2 Relative ages of bacterial clades.
4.3 Evolution of COG family repertoires and inferred genomes size of the bacterial tree.
4.4 Metabolic map of central carbohydrate pathways in surveyed nodes.
4.5 Metabolic map of acetogenesis and the WLP in surveyed nodes.

4.6 Components of the flagellum in surveyed nodes.

5.1 Biosynthesis pathway and composition of phospholipids in Bacteria and Archaea.
5.2 Bayesian consensus tree of archaeal enzymes.
5.3 Bayesian consensus tree of both G3PDH enzymes.
5.4 Bayesian consensus tree of GlpK, PlsC and PlsY enzymes.

List of Tables

2.1 Number of taxa sampled from each clade in the GTDB-independent analysis.
2.2 63 orthologues used to infer the species, with those used in the outgroup rooting analysis indicated.
2.3 Mean verticality by COG functional category.
2.4 Support for published hypotheses using outgroup rooting.
2.5 Support for published rooting hypotheses form our ALE analysis.
2.6 AU-test results for an ALE root analysis using 3595 COG families**.**
2.7 Singleton support on the credible set of rooted trees.

3.1 Estimated root origination rates and root presences by COG functional category.
3.2 PPs for presence of glycerolipids in LBCA.
3.3 PPs for presence of lipopolysaccharide genes in LBCA.

4.1 PPs for presence of glycerolipids in LBCA in nodes surveyed.
4.2 PPs for presence of lipopolysaccharide genes in nodes surveyed.

5.1 Distribution of phospholipid biosynthesis genes in bacteria and archaeal phyla.
5.2 PPs for RMC and MAD rooting methods.
5.3 AI scores for MAD roots.
5.4 BIC scores for outgroup rooted trees under IQ-Tree model selection.

List of Supplementary Tables

Supplementary Table 1 Protein family annotations (COG and KO) and root presence posterior probabilities (PPs) for all 3723 gene families under all three branches in the root region, Chapters 3 and 4 (Excel-formatted spreadsheet).

Supplementary Table 2 Protein family annotations (COG and KO) and root presence posterior probabilities (PPs) for key pathways used in reconstruction, Chapters 3 and 4 (Excel-formatted spreadsheet).

Supplementary Table 3 Table containing the results from the LOESS regression analysis of COG family members against genome size, Chapter 4 (tsv file).

Supplementary Table 4 COG families lost on the CPR stem, Chapter 4 (Excel-formatted spreadsheet).

Supplementary Table 5 Accession numbers for sequences used in phylogenetic analyses in Chapter 5 (Excel-formatted spreadsheet)..

Chapter 1

Bacteria and the challenges of reconstructing early evolution

This chapter is not part of any publication and has been written by GAC entirely.

Abstract

Reconstructing deep evolutionary history presents many challenges. This is especially evident in the case of Bacteria. Despite being one of the two primary domains of life, there has been little consensus on the deepest evolutionary relationships in the bacterial tree, especially the position of the root. A number of issues have impeded these endeavours, including difficulties in modelling such long stretches of evolutionary time, the use of inappropriate models and methods, problems with topological artefacts, and ultimately whether tree-like analogies are applicable to bacterial evolution at all. Understanding bacterial phylogeny is also necessary if we wish to understand the evolution of bacterial cells, metabolisms, and other traits. As Bacteria represent the most genetically and metabolically diverse lifeforms on the planet, there are a number of questions regarding the evolution of different metabolic pathways, physiologies and morphological characteristics. With the advent of new sequencing technologies, our knowledge of bacterial diversity has greatly expanded. This offers a wealth of new and important data, but also difficulties in how to integrate this data into our current frameworks of bacterial evolution.

1.1 The challenges of deep-time phylogenetics

Throughout human history we have attempted to classify the environment around us, including how various other life forms with which we share our planet fit into our concept of the wider world and our own place within it. Whether bound by religious dogma, or Enlightenment ideas about the continual march to perfection, most schemes concerning the natural world involved a classification of organisms in a progression of ever greater complexity, ending with humans at the pinnacle of the evolutionary scale. However, the publication of Charles Darwin's On the Origin of Species in 1859 saw a paradigm shift towards thinking about life not as a scale, but as a tree of interrelated organisms shaped by natural selection and evolution by common descent. The discovery of inheritable elements or "genes" (Mendel, 1866) and the molecule which could carry this information, DNA (Miescher, 1869; Miescher-Rüsch, 1871; Avery Oswald, Colin and MacLeod, 1944; Franklin and Gosling, 1953a, 1953b; Watson and Crick, 1953) gave a tangible mechanism to how evolution through descent can actually work, and thus evolutionary biology shifted from simply classifying things into groups, to trying to understand how different organisms were related to each other via their shared evolutionary history. With the development of gene sequencing techniques and advent of the computer age, we are now able to analyse genetic sequences and extract the evolutionary signal they hold within. As computers have become more powerful, and the amount of data ever expanding, we have been afforded the opportunity to resolve some of the deepest and most fundamental questions within evolutionary biology. However, we face a number of challenges in this endeavour.

Problems with evolutionary models

One of the primary issues within phylogenetics is selecting models that best describe the evolutionary process (Ripplinger and Sullivan, 2008; Hoff *et al.*, 2016). This is especially apparent with deep time phylogenies, where the process of evolution has continued for such extraordinary lengths of time that evolutionary signal is in danger of being overwritten and lost (Penny *et al.*, 2001; Gascuel, 2005; White *et al.*, 2007). All models which attempt to describe the evolutionary process are necessarily simplistic abstractions of what actually occurs, and therefore our reconstructions of the

evolutionary past will always be imprecise and lacking in resolution. Nonetheless, the use of poorly fitting or misspecified models may lead to less accurate results (Ripplinger and Sullivan, 2008; Hoff *et al.*, 2016; Naser-Khdour *et al.*, 2019), and therefore investing time in the understanding and development of better models is of great importance. Such a need has led the development from simple models, where frequencies of base-pairs or amino acids and the substitution of one for another have equal probability, as in the Juke-Cantor Model (Jukes, Cantor and Others, 1969), to more complex models which allow these probabilities to be unequal, such as the General Time Reversible (GTR) (Tavaré, 1986) and Le and Gascuel (LG) (Le and Gascuel, 2008) models. However, not all sites evolve at the same rate, with some evolving much faster than others. More simplistic models which model all sites homogeneously will be very susceptible to topological artefacts (Foster and Hickey, 1999; Foster, 2004). Two major areas of concern are the impact of composition-driven long branch attraction (LBA), and taxon sampling. LBA occurs when lineages with high substitution rates (that is, high rates of evolution) appear similar to each other due to convergence, causing the analysis to erroneously infer a close relatiohsip between these taxa and therefore “attracting” them to each other in the tree (Felsenstein, 1978; Lartillot, Brinkmann and Philippe, 2007a). A common example of this is when long branches are attracted to the base of the tree, often due to long-branching basal lineages or an outgroup separated by a long branch (discussed further below). Related to this, depending on the average branch lengths of taxa in a given dataset, changing the taxon sampling can further lead to such artefacts. Low taxonomic sampling is particularly susceptible to this, and improved sampling can help to resolve difficult phylogenetic problems (Graybeal, 1998; Hedtke, Townsend and Hillis, 2006). If increased taxonomic sampling is impractical, using multiple independent datasets may give insight into whether taxon sampling is causing artefacts or lack of resolution in the phylogeny. More complex models can account for among site variation using site-specific composition profiles, and thus produce more accurate phylogenies which may circumnavigate these topological artefacts (Le, Lartillot and Gascuel, 2008). Ultimately no model will truly describe the evolutionary process with complete accuracy, but careful selection of appropriate models, or extensive model testing where practical, will go some way to resolving issues in our phylogenies, or at least reducing very obvious errors and biases.

ward c wheeler long branch attraction

random outgourp will attach to the longest branch - largest target.

The issues of rooting deep phylogenies

Another issue in deep time phylogenetics lies in attempting to determine roots within phylogenetic trees. The root of a phylogenetic tree represents the first split in that tree, and thus the node at a given root represents the last common ancestor of the group in question. The standard approach to rooting phylogenies is to include an outgroup, i.e. a closely related organism that does not belong to the group under study, the ingroup (Penny, 1976). A tree is inferred with this outgroup, and the root placed on the branch leading to it. The resulting branch order within the ingroup gives us the position of its root. Several problems may arise when attempting to use outgroups. First, the choice of outgroup requires some prior phylogenetic knowledge about the placement of the outgroup with the relation to the ingroup. Specifically, there must be confidence that the outgroup is truly an outgroup and not actually part of the ingroup, while still being closely related enough to be phylogenetically informative. Second, if the outgroup is too distant, this may further exaggerate LBA artefacts and distort ingroup relationships (Gouy, Baurain and Philippe, 2015). Such is the case in many parts of the tree of life where the nearest outgroup to a clade is separated by a long branch. Third, analysing both the ingroup and the outgroup may reduce the number of genes that are conserved between the two groups, and therefore reduce the amount of data that can be used for tree inference. This will be especially true of clades with distant outgroups. Fourth, in the case of the entire tree of life, there is no outgroup, rendering outgroup rooting impossible. Alternatives to outgroup rooting have been employed, such as the relaxed molecular clock (Thorne, Kishino and Painter, 1998; Kishino, Thorne and Bruno, 2001), and the recently described MAD rooting method of (Tria, Landan and Dagan, 2017). The MAD algorithm finds the root position that minimises pairwise evolutionary rate variation, averaged over all pairs of taxa in the tree. However, both methods may be sensitive to both composition-driven LBA and to taxon sampling (Chapter 2 of this thesis).

Vertical or horizontal? The transmission of genetic information

A further major issue concerning phylogenetics concerns the type of genetic transmission and how this affects our modelling of evolution. Traditionally in

phylogenetics, evolution has been assumed to be dominated by vertical transmission of genetic information, with a bifurcating tree describing the majority of evolutionary relationships. The underlying assumption therefore is that there is ultimately a “true tree” which explicitly describes all evolutionary relationships. As part of this assumption, the use of concatenations to build phylogenies is common, where multiple gene alignments are appended together and modelled as a single gene, under the assumption that they all follow an underlying species tree. While such approaches are most likely appropriate for certain parts of the tree of life (evolution of animals for example), this may not be the case for others. For example, it is known that horizontal gene transfer (HGT) is extensive in prokaryotes (Ochman, Lawrence and Groisman, 2000; Koonin, Makarova and Aravind, 2002; Heuer and Smalla, 2007). Previously published analyses have indicated that the vast majority, if not all prokaryotic gene families have undergone HGT to some extent during their evolutionary history (Dagan and Martin, 2007; Williams *et al.*, 2017), implying that no single tree fully describes the evolution of all bacterial genes or genomes (Doolittle, 1999; Doolittle and Bapteste, 2007). This presents a problem to using concatenation as it reduces the number of genes that evolve on a single species tree and therefore reduces the number of genes available for use (Dagan and Martin, 2007). Alternatives to traditional tree construction methods have been used, including phylogenetic networks (Doolittle and Bapteste, 2007; Alvarez-Ponce *et al.*, 2013), which were the first methods to explicitly acknowledge non-vertical evolution. However, networks can be difficult to integrate with vertical data and can be difficult to interpret biologically. It is not clear how extensive horizontal transmission is compared to vertical transmission, with vertical inheritance still likely being an important part of evolutionary history. Being able to coherently model both vertical and horizontal signal in the data is therefore very important when attempting to understand and reconstruct the history of life.

1.2 Approaching deep-time evolution using whole genomes

To address the problems of rooting and prevalence of HGT, we may turn to whole-genome approaches. Such approaches initially began with attempts to root the tree of life, for which no outgroup exists, using gene duplications (Iwabe *et al.*, 1989; J. P.

Gogarten *et al.*, 1989; Brown and Doolittle, 1995). If a gene conserved across all life had a duplication before the last universal common ancestor (LUCA), and has copies preserved in modern taxa, each copy can reciprocally root the other. These methods were developed to include not just gene duplications, but also gene gains, losses and HGTs (Csurös, 2010; Abby *et al.*, 2012). Subsequent further development of methods augmented these models of gene duplication, transfer and loss (DTLs) with information from gene tree topologies (Abby *et al.*, 2012; Bansal, Alm and Kellis, 2012; Lafond, Swenson and El-Mabrouk, 2012; Szöllősi, Boussau and Abby, 2012; Szöllősi *et al.*, 2013; Szöllősi, Davín, *et al.*, 2015; Jacox *et al.*, 2016; Noutahi *et al.*, 2016; de Oliveira Martins and Posada, 2017; Comte *et al.*, 2019; Zwaenepoel and Van de Peer, 2019). The development of such probabilistic gene tree-species tree reconciliation methods allows us to calculate the joint likelihood of a reconciled gene family tree and species tree and rates of DTLs. Ideally, DTLs, rooted gene trees and a rooted species tree would be jointly modelled, but as this is not currently tractable, it is necessary to use a two step approach where we infer unrooted gene trees with a species tree-unaware model, and use the gene tree topologies for the reconciliation analyses. A potential problem with such approaches is that it relies on gene trees which may be poorly resolved, and therefore negatively affect the analysis. However, methods such as Amalgamated Likelihood Estimation (ALE), innovates on previous reconciliation methods by incorporating uncertainty in the gene trees using conditional clade probabilities to down-weight poorly resolved regions of the gene trees so they do not unduly affect the analysis (Szöllősi *et al.*, 2013). These whole-genome approaches improved on other rooting methods by incorporating a much larger amount of data, namely whole genomes as opposed to a small selection of conserved orthologues.

One application of these gene family likelihoods is as a measure to compare support for different rooted species trees. Each competing species tree topology chosen implies a particular evolutionary history for each gene family regarding transfer, loss or gain of genes, which can be compared using statistical tree selection tests such as an Approximately Unbiased (AU) test (Shimodaira, 2002). These likelihoods can be summed for each candidate rooted phylogeny, and compared to determine the likelihood of our gene trees given a candidate rooted phylogeny and our model of DTLs. As this method models the histories of the genes over the tree with regards to duplications, transfers and losses, it models both vertical and horizontal transmission

of genes. We can therefore also estimate rates of HGT over the tree. Furthermore, it can count the proportion of sampled reconciliations in which a given gene family is present in a given node, from which a probability of the presence of that gene can be calculated. This allows us to predict the gene content, and therefore reconstruct the metabolic capabilities, or any given node in the tree.

1.3 The case of Bacteria

Bacteria are one of the two primary domains of life and represent the most abundant and metabolically diverse cellular life forms. They inhabit almost all known habitats and ecosystems, and have evolved a staggering array of physiologies in order to adapt to such diverse environments. They have a profound effect on the environment around us and perform vital roles in many biogeochemical cycles. In recent years, our knowledge of bacterial diversity has greatly expanded due to the development of techniques for sequencing microbes directly from environmental samples, without the need for laboratory cultivation (Hug *et al.*, 2016; Mukherjee *et al.*, 2017; Parks *et al.*, 2017, 2018). Almost all bacterial phyla have seen an increase in what was previously hidden diversity, and many entirely new lineages and phyla have also been identified. Notably, this includes a large radiation of previous completely unknown phyla, known as the Candidate Phyla Radiation (CPR, also known as the Patesciacteria (Brown *et al.*, 2015; Hug *et al.*, 2016; Castelle and Banfield, 2018; Zhu *et al.*, 2019)). The CPR comprises lineages that are characterised by small cells and genomes and are suggested to have predominantly symbiotic or parasitic lifestyles, although little is still known about their ecology and physiology (Brown *et al.*, 2015; Castelle and Banfield, 2018; Castelle *et al.*, 2018; Beam *et al.*, 2020).

While the great expansion in known bacterial diversity has greatly increased our understanding about microbial evolution, integrating this new information into testing hypotheses about the evolution and history of Bacteria has been challenging. Due to this diversity and their ancient and long evolutionary history, phylogenetic analyses of Bacteria are highly susceptible to the challenges discussed above, including issues with LBA, difficulty in determining the root, and extensive HGT. Thus, there are many

fundamental questions about the nature of bacterial evolution which have yet to be answered and which are important in our understanding of the evolution of the early Earth. In this thesis, we use phylogenetic and whole genome approaches discussed above to answer the following questions regarding early prokaryotic evolution, which will be further discussed in the following sections:

1. Can bacterial evolution be described as tree-like and if so, where does the root lie?
2. How has core metabolism evolved over the course of bacterial evolution, and what metabolism was present in the last bacterial common ancestor (LBCA)?
3. How did the bacterial cell envelope evolve?
4. What can we say about the timing of bacterial diversification?
5. How have phospholipid membranes evolved across the tree of life?

1.4 A rooted tree of Bacteria

Due to the problems with rooting deep radiations, including attempting to model deep evolutionary change accurately and circumnavigating LBA artefacts, there is no consensus on where the root of the bacterial tree lies. A number of hypotheses have been advanced (Fig. 1.1). Many early attempts to root the bacterial tree have used Archaea as an outgroup, based on evidence that the root of all life lies between the two domains (Iwabe *et al.*, 1989; J. P. Gogarten *et al.*, 1989; Brown and Doolittle, 1995; Zhaxybayeva, Lapierre and Gogarten, 2005). Many of these proposed root positions place the thermophilic bacteria Aquificota and Themotogota at the base of the tree (Bocchetta *et al.*, 2000; Bern and Goldberg, 2005; Barion *et al.*, 2007; Battistuzzi and Hedges, 2009) (Fig 1.1). The basal placement of thermophiles would imply a thermophilic ancestral bacterium, and therefore has important implications for early prokaryotic evolution. Other analyses using archaeal outgroups found mesophilic Planctomycetes at the base of the tree (Brochier and Philippe, 2002) (Fig. 1.1). However, these are potentially susceptible to LBA due to the distant archaeal outgroup, as described above. Alternative approaches which avoid the use of an outgroup have also been employed, using gene flows and polarisation of changes in multimeric protein complexes and other complex characters to root the tree. Cavalier-

Smith used such approaches in his “transition analysis”, which takes various cellular, molecular and biochemical characters in order to polarise major transitions and systematically exclude lineages with derived characters, to suggest a root between Chloroflexota and all other life, with Archaea and Eukarya branching from within monoderm Bacteria (Cavalier-Smith, 2006) (Fig. 1.1). Lake *et al.* (2009) used analyses of insertions and deletions (indels) within genomes to root the tree within monoderm bacteria, with this root also representing the root of all life (Fig. 1.1).

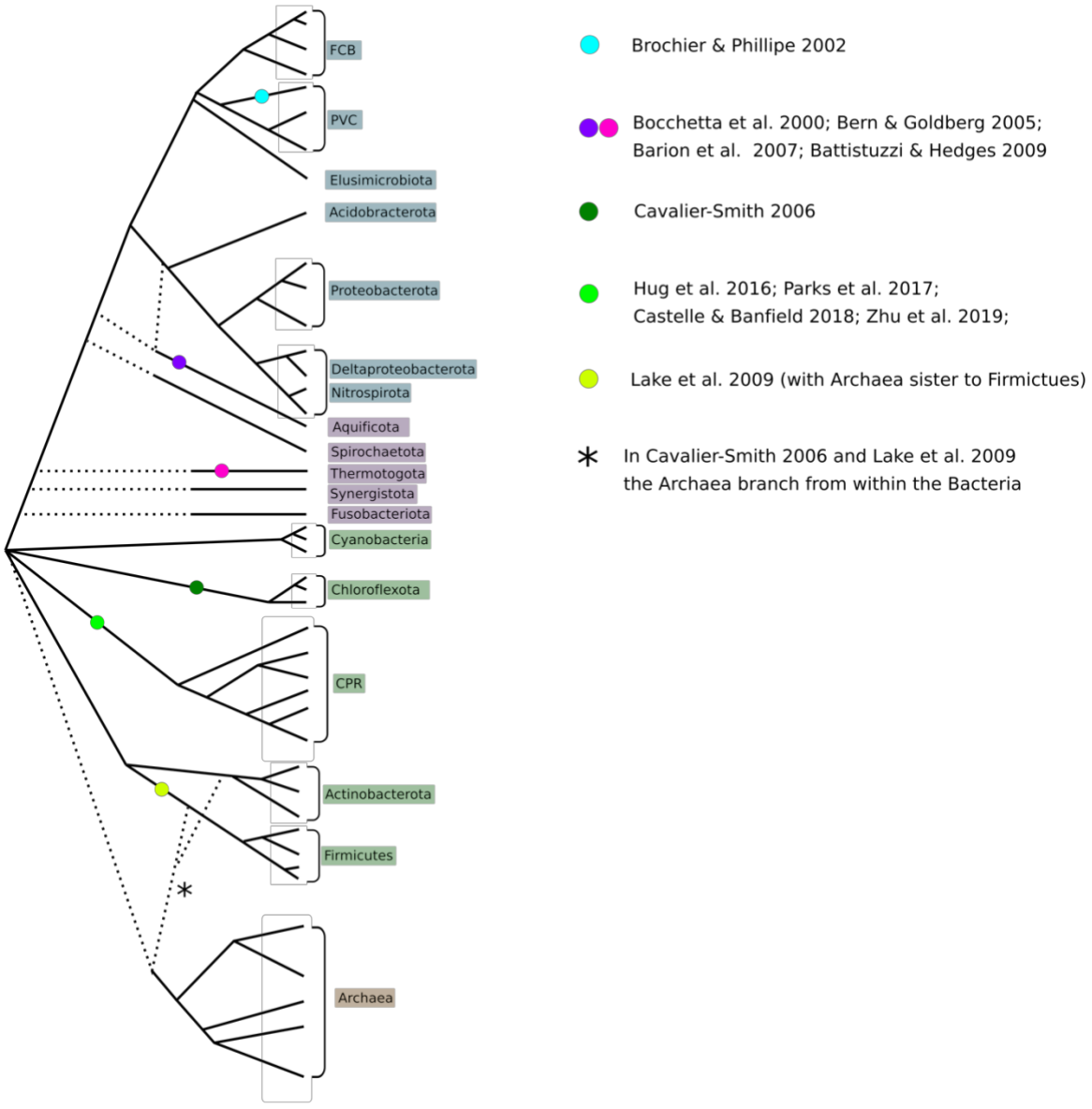


Fig. 1.1 *Schematic representing the bacterial tree, with various proposed root positions indicated. Note that in the trees of Cavalier Smith and Lake et al. the Archaea are a sister to Actinobacteriota and Firmicutes respectively. Lake et al. also describe their root as a "ring of life" (see in text below).*

Recent phylogenetic analyses of the whole tree of life, which incorporate the greatly expanded knowledge of microbial diversity, have place the bacterial root between CPR and all other Bacteria (Hug *et al.*, 2016; Castelle and Banfield, 2018; Zhu *et al.*, 2019) (Fig. 1.1). Given the reduced genomes and likely symbiotic nature of CPR, such early divergence of the clade would have important implications for our understanding of early prokaryote evolution. The DPANN superphylum is an archaeal clade analogous to CPR, and recent analyses suggest that the root of Archaea falls between this clade and other Archaea (Castelle *et al.*, 2015; Williams *et al.*, 2017). If both these root positions are correct, it would imply the presence of symbiotic, highly reduced prokaryotic life alongside more conventional prokaryotic cells even at the earliest stages of evolutionary history. Resolving the position of the root within Bacteria is therefore imperative if we wish to understand the nature of the earliest life and how it subsequently evolved.

These discussions on proposed root positions rely on the existence of some detectable tree-like structure. As discussed above, HTGs are common across prokaryotes, and it has been argued that thinking of early bacterial evolution in terms of a bifurcating species tree may be misleading. Indeed, Lake *et al.* (2009) suggested that his rooted tree, where roots were successively rejected based on the grouping of indels, only made sense when represented as a "ring of life", as many of the genomic relationships could not be adequately described by, or were incompatible with a tree diagram. Additionally, HGT has clearly had a profound effect on prokaryotic evolution. For example, it has been suggested the origin of many major clades within Archaea were driven by transfers from Bacteria, although transfers were less prevalent from Archaea to Bacteria (Nelson-Sathi *et al.*, 2015). Gene tree-species trees reconciliation methods outlined above integrate both tree and network based approaches by modelling both the vertical and horizontal components of genomes evolution, allowing us to measure the contribution of both to bacterial evolutionary history. To do this, we must quantify the amount of vertical evolution with the tree, i.e. the proportion of gene families which evolve vertically. Quantifying verticality will thus allow us to evaluate

how prevalent HGT has been in bacterial evolutionary history, and may give insights into the origins and drivers of innovation and adaptation in bacterial genomic evolution.

In chapter 2, we present a new rooted tree of Bacteria using ALE. We demonstrate that other rooting methods, especially outgroup rooting, are not robust and are susceptible to both composition-driven LBA and taxon sampling. In addition, we attempt to quantify the extent of HGT through the bacterial tree to determine the extent to which bacterial evolution can be described as tree-like.

1.5 Evolution of core metabolism in Bacteria

To fully understand the early evolution of life and the role it has played in shaping the environment around us, we must understand the physiology and metabolic capabilities of the earliest cells. Relatively little work has been done in reconstructing the ancestral metabolism of Bacteria, partly due to the complications with unclear phylogeny and rooting. Furthermore, it is difficult to disentangle such discussions from those concerning the metabolism and habitat of LUCA, depending on how distant LBCA is thought to be from LUCA and whether either resembles modern cells, or were both primitive proto-cells.

Possible paths to carbon fixation

Many scenarios concerning the early evolution of life posit that early prokaryotes would have been autotrophic, and therefore there are key questions regarding which carbon fixation pathway, electron donors and electron acceptors were used by LBCA. Decker *et al.* (1970) used comparative biochemistry to suggest that methanogenesis and acetogenesis were the oldest forms of energy metabolism in extant microbes. Both methanogens and acetogens are anaerobes without cytochromes and obtain organic carbon via the reduction of carbon dioxide by hydrogen, both gases thought to be abundant on the early Earth (Arndt and Nisbet, 2012). Evidence for ancient origins of methanogenesis have been seen in the geological record, showing biological methane production extending back to at least 3.4 Ga (Ueno *et al.*, 2006). Geological reactions that bear a striking resemblance to core metabolic reactions of methanogens are found

to occur spontaneously at hydrothermal vents (Lang *et al.*, 2010; Schrenk, Brazelton and Lang, 2013), in particular, the generation of methane by serpentinisation. The discovery of electron bifurcation (Li *et al.*, 2008), a mechanism of energy conservation, provides a mechanism for both acetogens and methanogens to reduce carbon dioxide with elections from hydrogen despite the initial part of the reaction being energetically uphill (Buckel and Thauer, 2013), and further points to ancient carbon fixation, and the ancient evolution of autotrophy. Methanogens and acetogens both use the Wood-Ljungdahl Pathway (WLP), which has been suggested as the most ancient carbon fixation pathway (Fuchs, 2011; Sousa and Martin, 2014; Williams *et al.*, 2017; Adam, Borrel and Gribaldo, 2018) and previous phylogenetic work has suggested its presence in both the archaeal (Williams *et al.*, 2017; Adam, Borrel and Gribaldo, 2018) and bacterial (Adam, Borrel and Gribaldo, 2018) common ancestors. In the WLP, carbon dioxide is sequentially reduced by hydrogen to methane and acetate respectively in methanogens and acetogens (Ferry and House, 2006; Lane and Martin, 2012; Liu, Beer and Whitman, 2012). The pathway can be divided into two stages, the methyl synthesis stage, and the acetyl synthesis stage. While superficially similar in both groups, different pterin cofactors for methyl synthesis are used. Tetrahydrofolate (H_4F) is used in acetogens and methanopterin tetrahydromethanopterin (H_4MPT) is used in methanogens (Escalante-Semerena, Rinehart and Wolfe, 1984; Jones, Donnelly and Wolfe, 1985; Maden, 2000). The methyl synthesis pathways of both groups also use differing, non-homologous enzymes. However, the key enzyme complex of the pathway, CODH/ACS (CO dehydrogenase/acetyl-CoA synthase), is conserved in both domains, and is predicted to have been present in both the archaeal (Williams *et al.*, 2017; Adam, Borrel and Gribaldo, 2018) and bacterial common ancestors (Adam, Borrel and Gribaldo, 2018). If the WLP were present in LUCA, it has been suggested the methyl branch would have been provided by geochemistry via serpentinisation, while the carbonyl branch would have been performed by CODH/ACS. The enzymes for the methyl pathway would have subsequently evolved in Bacteria and Archaea respectively as they diverged into independent lineages (Martin and Russell, 2003, 2007; Sousa *et al.*, 2013; Sousa and Martin, 2014; Adam, Borrel and Gribaldo, 2018).

Alternatives pathways utilising the WLP have also been suggested. For example, the earliest prokaryotic lineages may have had a denitrifying methanotrophic WLP, with

methanogenesis arising late and independently from acetogenesis (Nitschke and Russell, 2013). However, this hypothesis has a number of problems. The late evolution of methanogens is not compatible with studies of deep phylogeny or other evidence of early biological methanogenesis (Ueno *et al.*, 2006; Martin and Russell, 2007). It is also the case that the denitrifying methanotrophy model must take place under oxidising conditions in the oceans (Sousa *et al.*, 2013), but that under even very mildly oxidising settings, the accumulation at the vent-ocean interface of reduced organic compounds ceases to be thermodynamically favourable (McCollom and Amend, 2005). Furthermore, biological methanogenesis also has a geochemical homologue observed at hydrothermal vents, namely the formation of methane (among other organic compounds) in serpentising systems (Proskurowski *et al.*, 2008; Lang *et al.*, 2010; Etiope, Schoell and Hosgörmez, 2011). However, despite the oxic atmosphere of the present, the geochemical methane oxidation (required for Nitschke and Russell's model) has not been observed.

Sulphate reduction is another possible alternative to methanogenesis. Modern sulphate reducing bacteria respire sulphate to sulphide in a reaction which takes place in two steps. The first step is the reduction of sulphate to sulphite which requires energy, and a second step reduces sulphite to sulphide, where energy is released via a simple respiratory chain. This second part of the process is important as it requires no energy, and sulphite is thought to have been in abundant supply on the early Earth, formed by the reaction of SO_2 from volcanoes, with water. Many modern autotrophic sulphur-reducing bacteria also have the WLP for carbon fixation (Rabus, Hansen and Widdel, 2006). There is geological evidence for the early appearance of this metabolic pathway, with stable isotopes supporting the origin of sulphate respiration as early as 3.47 Ga (Shen, Buick and Canfield, 2001). This, along with the abundant supply of sulphate on the early Earth makes the early appearance of this metabolism very plausible. Further evidence comes from the enzyme dissimilatory sulphite reductase (Dsr), which seems to be highly conserved across many disparate prokaryotic lineages, suggesting an ancient origin of this pathway (Wagner *et al.*, 1998). The trees generated from Dsr were congruent with the 16S rRNA phylogeny of the tree of life, and which were taken as evidence of vertical inheritance rather than horizontal gene transfer (Wagner *et al.*, 1998). Others (Klein *et al.*, 2001) found the gene tree to not be fully compatible with the 16S rRNA tree and therefore inferred horizontal gene

transfer, and others still suggesting both vertical descent and horizontal transfer in different groups (Zverlov *et al.*, 2005). Therefore, it is not completely clear whether sulphate reduction was present in the earliest life, but the evidence points to this as an intriguing possibility.

Alternative carbon fixation pathways to the WLP have also been posited. The reverse TCA cycle has been suggested as a possible ancient carbon fixation pathway (Wächtershäuser, 1990; Cody *et al.*, 2001; Smith and Morowitz, 2004; Nunoura *et al.*, 2018), given the widespread presence of the TCA cycle in modern Bacteria, and that it may function in both the oxidative and reductive direction. Based on a basal position of the Aquificae, and using biomimetic analysis, Marakushev and Belonogova inferred a free-living, chemoautotrophic bacterial ancestor, with an 'archaic metabolic network' coupling reductive tricarboxylic acid, oxidative tricarboxylic acid and 3-hydroxypropionic cycles (Marakushev and Belonogova, 2011, 2013). Braakman and Smith (2012) suggested a combined system of the WLP and reductive tricarboxylic acid cycle in both LBCA and LUCA.

Generation of energy

In addition to specific metabolic pathways, the evolution of energy production and ion pumping is an essential step in the evolution of physiological capabilities of modern Bacteria. Herrmann *et al.* (2008) have suggested that the reduced ferredoxin whose FeS cluster acts as an "energised coupler" in methanogenesis, has energy currency characteristics more primitive than those of ATP. The origin of chemical osmotic coupling, which was hitherto seen as an impossibly large leap in complexity, may have developed from naturally occurring proton gradients at alkaline hydrothermal vents (Russell *et al.*, 1994; Russell and Hall, 1997). However, how was this naturally occurring geochemical ion gradient replaced by ion pumping in order for life to become independent from geochemical ion gradients, and have the ability to produce their own ion gradients across membranes? In Archaea, MtrA-H complex found in methanogens is a potential candidate for ancestral pumping systems, with the Rnf complex in acetogens a similar candidate for Bacteria (Sousa *et al.*, 2013; Sousa and Martin, 2014). In methanogens, the MtrA-H complex pumps out sodium, whilst transferring the methyl group from methyl-H_4MPT to methyl-CoM, whilst in acetogens Rnf pumps out sodium whilst taking electrons from reduced ferredoxin to reduce NAD^+ (Thauer *et al.*,

2008; Lane and Martin, 2012). The synthesis of low-potential ferredoxin, crucial for carbon dioxide reduction in both groups, is dependent on electron bifurcation (Buckel and Thauer, 2013). It is therefore possible to hypothesise the transfer of methyl groups via MtrA-H complex producing a form of substrate-level pumping, using the abundant methyl groups and ion gradients at the hydrothermal vent, which could have developed into an active pumping mechanism without much evolutionary innovation. This may be the most ancient form of pumping in Archaea. Analogous to this, a similar scenario could have happened with the bacterial Rnf complex, similarly utilising naturally occurring ion gradients (Sousa and Martin, 2014). The Rnf complex may therefore present the most ancient form of ion pumping in Bacteria.

In Chapters 3 and 4 we investigate these questions surrounding the evolution of metabolism in the earliest Bacteria, and determine which pathways were present at different ancestral nodes in the tree. In Chapter 3, we present a reconstruction of the central metabolic pathways present in LBCA. In Chapter 4, we extend this to several deep nodes in the bacterial tree in order to evaluate the evolution of these pathways in the deepest parts of the tree.

1.6 One membrane or two? The evolution of the cell envelope

Monoderms vs diderms

Bacteria have been classically divided into two groups based on their response to Gram staining, with some Bacteria resisting the decolourisation step of the process (Gram 1884). These “gram-negative” Bacteria were shown to resist the decolourisation by way of a secondary out membrane, exhibiting a “diderm” architecture, as opposed to those which had a single membrane and did not resist the decolourisation step (Bladen and Mergenhagen, 1964). The two model organisms, *Bacillus subtilis* (a firmicute), and *Escherichia coli* (a gammaproteobacterium) respectively epitomise the classic monoderm and diderm phenotypes (Silhavy, Kahne and Walker, 2010, Megrian *et al.*, 2020). Monoderms typically exhibit a single lipid cell membrane and a thick peptidoglycan wall with teichoic and lipoteichoic acids. Diderms instead have a thin peptidoglycan wall with an inner and an outer lipid membrane,

often embellished with lipopolysaccharides (LPS). A number of systems are involved in the assembly of the classic diderm envelope, including LPS synthesis carried out by the Lpx and Kds enzymes, transport across the inner membrane by MsbA and transport to the outer membrane via the Lpt system, the assembly and insertion of proteins into the outer membrane by the Bam and Tam systems, the insertion of lipoproteins by the Lol system, and the maintenance of lipid asymmetry between the inner and outer membranes by the Mla system (Antunes *et al.*, 2016; Megrian *et al.*, 2020). Other machineries that are found in both monoderms and diderms have specific proteins to anchor themselves to the outer membrane, including the P and L rings (FlgAHI) in flagella, and secretin (PilQ) for type IV pili. Many bacteria, however, may exhibit cell envelopes with a mixture of characteristics which do not follow the classic gram-negative/gram-positive divide (Sutcliffe, 2010).

Scenarios for the origin of the outer membrane

A number of different scenarios have been proposed for the evolution of the double membrane (Fig. 1.2), of which there are two main camps, Monoderm-first and Diderm-first. Under Monoderm-first hypotheses, the monoderm cell envelope, seen as ancestral or “primitive” in its architecture, would have been the ancestral state, with the emergence of the diderm envelope later in evolutionary history as a derived trait. One scenario, proposed by Lake *et al.* (2009), suggests that diderms originated as a fusion between two monoderm bacteria, a firmicute and an actinobacterium (Fig. 1.2). This has been criticised (Gupta, 2011), especially on the grounds that there appears to be no evidence to group all diderms form a single clade to the exclusion of all monoderms. An alternative scenario posits that the outer membrane evolved under antibiotic selection pressure, where diderm Bacteria lacking LPS represent evolutionary intermediates to classical diderms (Gupta, 2011) (Fig. 1.2). Specifically, within this scenario, two gene insertions (Hsp70 and Hsp60 respectively) lead to the classic diderm envelope seen in most modern species, with the Chloroflexota (monoderm but having the Hsp70 insert) and the “simple diderm” (i.e. lacking in LPS) Deinococcota representing transitional stages. Fusobacteriota, Synergistota, Elusimocrobiota, and the two diderm classes of Firmicutes (the Negativicutes and Halanaerobiales), represent “atypical diderms” in this scenario, as they exhibit classic LPS-membranes, but lack the Hsp70 and Hps60 gene insertions (Gupta, 2011), and would have presumably evolved their membranes independently of other diderms, or

through HGT. However, this scenario may not be compatible with current ideas of bacterial phylogeny (Hug *et al.*, 2016; Parks *et al.*, 2017). Tocheva *et al.* (2016) have suggested that the outer membrane may have formed via sporulation, stemming from the observation that a temporary outer membrane is formed during endospore formation in Firmicutes before being lost during spore germination (Fig. 1.2). The outer membrane would have therefore originated once in a spore-forming monoderm ancestor (Tocheva *et al.*, 2011; Errington, 2013; Tocheva, Ortega and Jensen, 2016). However, such sporulation seems to be specific to Firmicutes, and therefore this hypothesis is incompatible with current knowledge of bacterial phylogeny and evolution.

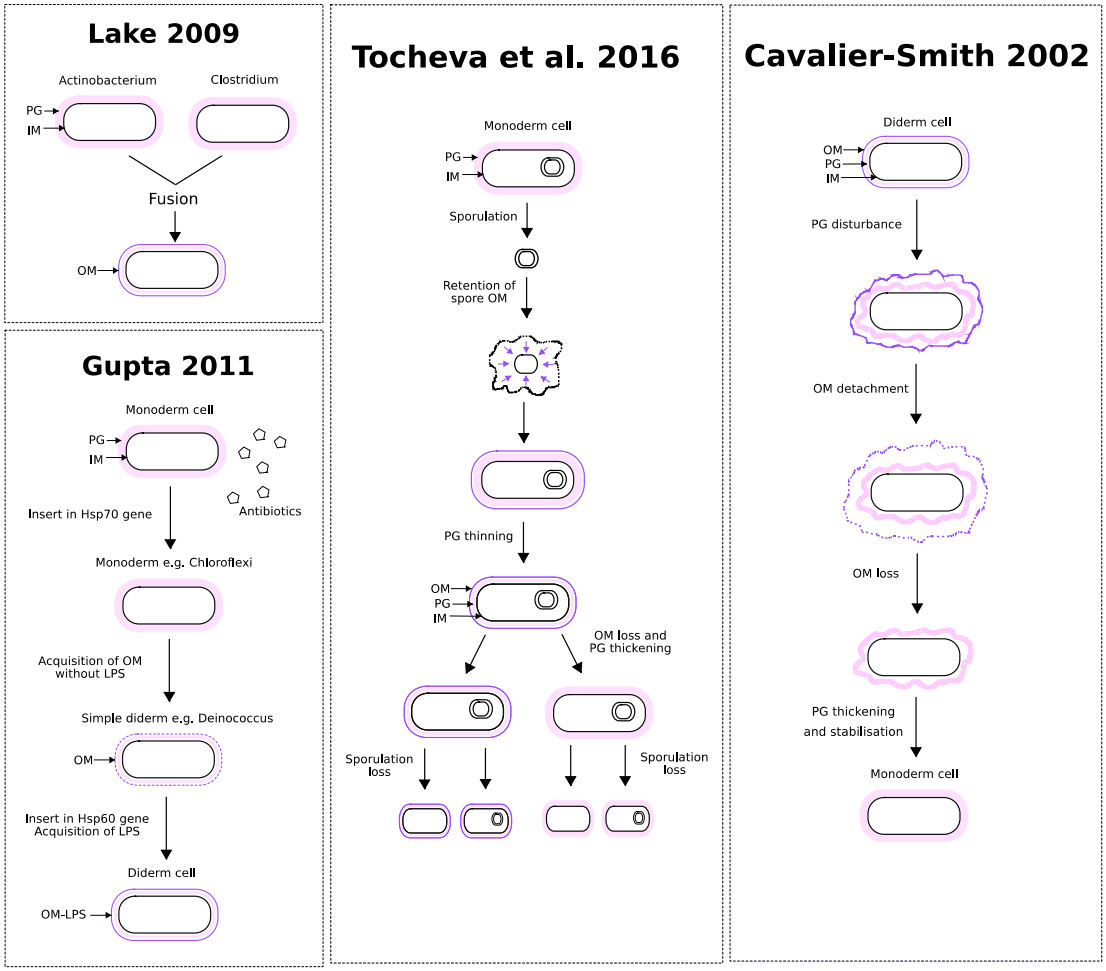


Fig. 1.2 *Summary of different evolutionary hypotheses for the origin of the outer membrane, including Monoderm-first scenarios proposed by Lake (2009), Gupta (2011) and Tocheva et al. (2016), and a Diderm-first hypothesis proposed by Cavalier-*

Smith (2002). Based on Figure 2 form Megrian et al. (2020). PG=peptidoglycan, IM=inner membrane, OM=outer membrane, LPS=lipopolysaccharides.

Diderm-first hypotheses have also been advocated. It has been suggested that the earliest Bacteria were diderm (Cavalier-Smith, 2002), with the loss of the outer membrane occurring due to a mutation which increased the thickness of the peptidoglycan wall, causing the outer membrane attachments to break (Cavalier-Smith, 2006) (Fig. 1.2). Within this scenario, the root of life is within Bacteria, with a single clade of monoderms including Archaea and Eukaryotes, and diderm Chloroflexota at the base of the tree. However, subsequent analysis have shown Chloroflexota to be monoderms (Sutcliffe, 2010), and the basal position of the phylum is contentious (Raymann, Brochier-Armanet and Gribaldo, 2015). More recently, it has been demonstrated based on phylogenetic analyses of associated genes that the classically monoderm Firmicutes are ancestrally diderm (Antunes *et al.*, 2016), and that the monoderm phenotype has arisen multiple times within the phylum. Given the extensive distribution of the diderm phenotype across the tree, it has been argued that this scenario with Firmicutes is analogous to what happened across the bacterial tree, namely a diderm ancestor followed by lineages specific losses (Megrian *et al.*, 2020).

In Chapters 3 and 4 we reconstruct the evolution of cell envelope architecture in Bacteria. In Chapter 3, we present a reconstruction of the cell envelope of LBCA. In Chapter 4, we extend this to several deep nodes in order to evaluate the evolution of cell envelope architecture across the bacterial tree.

1.7 Timing of bacterial evolution

Relatively little research has been carried out attempting to date the bacterial tree, either absolutely using molecular clocks, or using some form of relative dating. Using molecular clocks for Bacteria is difficult due to the need for fossil calibrations, which are sparse and diagnostically uninformative. However, understanding the timing of evolutionary events within the tree is crucial in understanding the evolution of metabolism and physiology through bacterial evolution. As discussed above,

methanogenesis and acetogenesis, and the associated WLP, are often posited as some of the earliest emerging metabolisms (Battistuzzi, Feijao and Hedges, 2004; Ueno *et al.*, 2006; Sousa, Nelson-Sathi and Martin, 2016; Weiss *et al.*, 2016; Williams *et al.*, 2017; Adam, Borrel and Gribaldo, 2018; Wolfe and Fournier, 2018). Being able to infer the timing of the emergence of bacterial clades which use this pathway would be able to lend some support to its ancient origins, whether or not we find evidence for its presence in LBCA.

Another important debate in bacterial evolution revolves around the timing of the emergence of Cyanobacteria and its relation to the Great Oxidation Event (occurring ~2.4 Ga), and by extension the origin of oxygenic photosynthesis. The GOE has often been causally linked to the emergence of the Cyanobacteria (Schirrmeister *et al.*, 2013; Knoll and Nowak, 2017; Sánchez-Baracaldo *et al.*, 2017), although some evidence has placed their emergence later in time (Betts *et al.*, 2018). Yet another debate is the emergence of eukaryotic cells, a key moment in evolutionary history. It has been suggested to have happened early, possibly contemporaneously with or even predating the emergence of prokaryotes (Kurland, Collins and Penny, 2006). It has also been linked to the GOE (Knoll and Nowak, 2017). However, other studies have suggested that eukaryotes evolved relatively late (Chernikova *et al.*, 2011; Parfrey *et al.*, 2011; Eme *et al.*, 2014a; Knoll, 2014; N. J. Butterfield, 2015; Betts *et al.*, 2018). There is now a growing consensus, based on phylogenetics and comparative genomic evidence, that eukaryotic cells arose from a symbiosis between an archaeal host cell and a bacterial endosymbiont that evolved into the mitochondrion (Embley and Martin 2006; Martin et al. 2015; Eme et al. 2017; Roger et al. 2017). Eukaryotic cells would have to postdate the emergence of Alphaprotebacteria.

Some of the questions above may be partially answered by relative dating, that is inferring the order in which clades emerged within the bacterial tree. This can be done using whole-genome approaches such as ALE as they model HGT. Transfers contain information about the relative divergences because donor lineages are necessarily as old as the recipient lineages (Chauve *et al.*, 2017; Davín *et al.*, 2018). This approach has been used to infer that methanogenic Euryarchaeota were the earliest radiating lineages within Archaea, supporting the ancient origin of methanogenesis (Davín *et al.*, 2018). The same study also inferred a relatively late radiation of crow-group

DPANN, a superphylum of Archaea with highly reduced genomes analogous to CPR in Bacteria, despite their early divergence within the tree. Inferring the relative age of diversification of CPR is important in understanding their role, and indeed the role of highly reduced, streamlined cells, in the early evolution of cellular life.

In Chapter 4, we use HGTs to relatively date the emergence of different clades within the bacterial tree, allowing us to infer the order of major events in bacterial evolution. It must be stressed that the analyses and results presented in Chapter 4 are not absolute dates, that is they only tell the order of the events, not when they occurred or the time that elapsed between them. However, such relative time information is still of great use in our attempts to reconstruct the evolutionary history of Bacteria.

1.8 The Lipid divide

Membrane phospholipids across the tree of life

A striking difference between Bacteria and Archaea lies in the phospholipid composition of the cell membranes (Fig. 1.3). Canonically, Bacteria, along with Eukaryotes, have acyl (fatty-acid) chains attached to a glycerol-3-phosphate (G3P) backbone via ester bonds and form bilayers (Lombard, López-García and Moreira, 2012b). Archaea, on the other hand, typically possess isoprenoid chains attached to a glycerol-1-phosphate (G1P) backbone via ether bonds and can have either membrane spanning or bilayer-forming phospholipids (Lombard, López-García and Moreira, 2012b). Bacterial and archaeal phospholipids are synthesised by non-homologous enzymes by different biosynthetic pathways, implying independent evolution of these pathways in each domain. This “lipid-divide” (Koga, 2011) raises important questions regarding the nature of the earliest cells and the evolution of their membranes.

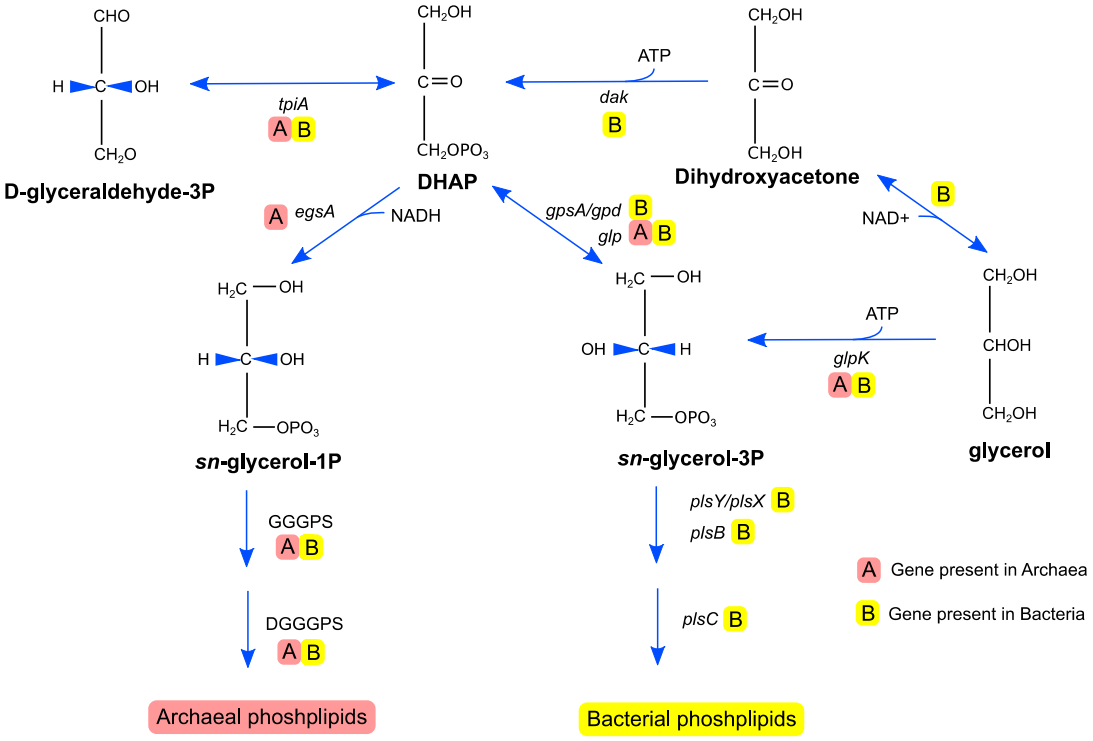


Fig. 1.3 *Schematic representation of the phospholipid biosynthesis pathways in Archaea and Bacteria, based partially on Figure 1 from Peretó et al. (2004).*

Despite the lipid divide being important for our understanding of early cellular evolution, relatively little experimental work has been done to determine the stereochemistry of phospholipids in individual lineages, with most studies assuming that bacterial and archaeal lineages will have their respective stereochemistry as a matter of course. While the limited studies which have determined glycerol stereochemistry seem to support this divide (Sinninghe Damsté *et al.*, 2002; Weijers *et al.*, 2006), there is some evidence to suggest that certain Bacteria have the ability to produce G1P-linked ether lipids. Notably, it has been demonstrated experimentally that *B. subtilis* possesses homologues of archaeal G1P dehydrogenase (G1PDH) and geranylgeranylglyceryl phosphate synthase (GGGPS) (Guldan, Sterner and Babinger, 2008; Guldan *et al.*, 2011), which allow it to produce an archaeal-like phospholipids, although it is unknown if these are used in the *B. subtilis* membrane. Aside from stereochemistry, other characteristics of membrane phospholipids appear to be variable, often exhibiting a mixture of bacterial and archaeal features. For example, plasmalogens found in both Eukaryotes and Bacteria have ether bonds (Goldfine, 2010) and some Archaea have been shown to produce membrane lipids with fatty-

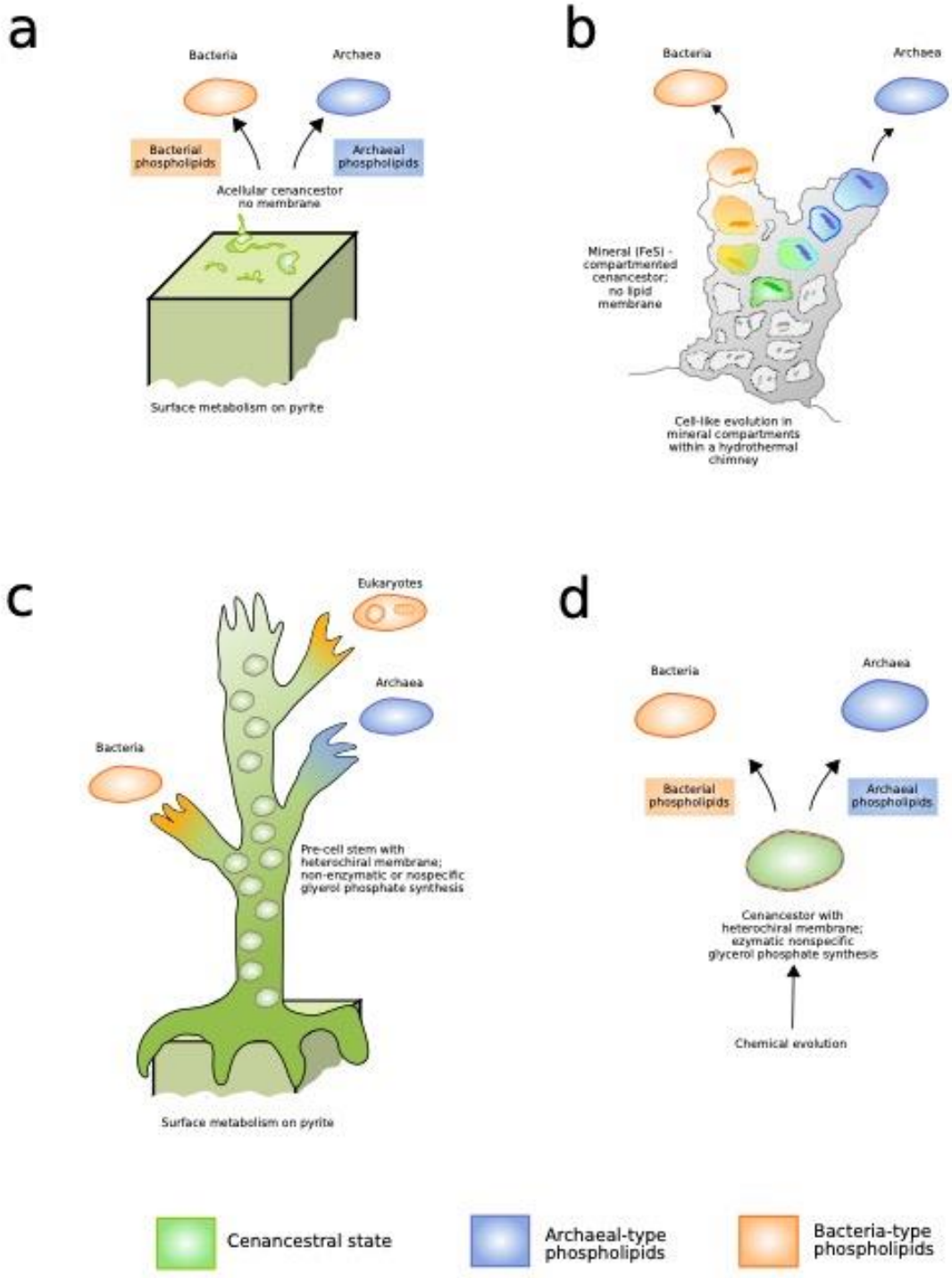
acids (Gattinger, Schloter and Munch, 2002). Of great interest are branched glycerol dialkyl glycerol tetra-ethers (brGDGTs) found in the environment, which exhibit bacterial glycerol stereochemistry, and use branched alkyl chains (rather than archaeal isoprenoid chains), but which have ether bonds and are membrane spanning, characteristics usually associated with archaeal lipids (Schouten *et al.*, 2000; Weijers *et al.*, 2006). These brGDGTs are particularly abundant in peat bogs, where their unusual mixture characteristics were thought to be bacterial adaptations to low pH environments (Weijers *et al.*, 2006; Damsté, Sinninghe Damsté, *et al.*, 2007), although they are now known to occur in a wide range of soil and aquatic environments (Schouten, Hopmans and Sinninghe Damsté, 2013). The biosynthetic pathways and associated enzymes for these mixed-type membrane lipids remain enigmatic, but given the frequency of prokaryotic HGT (Hemmi *et al.*, 2004), it is not unreasonable to assume that they may reflect pathways of mixed bacterial and archaeal origin. This indicates that the lipid-divide, thought to be such a defining difference between the two domains of life, may be less clear-cut than previously thought.

Possible scenarios for the evolution of membrane phospholipids

Several different hypotheses have been suggested to explain the origins of the different pathways, and the nature of the membrane of LUCA, summarised here in Fig. 1.4. There is some debate as to whether LUCA was acellular, living on the surface of pyrite (Koga *et al.*, 1998) (Fig. 1.4a) or in mineral-bounded compartments within a hydrothermal chimney (Martin and Russell, 2003) (Fig. 1.4b), with lipid membrane evolving independently in each domain at a later point. However, the presence of some genes for lipid biosynthesis (Lombard and Moreira, 2011; Lombard, López-García and Moreira, 2012b; Koga, 2014; Weiss *et al.*, 2016) and, in particular, a membrane-bound ATPase (Sojo, Pomiankowski and Lane, 2014; Weiss *et al.*, 2016) in reconstructions of LUCA implies that it possessed a membrane, although its properties may have differed from those of modern, prokaryote cell membranes (Lombard, López-García and Moreira, 2012b; Koga, 2014; Sojo, Pomiankowski and Lane, 2014). Alternatively, archaeal and bacterial phospholipid biosynthesis may have evolved from a stem of pre-cells with heterochiral membranes (Wächtershäuser, 2003) (Fig. 1.4c) or a heterochiral LUCA with membranes synthesised via universal, substrate-nonspecific enzymes (Peretó, López-García and Moreira, 2004) (Fig. 1.4d). In the latter hypothesis, the heterochiral membrane would be less stable than a homochiral one,

putting selective pressure on the ancestral bacterial and archaeal populations to shift to a homochiral membrane with either phospholipid type, although there is evidence to suggest that heterochiral membranes are not less stable than homochiral ones (Fan *et al.*, 1995; Shimada and Yamagishi, 2011; Caforio *et al.*, 2018).

Fig. 1.4 (below) *Representation of four different models of the origin and early evolution phospholipid biosynthesis in Archaea and Bacteria. a) independent evolution of archaeal and bacterial pathways from an acellular cenancestor (Kog et al. 1998); b) independent evolution of archaeal and bacterial pathways from mineral bound compartments (Martin and Russel 2003); c) evolution of domain specific pathways form a stem of heterochiral pre-cells (Wächtershäuser, 2003); d) evolution of domain specific pathways form a fully cellular, heterochiral ancestor (Peretó, López-García and Moreira, 2004). Based on Figure 2 from Peretó et al. (2004)*


a
Bacteria
Archaea
Bacterial phospholipids
Archaeal phospholipids
Acellular cenancestor no membrane
Surface metabolism on pyrite
b
Bacteria
Archaea
Mineral (FeS) - compartmented cenancestor; no lipid membrane
Cell-like evolution in mineral compartments within a hydrothermal chimney
c
Eukaryotes
Archaea
Bacteria
Pre-cell stem with heterochiral membrane; non-enzymatic or nospecific glycerol phosphate synthesis
Surface metabolism on pyrite
d
Bacteria
Archaea
Bacterial phospholipids
Archaeal phospholipids
Cenancestor with heterochiral membrane; ezymatic nonspecific glycerol phosphate synthesis
Chemical evolution
Cenancestral state
Archaeal-type phospholipids
Bacteria-type phospholipids

It is also possible that LUCA was homochiral with either type of phospholipid, with the evolution of the other in its respective lineage later in evolutionary history (Yokobori *et al.*, 2016), although it is unclear what would prompt such a change. If archaeal phospholipids are ancestral (Daiyasu *et al.*, 2002; Peretó, López-García and Moreira, 2004; Carbone *et al.*, 2015), the change to bacterial phospholipids within Bacteria may have been driven by the flexibility and adaptability afforded by bacterial lipid architecture. Namely, based on chemical considerations, bacterial phospholipids may be cheaper to make and break. They also allow a greater variety of fatty acyl moieties, varying in chain length, unsaturation, degree of branching and cyclisation compared to archaea-type phospholipids, allowing better adaptation to diverse environments. These characteristics may have given marginal benefits in various dynamic mesophilic environments, and would be a possible explanation to the relatively higher abundance of Bacteria compared to Archaea in most environments (Danovaro *et al.*, 2016; Hug *et al.*, 2016; Castelle and Banfield, 2018). Conversely, if bacterial-type phospholipids are ancestral (Yokobori *et al.*, 2016), the evolution of archaea membrane may have been driven by adaptation to high temperatures (Akanuma *et al.*, 2013; Akanuma, Yokobori and Yamagishi, 2013; Yokobori *et al.*, 2016), as ether bonds are more thermostable than esters (Vossenberg *et al.*, 1998; Koga, 2012) and are also found in the membranes of thermophilic Bacteria (Kaur *et al.*, 2015). It should be noted however that the widespread occurrence of bacterial-, archaeal- and mixed-type membranes suggest that, except in thermophilic or low pH environments, there seems to be little advantage to either membrane.

In chapter 3 and 4, we present evidence for the composition of lipid membranes present in LBCA and subsequent nodes. In Chapter 5, we expand our study to the whole tree of life. Using expanded taxon sampling, including environmental samples, and using the best evolutionary models available to us, we present a reconstruction of the evolutionary history of the gene families involved in phospholipid biosynthesis.

1.9 A model for the evolution of early life

As discussed above, there are still a number of fundamental questions about bacterial evolution that are unresolved. While there are many difficulties and challenges involved in such deep phylogenetic reconstructions, by using the best models and data currently available to us, this thesis attempts to answer such questions. We use new and innovative methods in order to overcome the problems associated with traditional phylogenetic methods, and which allows us to incorporate a much larger breadth of data. The results generated from these analyses will allow us to test and suggest novel models and hypotheses regarding the evolution of Bacteria. Specifically, the use of whole-genomes approaches, such as ALE, will allow us to root the bacterial tree, as well as model HGT over time and infer ancestral gene content. The inference of ancestral gene content will allow us to reconstruct the metabolic capabilities and habitat of the earliest Bacteria, and answer many of the questions discussed above relating to the evolution of particular physiologies and characters. There are of course many caveats to our analyses. Practical concerns must be considered and sometimes compromises have to be made in order to make the analyses computationally feasible e.g. the use of maximum likelihood instead of Bayesian analysis, or using reduced taxon sampling. These caveats and how they can be improved upon are further discussed through the thesis and in more detail in the Chapter 6. Ultimately, while none of our analyses are without their caveats, we believe by using innovative methods and different lines of evidence in this thesis, we can advance the field of evolutionary microbiology and shed greater light on the earliest period of the history of life.

Chapter 2

Phylogenomics produces a rooted tree of Bacteria

A version of this chapter forms part of a paper under revision in collaboration with Adrián A. Davín, Tara Mahendrarajah, Anja Spang, Philip Hugenholtz, Gergely J. Szöllősi, and Tom A. Williams. Gareth A. Coleman is the first author of the paper. The project was conceived by TAW, GJSz, PH, AS, GAC and AAD. Analyses for the GTDB dataset were performed by GAC, AAD, TAW and GJSz. GJSz developed new analytical methods. Pipelines for orthologue selection and gene family generation were developed by GAC. All authors contributed to interpretation and writing. All analyses, writing and interpretation for the non-GTDB dataset were performed by GAC. For the ToL rooting section, species selection was carried out by GAC, orthologues were given by TAW, HiFix trees generated by Edmund Moody, species trees inferred by GAC, and all ALE analyses carried out by GAC. All writing and interpretation for the ToL section was carried by GAC.

Paper preprint as:
Coleman, G.A., Davín, A.A., Mahendrarajah, T., Spang, A.A., Hugenholtz, P., Szöllősi, G.J. and Williams, T.A., 2020. A rooted phylogeny resolves early bacterial evolution. *bioRxiv*.

BioRxiv preprint for the paper can be found here:
https://www.biorxiv.org/content/10.1101/2020.07.15.205187v1

Abstract

Bacteria are the most abundant and metabolically diverse cellular lifeforms on Earth. A rooted bacterial phylogeny provides a framework to interpret this diversity and to understand the nature of early life. Inferring the position of the bacterial root is complicated by incomplete taxon sampling and the long branch to the archaeal outgroup. To circumvent these limitations, we model bacterial genome evolution at the level of gene duplication, transfer and loss events, allowing outgroup-free inference of the root. We infer a rooted bacterial tree on which 68% of gene transmission events are vertical. Our analyses reveal a basal split between Terrabacteria and Gracilicutes, which together encompass almost all known bacterial diversity. However, the position of a few small phyla could not be resolved in relation to these two major clades. In contrast to recent proposals, our analyses strongly reject a root between the Candidate Phyla Radiation (CPR) and all other Bacteria. Instead, we find that the CPR is a sister lineage to the Chloroflexota within the Terrabacteria.

2.1 Introduction

Rooting deep radiations (Williams *et al.*, 2017) is among the greatest challenges in phylogenomics, and there is no consensus on the root of the bacterial tree. Based on evidence (Iwabe *et al.*, 1989; J. P. Gogarten *et al.*, 1989; Brown and Doolittle, 1995; Zhaxybayeva, Lapierre and Gogarten, 2005) that the root of the entire tree of life lies between Bacteria and Archaea, early analyses using an archaeal outgroup placed the bacterial root near Aquificales/Thermotogales (Bocchetta *et al.*, 2000; Battistuzzi and Hedges, 2009) or Planctomycetes (Brochier and Philippe, 2002). Alternative approaches, including analyses of gene flows and polarisation of changes in multimeric protein complexes and other complex characters (Cavalier-Smith, 2006), have instead suggested roots within the monoderm (single-membrane) Bacteria (Lake, 2009), or between Chloroflexi and all other cellular life (Cavalier-Smith, 2006). The development of techniques for sequencing microbes directly from environmental samples, without the need for laboratory cultivation, has greatly expanded the genomic representation of natural prokaryotic diversity (Hug *et al.*, 2016; Mukherjee *et al.*, 2017; Parks *et al.*, 2017, 2018). Recent phylogenomic analyses of that expanded diversity have placed the bacterial root between one of these new groups, the Candidate Phyla Radiation (CPR; also known as Patescibacteria (Brown *et al.*, 2015; Zhu *et al.*, 2019)) and all other Bacteria (Hug *et al.*, 2016; Castelle and Banfield, 2018; Zhu *et al.*, 2019). The CPR comprises lineages that are characterised by small cells and genomes, and are suggested to have predominantly symbiotic or parasitic lifestyles, but much remains to be learned about their ecology and physiology (Brown *et al.*, 2015; Castelle and Banfield, 2018; Castelle *et al.*, 2018; Beam *et al.*, 2020). If correct, the early divergence of CPR has important implications for our understanding of the earliest period of cellular evolution. Taken together with evidence that the root of the archaeal domain lies between the reduced and predominantly host-associated DPANN superphylum and the rest of Archaea (Castelle *et al.*, 2015; Williams *et al.*, 2017), the CPR root would imply that streamlined, metabolically minimalist prokaryotes have co-existed with the more familiar, self-sufficient lineages throughout the history of cellular life (Beam *et al.*, 2020).

Historically there has also been little agreement on the relationships between different bacterial phyla. In recent years, some superphyla-level groupings have become widely accepted, namely FCB/Sphingobacteria (Fibrobacteres, Chlorobi, Bacteroidetes, and candidate phyla Cloacimonetes, Gemmatimonadetes, Ignavibacteria, Latescibacteria, Marinimicrobia and Zixibacteria, and the genus *Caldithrix*) (Gupta, 2004; Hug *et al.*, 2016; Castelle and Banfield, 2018; Parks *et al.*, 2018) and PVC/Planctobacteria (Planctomycetes, Verrucomicrobia and Chlamydiae, Lentisphaerae, and candidate phylum Omnitrophica) (Cavalier-Smith, 2002; Wagner and Horn, 2006; Hug *et al.*, 2016; Parks *et al.*, 2017; Castelle and Banfield, 2018). More tentative higher level relationships have also been suggested. Battistuzzi and Hedges (2009) divide the bacterial tree into two major clades; one clade comprising the gram-positive Bacteria (Firmicutes and Actinobacteria), Cyanobacteria, Chloroflexi and Deinococcus-Thermus, which they name Terrabacteria (Battistuzzi, Feijao and Hedges, 2004), as they assert that this clade was ancestrally terrestrial; and another clade comprising PVC, FCB and Proteobacteria, which they name Hydrobacteria (Battistuzzi and Hedges, 2009). The Terrabacteria has received some support in subsequent analyses (Bern and Goldberg, 2005; Boussau, Guéguen and Gouy, 2008). Hydrobacteria has alternatively been described by Cavalier-Smith as the Gracilicutes (Cavalier-Smith, 2006), as well being supported by other studies (Boussau, Guéguen and Gouy, 2008). Yet, despite this progress, the deep relationships within the bacterial tree are still highly debated, with the position of the root being particularly unclear.

Improved taxon sampling can help to resolve difficult phylogenetic problems (Graybeal, 1998; Hedtke, Townsend and Hillis, 2006), and the enormous quantity and diversity of genome data now available presents an unprecedented opportunity to resolve long-standing questions about the origins and diversification of Bacteria. But deep phylogenetic divergences are difficult to resolve, both because the phylogenetic signal for deep relationships is overwritten by new changes through time, and also because the process of sequence evolution is more complex than the best-fitting models currently available. In particular, variation in nucleotide or amino acid composition across the sites of the alignment and the branches of the tree can induce long branch attraction (LBA) artefacts in which deep-branching, fast-evolving, poorly-sampled or compositionally biased lineages group together irrespective of their evolutionary history (Bergsten, 2005). These issues are widely appreciated (Hug *et*

al., 2016) but are challenging to adequately address, particularly when sequences from thousands of taxa (Hug *et al.*, 2016; Parks *et al.*, 2017, 2018; Castelle and Banfield, 2018; Zhu *et al.*, 2019) are used to estimate trees of global prokaryotic diversity, which precludes the use of the best available phylogenetic methods.

The traditional method of outgroup rooting has proven difficult as the closest outgroup to Bacteria are Archaea. This is highly problematic as Bacteria and Archaea represent the two fundamental domains of life, branching deep in evolutionary time, and creating long stem lineages to the crown groups. The information provided by archaeal outgroups are thus of not much use and may create artefactual results. Gene duplications are a useful way of resolving roots in clades without the use of an outgroup, and has been used in the rooting of the tree of life (Iwabe *et al.*, 1989; J. P. Gogarten *et al.*, 1989; Brown and Doolittle, 1995). Recently, it has also been shown that the use of gene gains, losses and HGTs can also be used alongside gene duplications as an effective method of rooting a species tree (Abby *et al.*, 2012; Szöllosi *et al.*, 2012; Szöllősi *et al.*, 2013). These can be integrated with probabilistic gene tree-species tree reconciliation methods, such as the recently developed Amalgamated Likelihood Estimation (ALE) method. When looking for the root in a species tree, ALE calculates the maximum likelihood of each gene family tree, given a chosen root position, and rates of gene duplication, transfer and loss (DTLs) (Szöllősi *et al.* 2013). Each chosen root position implies a particular evolutionary history of that gene family, with regards to transfer, loss, or gain of genes, which can be evaluated under maximum likelihood. ALE is innovative in that it can incorporate uncertainty in the underlying gene trees using conditional clade probabilities, and has been shown to be able to infer numbers of gene duplications, transfers, losses and ancestral genome sizes accurately over large phylogenetic depths (Szöllősi *et al.* 2013).

2.2 Methods

Taxon sampling

To obtain a representative taxon sampling from across known bacterial diversity, we sampled taxa according to the classification provided by the Genome Taxonomy

Database (GTDB r89) (Parks *et al.*, 2018). We sampled 265 genomes from the GTDB as follows. First, we filtered out the genomes with Quality < 0.75 (Quality is defined as Completeness - (5*Contamination) (Parks *et al.*, 2017)), and filtered out all phyla subsequently left with fewer than 10 species. Genomes were sampled from the remaining taxa on a per-class basis: for classes containing a single order, the genome with the highest quality score was sampled; for classes containing multiple orders, the highest quality genome from each of two randomly chosen orders was sampled. This protocol ensured that every class in the GTDB is represented in the final tree. We then manually added the genome of *Gloeomargarita litophora* given its importance in constraining the phylogeny and timing of chloroplast evolution (Appendix A, Table 1).

To sample representative bacterial taxa independently of the GTDB, we began with the bacterial portion of a recent global analysis of the tree of life (Hug *et al.*, 2016). We initially generated our subsample using an algorithm which maximised genetic difference between lineages in order to select 200 taxa that were evenly distributed across the tree. However, this selection process was biased in favour of long branches. We also explored other programmatic ways of selecting taxa based on genetic diversity, but these were found to be extremely sensitive to the phylogeny and initial choice of root position. Instead, we inferred a tree of the bacterial portion of the concatenate under the LG+G4+F model in IQ-Tree. We divided the tree into 7 major bacterial clades based on a literature search (Table 2.1) and additional environmental lineages with branch length diversity comparable to the known groups. For each group defined in this way, we manually subsampled taxa so as to maintain genetic diversity, while avoiding overly long or short branches. We selected 342 species, comprising 200 ‘classic’ bacteria, 125 CPR bacteria and one bacterial genome respectively from each of the 17 new phyla described by (Parks *et al.*, 2017) (Appendix A, Table 2). For all species in both datasets, proteomes were download as amino acid sequences, and contain plasmids.

Clade	No. of taxa sampled
Firmicutes	25
Actinobacteriota+Cyanobacteria+Chloroflexota	35
CPR	125

FCB+PVC+Elusimicrobiota	35
Proteobacteria	50
Deltaproteobacteria+Nitrospirota+Acidobacterota+Aquificota	30
FASST+environmental lineages	25
New Phyla (Parks et 2018)	17

Table 2.1 *Number of taxa sampled from each clade in the GTDB-independent analysis.*

Orthologue selection

We used OMA 2.1.1 (Roth, Gonnet and Dessimoz, 2008) to identify candidate single-copy bacterial orthologues, and retained those with at least 75% of all species represented in each family. Sequences were aligned in Mafft using the -auto option, and trimmed in BMGE 1.12 (Criscuolo and Gribaldo, 2010) using the BLOSUM30 model. Initial trees were inferred for each candidate orthologue under the LG+G+F model in IQ_TREE 1.6.10. The trees were manually inspected, and we selected orthologues where the monophyly of 14 pre-defined major lineages was not violated with bootstrap support >70%, resulting in 63 final orthologues (Table 2.2). The same selection process was used on both datasets, generating the same set of orthologues.

KO number	Gene name	Annotation	Used in outgroup tree?
K03046	rpoC	DNA-directed RNA polymerase subunit beta'	**y**
K03043	rpoB	DNA-directed RNA polymerase subunit beta	
K02337	dnaE	DNA polymerase III subunit alpha	**y**
K03070	secA	Protein translocase subunit SecA	
K01873	VARS, valS	Valine--tRNA ligase	
K02335	polA	DNA polymerase I	**y**
K01872	AARS, alaS	Alanine tRNA ligase	
K02469	gyrA	DNA gyrase subunit A	
K00962	pnp, PNPT1	Polyribonucleotide nucleotidyltransferase	**y**
K02355	fusA, GFM, EFG	Translation elongation factor G	**y**
K01972	E6.5.1.2, ligA, ligB	DNA ligase NAD	
K03702	uvrB	Excinuclease ABC subunit B	
K02470	gyrB	DNA gyrase subunit B	

K04077	groEL, HSPD1	Molecular chaperone GroEL	
K01937	pyrG, CTPS	CTP synthase	**y**
K02313	dnaA	Chromosomal replication initiator protein	
K02314	dnaB	Replicative DNA helicase	
K02433	gatA, QRSL1	aspartyl-tRNA(Asn)/glutamyl-tRNA(Gln) amidotransferase subunit A	
K03076	secY	Protein translocase subunit SecY	**y**
K04485	radA, sms	DNA repair protein RadA/Sms	
K02112	ATPF1B, atpD	F-type H+/Na+-transporting ATPase subunit beta	**y**
K03590	ftsA	Cell division protein FtsA	
K02358	tuf, TUFM	Elongation factor Tu	**y**
K06942	ychF	Redox Regulated ATPase YchF	
K00927	PGK, pgk	Phosphoglycerate kinase	
K01889	FARSA, pheS	Phenylalanine--tRNA ligase alpha subunit	
K03551	ruvB	Holliday junction branch migration DNA helicase RuvB	**y**
K04485	radA, sms	DNA recombination repair protein RecA	**y**
K02835	prfA, MTRF1, MRF1	Peptide chain release factor 1	
K02886	RP-L2, MRPL2, rplB	50S ribosomal protein L2	**y**
K01803	TPI, tpiA	Triose phosphate isomerase	**y**
K03438	mraW, rsmH	16S rRNA (cytosine1402-N4)-methyltransferase	
K00554	trmD	tRNA (guanine37-N1)-methyltransferase	
K02863	RP-L1, MRPL1, rplA	50S ribosomal protein L1	**y**
K03685	rnc, DROSHA, RNT1	Ribonuclease III	
K02967	RP-S2, MRPS2, rpsB	30S ribosomal protein S2	**y**
K02982	RP-S3, rpsC	30S ribosomal protein S3	
K02906	RP-L3, MRPL3, rplC	50S ribosomal protein L3	**y**
K03470	rnhB	Ribonuclease HII	
K01358	clpP, CLPP	ATP dependent Clp protease proteolytic subunit	
K06187	recR	Recombination protein RecR	
K15034	yaeJ	Aminoacyl tRNA hydrolase, ribosome-associated protein	
K02931	RP-L5, MRPL5, rplE	50S ribosomal protein L5	**y**
K02933	RP-L6, MRPL6, rplF	50S ribosomal protein L6	**y**
K02601	nusG	Transcription termination antitermination protein NusG	
K02988	RP-S5, MRPS5, rpsE	30S ribosomal protein S5	**y**
K02992	RP-S7, MRPS7, rpsG	30S ribosomal protein S7	**y**
K03664	smpB	SsrA binding protein	

K02838	frr, MRRF, RRF	Ribosome recycling factor	
K02867	RP-L11, MRPL11, rplK	50S ribosomal protein L11	
K02878	RP-L16, MRPL16, rplP	50S ribosomal protein L16	**y**
K02871	RP-L13, MRPL13, rplM	50S ribosomal protein L13	**y**
K02994	RP-S8, rpsH	30S ribosomal protein S8	**y**
K02948	RP-S11, MRPS11, rpsK	30S ribosomal protein S11	**y**
K02952	RP-S13, rpsM	30S ribosomal protein S13	**y**
K02935	RP-L7, MRPL12, rplL	50S ribosomal protein L7/12	
K02996	RP-S9, MRPS9, rpsI	30S ribosomal protein S9	
K02874	RP-L14, MRPL14, rplN	50S ribosomal protein L14	**y**
K02887	RP-L20, MRPL20, rplT	50S ribosomal protein L20	
K02946	RP-S10, MRPS10, rpsJ	30S ribosomal protein S10	**y**
K02965	RP-S19, rpsS	30S ribosomal protein S19	**y**
K02956	RP-S15, MRPS15, rpsO	30S ribosomal protein S15	
K02518	infA	Translation initiation factor IF 1	

Table 2.2 *63 single-copy orthologous used to infer the species tree, with those used in the outgroup rooting analysis indicated.*

Species tree inference

For the GTDB dataset, we concatenated the 63 orthologues resulting in an alignment of 18,234 amino acids. We inferred an unrooted phylogeny from this concatenate under the LG+C60+R8+F model, which was chosen as the best-fitting model by the BIC criterion in IQ-TREE (Nguyen *et al.*, 2015). We additionally removed the most compositionally heterogeneous sites from the sequence alignment using Alignment Pruner (Dombrowski *et al.*, 2020) (20%, 40%, 60% and 80% respectively) and inferred trees using the same procedure described above in order to compare the resulting topologies.

For the GTDB-independent analysis, we also concatenated the 63 orthologues resulting in an alignment of 17,428 amino acids. A tree was inferred in IQ-Tree using the LG+C20 model, with PMSF (Wang *et al.*, 2018) - an ML implementation based on a finite mixture of site specific amino acid profiles of a CAT model, which can circumnavigate problems of long branch attraction (LBA), and therefore suitable for

trees with deeply diverging taxa. However, the deep relationships with Bacteria were not well resolved. We therefore performed sensitivity analyses to evaluate the robustness of our topology. We created concatenated alignments with a sub-sample of 17 CPR and with no CPR respectively, and inferred LG+PMSF trees from these alignments in IQ-Tree. We created two further alignments; one recoded using the four category scheme of Susko and Roger (2007); and another with a reduced taxon sample of 98 species. Bayesian CAT+GTR+G4 (Lartillot and Philippe, 2004) trees were inferred from both of these trees in PhyloBayes (Lartillot and Philippe, 2004; Lartillot, Brinkmann and Philippe, 2007b). Additionally, we inferred a tree under the multispecies coalescent model in ASTRAL (Zhang, Sayyari and Mirarab, 2017).

In order to further test hypotheses on topology, we performed a series of approximately unbiased (AU) tests (Shimodaira, 2002) on the GTDB-independent dataset, comparing various competing unrooted tree topologies. All AU tests in this chapter were performed in IQTree, using 1000 RELL boostrap replicates. We compared the unrooted topologies from the various tree construction methods described above, as well as trees constrained to test the validity of the "FASSyT" group (a grouping of five phyla, Fusobacteriota, Aquificota, Synergystota, Spirochaetota and Thermotogota). We additionally constrained the tree to maintain the monophyly of Terrabacteria without the CPR. Further, we constrained all monoderms and all diderms to form clades respectively. We compared all trees to the LG+PMSF tree to determine which tree topologies could be rejected (AU p-value > 0.05).

Outgroup rooting

To root the bacterial tree using an archaeal outgroup, we used a representative sampling of 148 archaeal genomes, with the concatenated alignment including a subset of 30 out of the 63 bacterial orthologues that were shared between bacteria and archaea, as determined by HMM searches and manual inspection of single orthologue trees. For the GTDB dataset, we inferred the ML tree in IQ-TREE under the best-fitting LG+C60+R8+F model. We performed an AU test to determine whether a range of published alternative rooting hypotheses could be rejected, given the model and data (AU p-value > 0.05). For the non-GTBD dataset, we inferred a tree under the LG+PMSF model in IQ-Tree for the full alignment, an alignment with reduced CRP

and an alignment with no CPR. We additionally inferred a tree under a CAT+GTR+G4 model in PhyloBayes for a recoded alignment.

Outgroup-free rooting methods

We rooted the trees derived from the non-GTDB dataset using a lognormal uncorrelated relaxed molecular clock, with an LG (Le and Gascuel, 2008) substitution model in BEAST2 (Drummond and Rambaut, 2007; Drummond and Bouckaert, 2015). We further rooted our trees using minimal ancestor deviation (MAD) (Tria, Landan and Dagan, 2017), which finds the root position that minimises pairwise evolutionary rate variation, averaged over all pairs of taxa in the tree, and uses the ambiguity index (AI) which is defined as the ratio of the MAD value to the second smallest value. The higher the AI value obtained for a root position, the less statistically distinguishable it is from the next best root.

Gene family clustering and ALE analysis

For the GTDB dataset, we used the protein annotations provided by GTDB, which were originally obtained using Prodigal. To infer homologous gene families for ALE inference, we performed an all vs all similarity search using Diamond (Buchfink, Xie and Huson, 2015) with an E-value threshold of $<10^{-7}$ to avoid distant hits and k = 0 to report all the relevant hits.

We performed clustering using MCL (van Dongen, 2000) with an inflation parameter of 1.2. Current clustering methods are not consummate and the parameters that determine the granularity of clustering do not have a direct biological motivation. Setting the value of the MCL inflation parameter therefore involves a trade-off between inferring large, inclusive clusters that will contain false positives (sequences that are not part of the real gene family) and small, conservative clusters that may divide real gene families into several subclusters. An additional practical concern for phylogenomics is that overly large clusters can align poorly and result in low-quality single protein trees. In our rooting analysis, we experimented with a range of values for the inflation parameter, and chose 1.2 because the clusters were inclusive without a substantial reduction in post-masking alignment length compared to more granular settings. This resulted in 186,827 gene families and a total of 11,765 families with 4 or more sequences. We aligned the 11,765 gene families using Mafft with the --auto

option, and filtered with BMGE using the BLOSUM30 model. After filtering, 260 alignments contained no high-quality columns and were discarded. We filtered out sequences containing more than 80% of gaps to produce the final set of alignments. We also discarded all alignments with less than 30 columns, leaving a total of 11,272 families. The gene trees were computed using IQ-TREE v 1.6.10 using model testing. Conditional clade probabilities (CCPs) were computed using ALEobserve and the resulting ALE files were reconciled with the species tree. Loss rates were corrected by genome completeness, estimated using CheckM (Parks *et al.*, 2015). We tested 62 roots.

For the non-GTDB dataset, we downloaded the genomes from NCBI GeneBank and used the same pipeline as described above, resulting in 11,781 gene families with 4 or more sequences. The trees and conditional clade probabilities were calculated as described above. We tested 15 roots using four unrooted topologies: the topology from the unaltered concatenate, the topology from the recoded concatenate, a topology constrained to resemble that of the concatenate with reduced sample of CPR, and the topology from the MSC tree.

Quantifying vertical and horizontal signals in bacterial genome evolution

In the context of our analyses, “verticality” is the proportion of inferred evolutionary events that reflect vertical descent, estimated using gene tree-species tree reconciliation. We considered two kinds of verticality: branch-wise verticality, the proportion of vertical evolutionary events on a branch in the species tree; and family-wise verticality, the proportion of vertical events during the evolution of a specific gene family. We defined branch-wise verticality as V/(V+O+T), where V is the inferred number of vertical transmissions of a gene from the ancestral to descendant ends of the branch; O is the number of new gene originations on the branch; and T is the number of gene transfers into the branch. We defined family-wise verticality as V/(V+T), where V and T refer to inferred numbers of events within the history of a gene family (Table 2.3). The numbers reported here have been averaged over the reconciliations obtained using the three possible roots.

COG	Annotation	Number of families	Median verticality	Mean verticality

V	Defense mechanisms	32	0.5280122078075817	0.5638668252634967
T	transduction	88	0.5799451471715107	0.604432180596873
G	Carbohydrate	186	0.5913637562346645	0.6015642843633491
Q	Secondary metabolites	73	0.5964080412431355	0.6073596199333191
L	Replication	179	0.5987938232160366	0.6150912826158655
P	Inorganic ion	199	0.5988930441950502	0.6115949584832507
O	Post-translational modification	123	0.6013500360389593	0.6131412846715851
K	Transcription	131	0.6078333513305463	0.6340799581816328
I	Lipid	77	0.6124862586859439	0.6216075914824997
C	Energy	239	0.6145542409023125	0.6275371657010402
M	Cell wall/membrane	141	0.6155364029228826	0.625358402235949
H	Coenzyme	165	0.6233926976132386	0.6295188636764987
E	Amino acid	226	0.6235372737809106	0.631566055802421
F	Nucleotide	102	0.64234381501306	0.6370022274662093
N	Cell motility	70	0.6825519330347292	0.6801963045783017
D	Cell cycle	45	0.6849512300407962	0.6905029550824039
U	Intracellular trafficking	83	0.6854327621149885	0.6880748036306573
J	Translation	177	0.6906677776226816	0.6870530339985472

Table 2.3 Mean verticality V/(V+T) by COG functional category.

2.3 Results and Discussion

An unrooted phylogeny of Bacteria

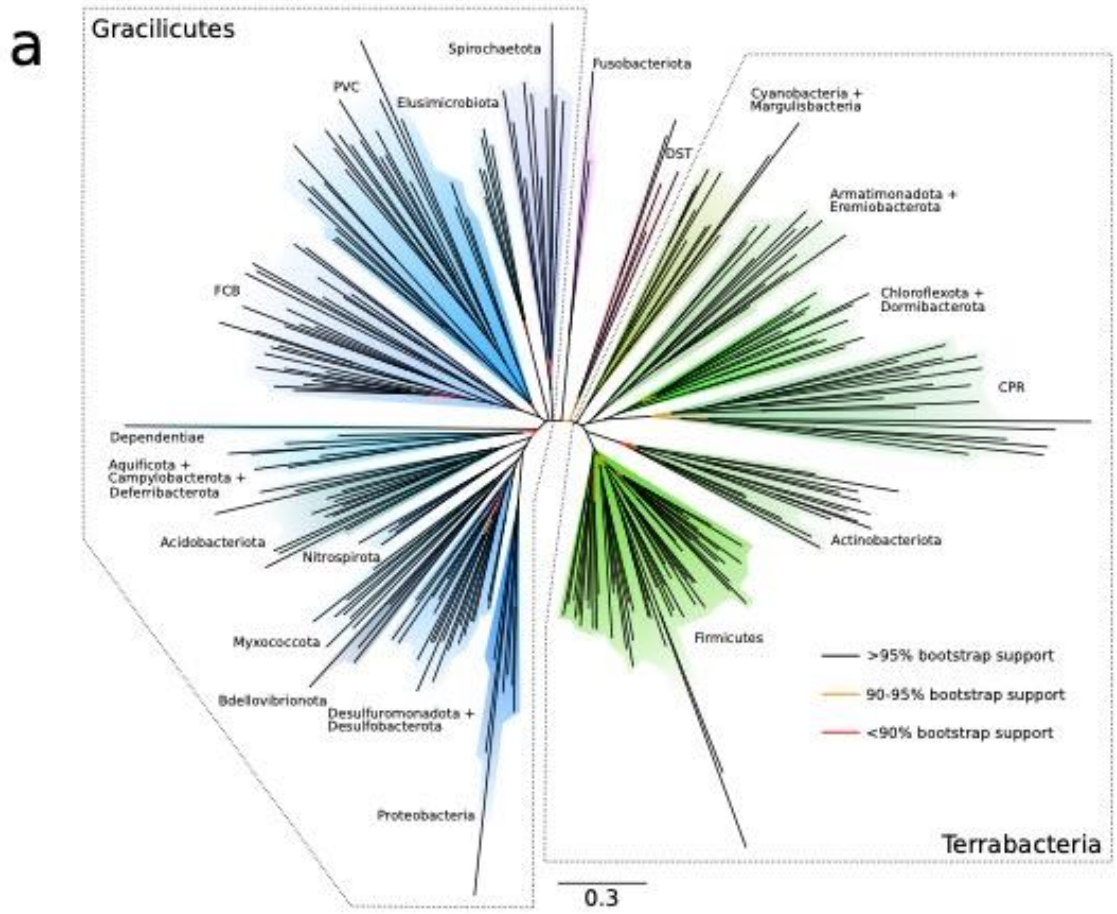
We inferred an unrooted species tree from the concatenation of the 63 markers using the LG+C60+R8+F model in IQ-Tree 1.6.10 (Fig. 2.1a), which was chosen as the best-fitting model using the Bayesian Information Criterion. We obtained highly congruent trees when removing 20-80% of the most compositionally heterogeneous sites from the alignment (Fig. 2.1b-e), suggesting that the key features of the topology are not composition-driven LBA artefacts. All trees were consistent with the GTDB taxonomy,

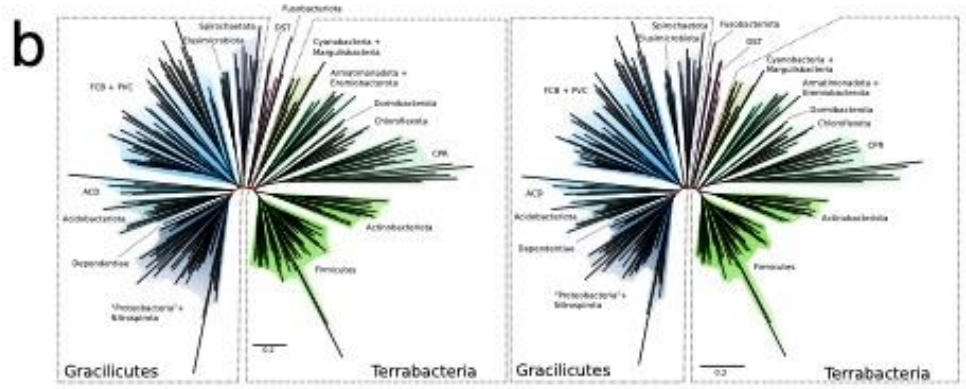
with all widely accepted phyla being resolved as monophyletic lineages, including the proposal that the Tenericutes branch within the Firmicutes (Davis *et al.*, 2013). Higher-level associations of phyla were also resolved, notably PVC (Wagner and Horn, 2006), FCB (Gupta, 2004), Cyanobacteria-Magulisbacteria (Anantharaman *et al.*, 2016), Chloroflexota-Dormibacterota (Ji *et al.*, 2017) and the CPR (Brown *et al.*, 2015).

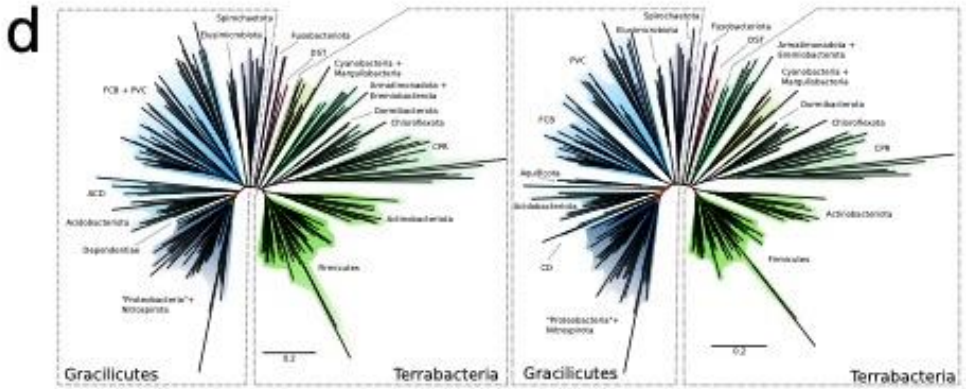
The largest stable groups in the unrooted tree were the Gracilicutes (Cavalier-Smith, 2006), comprising the majority of diderm lineages; and the Terrabacteria (Battistuzzi, Feijao and Hedges, 2004), which comprise some diderm lineages in addition to monoderm and atypical monoderm lineages, and which in our analyses include the CPR. The position of Fusobacteriota was unstable in the compositionally-stripped trees, either branching as in the focal tree (40%, 60%, 80% most compositionally heterogeneous sites removed) or with Deinococcota-Synergistota-Thermotogota (DST; 20% of sites removed). A clade comprising Aquificota, Campylobacterota and Deferribacterota (ACD) was recovered in three of the composition-stripped trees (20%, 40%, 60% most compositionally heterogeneous sites removed), but not in the focal tree or when 80% of the most compositionally heterogeneous sites were removed. Further, in the tree with 80% of the most compositionally heterogeneous sites removed, the clade comprising Aramtimonadota and Eremiobacterota, and the clade comprising Cyanobacteria and Margulisbacteria exchange places respectively.

Fig. 2.1 (below) Maximum likelihood unrooted bacterial phylogeny under the best-fitting substitution model (LG+C60+R8+F) for the full tree (a), and following the removal of the 20%-80% most compositionally heterogeneous sites (b-e). *(a) We used gene tree-species tree reconciliation to infer the root of the bacterial tree. The unrooted phylogeny was inferred from a concatenation of 63 marker genes under the best-fitting LG+C60+R8+F model, which accounts for site-heterogeneity in the substitution process and uses a mixture of 8 substitution rates estimated from the data to model across-site evolutionary rate variation. Branches are coloured according to bootstrap support value. Sites were identified and removed using Alignment Pruner. (b) 20% most compositionally heterogeneous removed, with 14580/18234 sites remaining following site stripping; (c) 40% most compositionally heterogeneous removed, with 10941/18234 sites remaining following site stripping; (d) 60% most compositionally heterogeneous removed, with 7294/18234 sites remaining following*

site stripping; (e) 80% most compositionally heterogeneous removed, with 3647/18234 sites remaining following site stripping; Branch supports are ultrafast bootstraps, branch lengths are proportional to the expected number of substitutions per site. FCB are the Fibrobacterota, Chlorobiota, Bacteroides, and related lineages; PVC are the Planctomycetes, Verrucomicrobia, Chlamydiae, and related lineages; FASSyT are Fusobacteriota, Aquificota, Synergystota, Spirochaetota and Thermotogota.


a
Gracilicutes
Spirochaetota
Fusobacteriota
PVC
Elusimicrobiota
Cyanobacteria +
Margulisbacteria
DST
Armatimonadota +
Eremiobacterota
FCB
Chloroflexota +
Dormibacterota
CPR
Dependentiae
Aquificota +
Campylobacterota +
Deferribacterota
Acidobacteriota
Actinobacteriota
Nitrospirota
Firmicutes
Myxococcota
>95% bootstrap support
90-95% bootstrap support
<90% bootstrap support
Bdellovibrionota
Desulfuromonadota +
Desulfobacterota
Proteobacteria
Terrabacteria
0.3


b
c
Gracilicutes
Terrabacteria
Gracilicutes
Terrabacteria

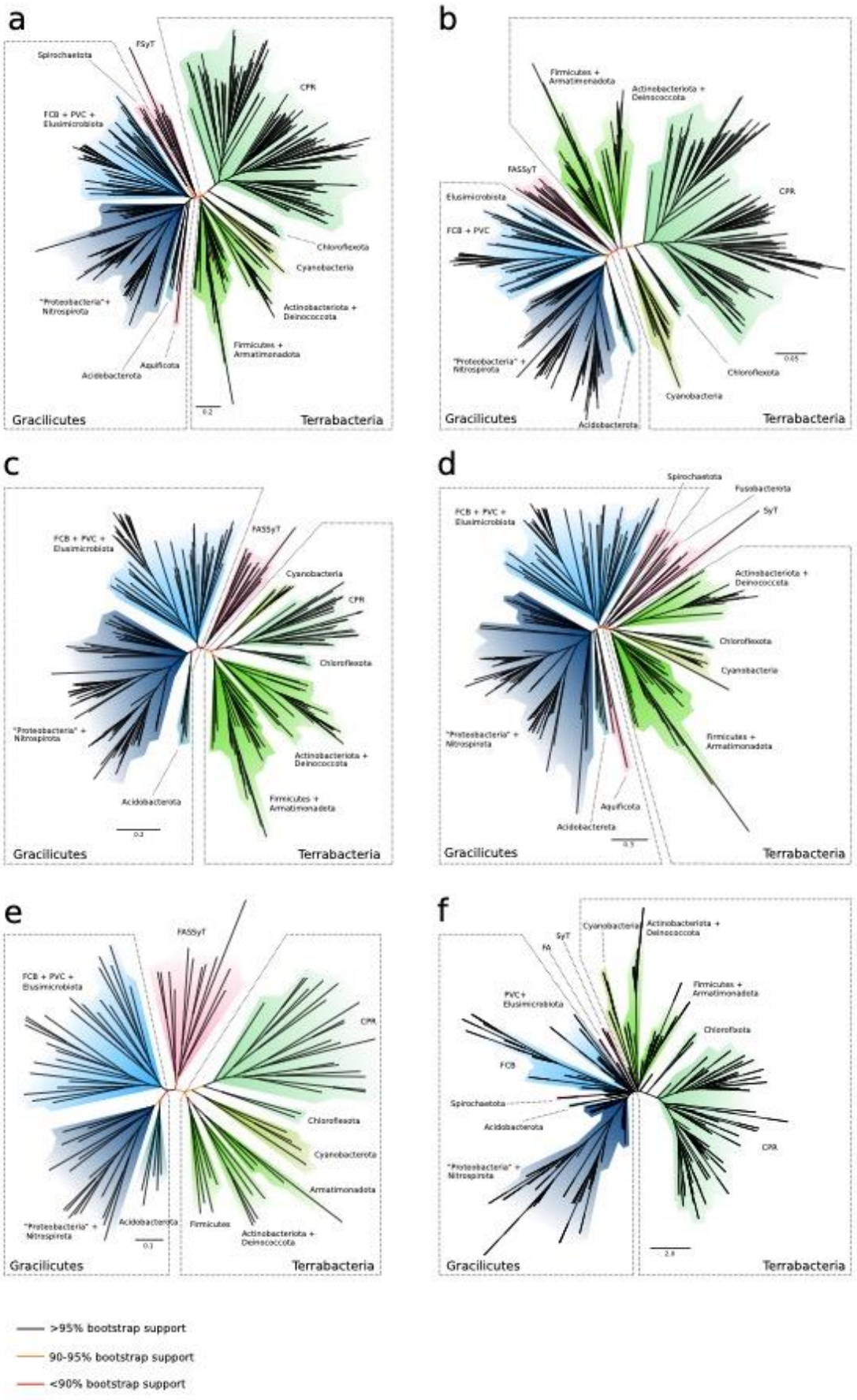

d
e
Gracilicutes
Terrabacteria
Gracilicutes
Terrabacteria

For the non-GTDB dataset, we inferred a tree in IQ-Tree using the LG+PMSF model, a model which accounts for across-site heterogeneity by assigning a conditional mean amino acid frequency profile for each site, which is calculated from a mixture model fitted to the data using a guide tree. This allows the model to mitigate possible LBA artefacts, while using fewer parameters than a full Bayesian CAT model, making it tractable for ML. In addition to the focal tree for this dataset, we also carried out a series of topological tests. First, we recorded our alignment using the four-category scheme of Susko and Roger (2007) and inferred a tree in PhyloBayes using the CAT+GTR+G4 model (Lartillot and Philippe, 2004), one of the best models for combating LBA. We further tested the impact of CPR taxon sampling by inferring LG+PMSF trees from alignments with the number of CPR reduced to 17 taxa, and with CPR removed completely. Additionally, we reduced the entire sample to 98 taxa and inferred another CAT+GTR+G4 tree in PhyloBayes. In order to explore a non-concatenation based approach, we also inferred multispecies coalescent (MSC) tree from the individually inferred trees of our 63 orthologues in ASTRAL (Zhang, Sayyari and Mirarab, 2017). Fig. 2.2 gives an overview of the results of these phylogenetic analyses.

All trees are broadly congruent with those recovered from the GTDB dataset. Major superphyla, such as the FCB, PVC and CPR are recovered, as are Terrabacteria and Gracilicutes. Deinococcota, which consistently falls within the “DTS" clade in trees derived from the GTDB dataset, is instead recovered as a sister clade to the Actinobacteriota in all trees. As in the GTDB dataset Fusobacteriota is unstable across the different trees, as are Aquificota, Spirochaetota, Synergystota, and Thermotogota. In three of the trees (recoded, reduced CPR sample, and reduced overall taxon sample concatenates respectively, Fig. 2.2(b,c,e)), these phyla are recovered as monophyletic (“FASSyT”). In the other trees, Synergistota and Thermotogota are always recovered as monophyletic, with Fusobacteriota being found either as a sister phylum (the unaltered concatenate, Fig. 2.2(a)), or on an adjacent branch (concatenate with CPR removed and the MSC tree Fig. 2.2(d,f)). Aquificota is found within the Gracilicutes in the unaltered concatenate and the concatenate with CPR removed (Fig. 2.2(a,d)), but found as a sister phylum to Fusobacteriota is the MSC tree (Fig. 2.2(f)).

We constrained the unaltered tree to match the topologies of the various trees and performed AU tests to attempt to distinguish between the different topologies. We additionally tested the monophyly of Terrabacteria with and without CPR, the validity of FASSyT, and a tree where all diderms and monoderms were monophyletic respectively. In all cases, the tree topology from the unaltered concatenate was supported, and other topologies rejected (AU p-value > 0.05), including no significant support for FASSyT as a clade. Indeed, when looking at the orthologous gene trees, none of them recover FASSyT monophyly. This suggests that some minor aspects of the tree topology may be affected by composition-driven LBA and taxon sampling. However, while there is some instability with regards to small phyla, the broad tree topologies are consistent across all datasets. Most notably, Gracilicutes and Terrabacteria are both stable across all trees, and the internal relationships between major phyla are largely consistent. This demonstrates that the key features of the topology are not due to composition-driven LBA artefacts, problems with concatenation, or problems with taxon sampling.

Fig. 2.2 (below) Unrooted bacterial phylogenies inferred from the GTDB-independent dataset. *(a) Maximum likelihood unrooted phylogeny inferred under the LG+PMSF model; (b) Bayesian unrooted phylogeny inferred under CAT+GTR+G4 model from a recoded concatenate using the four category scheme of Susko and Rogers (2007); (c) Maximum likelihood unrooted phylogeny inferred under the LG+PMSF model with a reduced CPR sampling; (d) Maximum likelihood unrooted phylogeny inferred under the LG+PMSF model with CPR removed; (e) Bayesian unrooted phylogeny inferred under CAT+GTR+G4 model from a reduce sampling of 98 taxa; (f) Unrooted phylogeny inferred under the multispecies coalescent (MSC) model in ASTRAL. FCB are the Fibrobacterota, Chlorobiota, Bacteroides, and related lineages; PVC are the Planctomycetes, Verrucomicrobia, Chlamydiae, and related lineages; FASSyT are Fusobacteriota, Aquificota, Synergystota, Spirochaetota and Thermotogota.*


a
Spirochaetota
FSyT
CPR
FCB + PVC + Elusimicrobiota
Chloroflexota
Cyanobacteria
"Proteobacteria" + Nitrospirota
Actinobacteriota + Deinococcota
Firmicutes + Armatimonadota
Aquificota
Acidobacterota
0.2
Gracilicutes
Terrabacteria
b
Firmicutes + Armatimonadota
Actinobacteriota + Deinococcota
FASSyT
CPR
Elusimicrobiota
FCB + PVC
"Proteobacteria" + Nitrospirota
Chloroflexota
Cyanobacteria
Acidobacterota
0.05
Gracilicutes
Terrabacteria
c
FCB + PVC + Elusimicrobiota
FASSyT
Cyanobacteria
CPR
Chloroflexota
"Proteobacteria" + Nitrospirota
Actinobacteriota + Deinococcota
Acidobacteriota
Firmicutes + Armatimonadeta
0.2
Gracilicutes
Terrabacteria
d
Spirochaetota
Fusobacteriota
FCB + PVC + Elusimicrobiota
SyT
Actinobacteriota + Deinococcota
Chloroflexota
Cyanobacteria
"Proteobacteria" + Nitrospirota
Firmicutes + Armatimonadota
Aquificota
Acidobacteriota
0.1
Gracilicutes
Terrabacteria
e
FASSyT
FCB + PVC + Elusimicrobiota
CPR
Chloroflexota
Cyanobacterota
Armatimonadota
"Proteobacteria" + Nitrospirota
Acidobacteriota
Firmicutes
Actinobacteriota + Deinococcota
0.1
Gracilicutes
Terrabacteria
f
Cyanobacteria
Actinobacteriota + Deinococcota
FA
SyT
Firmicutes + Armatimonadeta
PVC + Elusimicrobiota
Chloroflexota
FCB
Spirochaetota
Acidobacteriota
CPR
"Proteobacteria" + Nitrospirota
2.8
Gracilicutes
Terrabacteria
>95% bootstrap support
90-95% bootstrap support
<90% bootstrap support

Rooting the bacterial tree using outgroups

The standard approach to rooting is to include an outgroup in the analysis, and all published bacterial phylogenies in which CPR form a basal lineage (Hug *et al.*, 2016; Castelle and Banfield, 2018; Zhu *et al.*, 2019) have made use of an archaeal outgroup. Outgroup rooting on the bacterial tree, however, has three serious limitations. First, interpretation of the results requires the assumption that the root of the tree of life lies between Bacteria and Archaea. While this is certainly the consensus view, the available evidence is limited and difficult to interpret (Iwabe *et al.*, 1989; J. P. Gogarten *et al.*, 1989; Brown and Doolittle, 1995; Zhaxybayeva, Lapierre and Gogarten, 2005; Gouy, Baurain and Philippe, 2015), and alternative hypotheses in which the universal root is placed within Bacteria have been proposed on the basis of indels (Skophammer *et al.*, 2007; Lake *et al.*, 2009) or the analysis of slow-evolving characters (Cavalier-Smith, 2006). Second, the long branch leading to the archaeal outgroup has the potential to distort within-Bacteria relationships because of LBA. Third, joint analyses of Archaea and Bacteria are based on the smaller number of genes that are widely conserved and have evolved vertically since the divergence of the two lineages, and sequence alignment is more difficult because of the great evolutionary distance between the domains.

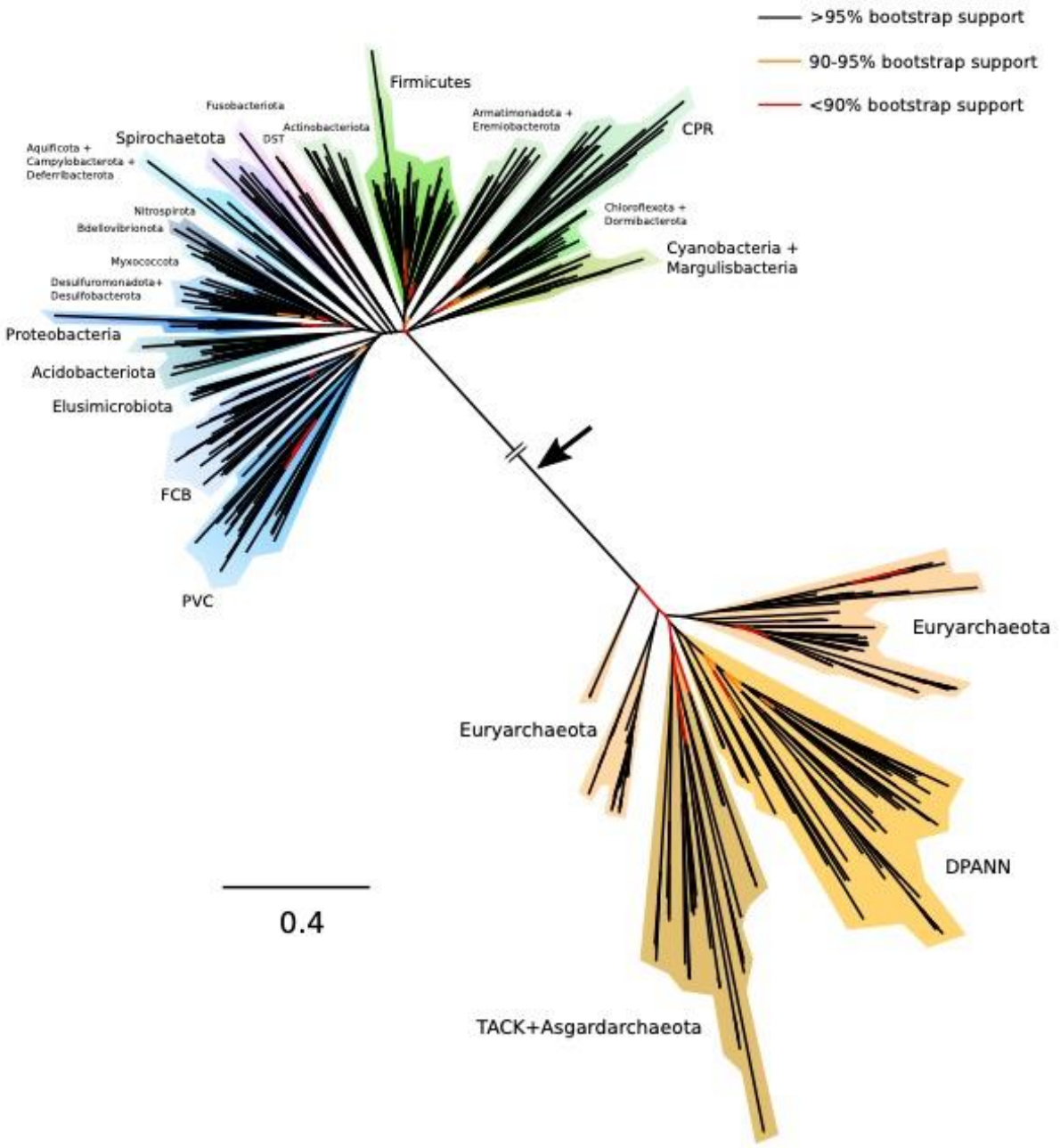
We began by evaluating the performance of outgroup rooting on the GTDB derived bacterial tree using 143 Archaea and a shared subset of 30 of our phylogenetic markers (Table 2.2). Using this archaeal outgroup, the ML phylogeny under the best-fitting model (LG+C60+R8+F, which accounts for site-heterogeneity in the substitution process) placed the bacterial root between a clade comprising Cyanobacteria+Margulisbacteria and CPR+Chloroflexota+Dormibacterota on one side of the root, and all other taxa on the other (Fig. 2.3). However, bootstrap support for this root, and indeed many other deep branches in both the bacterial and archaeal subtrees was low (50-80%). We therefore used AU tests to determine whether a range of published alternative rooting hypotheses (Table 2.4) could be rejected, given the model and data. The AU test asks whether the optimal trees that are consistent with these other hypotheses have a significantly worse likelihood score than the maximum likelihood tree. In this case, the likelihoods of all tested trees were statistically indistinguishable (AU > 0.05, Extended Data Table 2?). This indicates that outgroup rooting cannot resolve the bacterial root on this alignment of 30 conserved genes.

Root hypothesis	log-likelihood difference to ML	*p*-value	Study
Observed outgroup root (Fig. S2)	0	0.71	This study (ML tree)
Between Gracilicutes and Terrabacteria	-7.6	0.55	This study (ALE root, see below)
Thermotogota/Synergistota/Deinococcota basal*	-11.5	0.48	-
Chloroflexota basal	-11.6	0.46	Cavalier-Smith (2006)
Planctomycetes basal	-13.4	0.47	Brochier and Philippe (2002)
DPANN basal within archaeal outgroup	-19.9	0.41	(Castelle *et al.*, 2015; Williams *et al.*, 2017)
CPR basal	-20.4	0.35	(Hug *et al.*, 2016; Zhu *et al.*, 2019)
Between Firmicutes and Actinobacteriota	-26.8	0.36	Lake et al. (2009)
Fusobacteriota basal	-27.3	0.32	-

Table 2.4 *Support for published hypotheses using outgroup rooting. Our unrooted topology was incompatible with some published hypotheses, including a clade of Thermotogales and Aquificales at the root (Bocchetta et al., 2000; Battistuzzi and Hedges, 2009).*

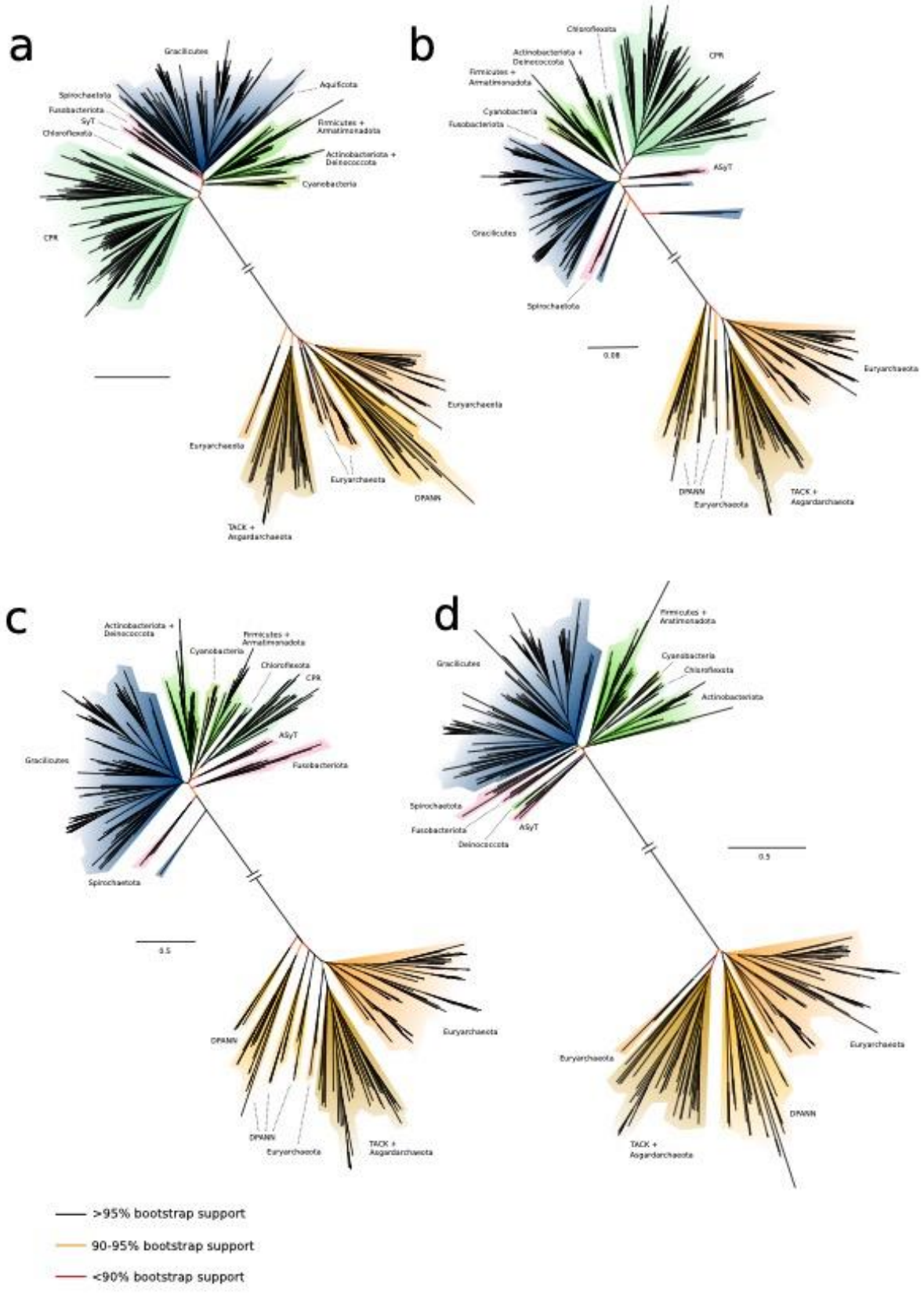
Fig. 2.3 (below) Maximum likelihood outgroup-rooted bacterial phylogeny from the GTDB dataset. *The maximum likelihood phylogeny obtained under the best-fitting LG+C60+R8+F model on a concatenation of 30 marker genes shared between Bacteria and Archaea. The bacterial root (marked by a black arrow) separates CPR, Cyanobacteria+Margulisbacteria, and Chloroflexota+Dormibacterota from the rest of the bacterial tree, but this position has poor bootstrap support and a range of alternative hypotheses could not be rejected statistically; note also that a basal position for DPANN within Archaea (Williams et al., 2017; Dombrowski et al., 2020) could not be rejected using an AU test (Table 2.4). FCB are the Fibrobacterota, Chlorobiota, Bacteroides, and related lineages; PVC are the Planctomycetes, Verrucomicrobia, Chlamydiae, and related lineages; DST are the Deinococcota, Synergistota, and Thermatogota; ACD are Aquificota, Campylobacterota, and*

Deferribacterota; Branch supports are ultrafast bootstraps, as indicated by the colour key. Branch lengths are proportional to the expected number of substitutions per site.



We performed outgroup rooting analyses on the non-GTDB dataset using the same 30 markers, producing an ML tree under the LG+PMSF model. The resulting tree placed the bacterial root between the CPR and all other Bacteria (Fig. 2.4a), replicating the results found in several other studies (Hug *et al.*, 2016; Parks *et al.*, 2017; Castelle and Banfield, 2018). We repeated this analysis recording our alignment using four Susko and Rogers groups and inferred a tree in PhyloBayes under a CAT+GTR+G4 model. In this case, the root was placed between the two environmental lineages with long branches, recovered in the Gracilicutes in the unrooted tree, and all other Bacteria (Fig. 2.4b), with Spirochaetota also being close to the root. We additionally carried out analyses testing the effect of taxon sampling. When reducing the number of CPR to 17, the root is similarly placed between some long branching environmental Gracilicutes and other Bacteria, with Spirochaetota being the next most basal lineage (Fig. 2.4c). When CPR are removed completely, the root is placed between a clade comprising Aquificota, Synergistota and Thermotogota on the one hand, and the rest of Bacteria on the other (Fig. 2.4d). The above analyses on both datasets demonstrate that outgroup rooting performs poorly when dealing with such ancient and divergent lineages. Outgroup rooting is sensitive to both composition-driven LBA and to taxon sampling. Previous studies which obtain a root between the CPR and all other Bacteria cannot be reliably replicated in our datasets and cannot be statistically distinguished from other root positions, raising the possibility that it may be artefactual.

Fig. 2.4 (below) Outgroup-rooted bacterial phylogenies form the GTDB-independent dataset. *Trees obtained from a concatenation of 30 marker genes shared between Bacteria and Archaea. (a) Maximum likelihood phylogeny obtained under the LG+PMSF model; (b) Bayesian phylogeny inferred under CAT+GTR+G4 model from a recoded concatenate using the four category scheme of Susko and Rogers (2007); (c) Maximum likelihood phylogeny inferred under the LG+PMSF model with a reduced CPR sampling; (d) Maximum likelihood phylogeny inferred under the LG+PMSF model with CPR removed. FCB are the Fibrobacterota, Chlorobiota, Bacteroides, and related lineages; PVC are the Planctomycetes, Verrucomicrobia, Chlamydiae, and related lineages; FASSyT are Fusobacteriota, Aquificota, Synergystota, Spirochaetota and Thermotogota.*

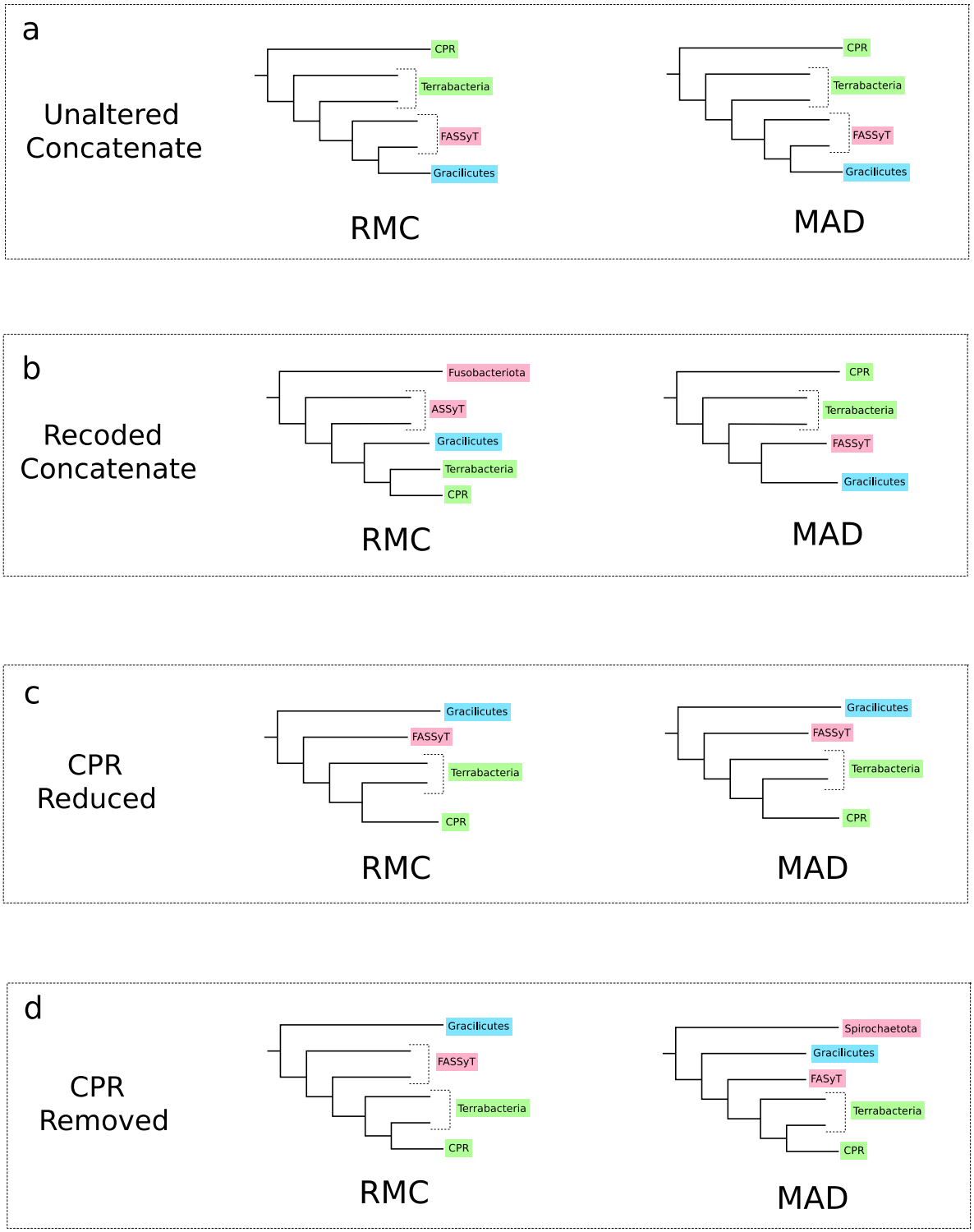


Attempting to root the bacterial tree without outgroups

We initially tried two rooting methods without outgroups, the relaxed molecular clock (RMC) in BEAST2 (Drummond and Rambaut, 2007; Drummond and Bouckaert, 2015), and minimal ancestor deviation (MAD) (Tria, Landan and Dagan, 2017) (see methods). For the RMC, we performed a lognormal uncorrelated relaxed molecular clock analysis, with the root posterior probabilities (PPs) averaged over the trees sampled during the Bayesian molecular clock analysis using RootAnnotator. When performed on the unaltered concatenate, we obtained a root position between CPR and the rest of Bacteria with a PP of 0.61 (Fig. 2.5(a)). However, several other rooted positions had comparable PPs. We performed the analysis on the recoded alignment, with CPR reduced, and with CPR removed. The recoded concatenate recovered a root between Fusobacteriota and the rest of Bacteria with a PP of 0.58 (Fig. 2.5(b)). A root between the Gracilicutes and Terrabacteria+FASSyT was found for both the concatenates with CPR reduced and removed respectively (PP = 0.66 and 0.55 respectively), although FASSyT is monophyletic in the former and paraphyletic in the later (Fig. 2.5(c-d)). In each case, the PPs were low, with several other root positions recovered with comparable PPs.

When rooted using MAD, the CPR-root is obtained from the tree on the unaltered and recoded concatenates (Fig. 2.5(a-b)). When CPR is reduced, the root is between the Gracilicutes on one side, and the Terrabacteria+FASSyT on the other (Fig. 2.5(c)). When CPR are removed, the root falls between the Spirochaetota and the rest of Bacteria (Fig. 2.5(d)). In all cases, the AI was high (>0.9), meaning that the best root position was not significantly better than the next best. These analyses demonstrate that, similar to outgroup rooting, the RMC and MAD are susceptible to artefacts due to LBA, or are sensitive to taxon sampling.

Fig. 2.5 (below) schematic representations of rooted bacterial trees under the lognormal uncorrelated relaxed molecular clock (RMC) using an LG substitution model, and minimal ancestor deviation (MAD) rooting. *(a) RMC and MAD roots for the unaltered concatenate; (b) RMC and MAD roots for the recoded concatenate using the four category scheme of Susko and Rogers (2007); (c) RMC and MAD roots with CPR reduced; (d) RMC and MAD roots with CPR removed. FASSyT are Fusobacteriota, Aquificota, Synergystota, Spirochaetota and Thermotogota.*



Whole-genome approaches to rooting

Given the limitations of the above methods to establish the root of the bacterial tree, we explored another outgroup-free rooting approach using gene tree-species tree reconciliation (David and Alm, 2011; Szöllosi *et al.*, 2012; Szöllősi *et al.*, 2013; Williams

et al., 2017). This approach has recently been applied to root the archaeal tree (Williams *et al.*, 2017), and similar approaches have been applied to investigate the root of eukaryotes (Katz *et al.*, 2012; Emms and Kelly, 2017) and to map and characterise whole genome duplications in plants (Zwaenepoel and Van de Peer, 2019). The method works by explaining the histories of individual gene families in the context of a shared species tree with a series of speciation, gene origination, duplication, transfer and loss events. Since these histories depend on the position of the root, reconciliation likelihoods can be used to estimate the most likely root, in what can be viewed as a genome-wide extension of the classical approach used to root the tree of life based on ancient gene duplications (Iwabe *et al.*, 1989; Johann Peter Gogarten *et al.*, 1989). In addition to leveraging genome-wide data, a further advantage is the ability to extract root signal from both gene duplications and transfers (Szöllosi *et al.*, 2012; Williams *et al.*, 2017). Amalgamated Likelihood Estimation (ALE) improves on earlier approaches by explicitly accounting for uncertainty in the gene tree topologies and in the events leading to those topologies, while also estimating rates of gene duplication, transfer and loss directly from the data (Szöllősi *et al.*, 2013). Simulations suggest that root inferences under ALE are robust to variation in taxon sampling and that the method finds the correct root even under high levels of gene transfer (Szöllosi *et al.*, 2012; Williams *et al.*, 2017), suggesting that the approach is appropriate for the problem at hand.

The ability of the ALEml_undated algorithm to infer the correct gene tree root in the presence of gene duplications, transfers and losses was previously investigated using simulations (Williams *et al.*, 2017). Briefly, gene families were simulated on a rooted species tree using a continuous-time ODTL process (that is, a more complex model of genome evolution than that implemented in ALEml_undated), and ALEml_undated was used to estimate the root from subsamples of the simulated families. The maximum likelihood root according to ALE was the correct root in 95/100 replicates, and the log likelihood of alternative roots decreased with nodal distance from the correct root (consistent with the pattern observed in our empirical data, see Fig. 2.7(c), see below). In the remaining 5 cases, the maximum likelihood root was one branch away from the true root. Analysis of empirical data suggested that ALE root inferences are robust to (that is, consistent across) subsets of the data that vary in terms of the rate of horizontal gene transfer or species representation in gene families (Williams *et*

al., 2017). These properties make the ALE approach appropriate for inferring the root of Bacteria.

More broadly, species tree-aware phylogenetic methods (such as ALE) have been shown to be of use in fixing gene tree errors (Szöllősi *et al.*, 2013), and for ancestral state (Williams *et al.*, 2017) and protein (Groussin *et al.*, 2015) inference. The additional power of these methods derives from the use of information in the species tree to decide between gene trees that are statistically equivalent from the point of view of the phylogenetic likelihood. Recently, a study of gene tree rooting performance suggested that a parsimony-based, species tree unaware DTL method (RANGER-DTL) provided more accurate gene tree root estimates than species tree-aware, probabilistic methods such as ALE and GeneRax (Wade *et al*., 2020). Gene tree rooting accuracy is not directly related to species tree rooting accuracy, because the information on the species tree root in ALE derives from finding the rooted species tree that maximises the sum of reconciliation likelihoods across gene families. We nevertheless decided to investigate, in order to understand which properties of the available methods contribute to, and detract from, rooting accuracy more generally.

To investigate, we re-analysed the data from Figure 5 of Morel *et al.* (2019), who simulated sequence alignments on known rooted gene trees. This setup differs from that of the original study (Wade *et al.*, 2020) in that, where possible, we start directly from the alignment and not from independently reconstructed gene trees. We chose this simulation setup because empirical analyses typically proceed from sequence alignments to inferred trees and then roots. We inferred rooted gene trees roots using ALE, GeneRax (Morel *et al.*, 2019) (a species-tree *aware* method that maximises a joint reconciliation and phylogenetic likelihood to infer rooted gene trees), TreeRecs (a species-tree aware method that implements a parsimony approach to DTL to infer rooted gene trees), RANGER-DTL (Bansal *et al.*, 2018) (a species-tree *unaware* method that implements a DTL parsimony approach to root input gene trees), MAD (Tria, Landan and Dagan, 2017) (a species-tree unaware method that roots input gene trees using minimal ancestor deviation). To quantify gene tree rooting accuracy, we plotted the rooted Robinsons-Fold (RF) score between the inferred and true rooted gene trees; this provides an interpretable measure of rooting accuracy even when the true root bipartition does not appear in the inferred gene tree, as demonstrated recently

(Wade *et al.*, 2020). The results (Fig. 2.6) indicate that probabilistic species tree-aware methods (GeneRax and ALE) provide the most accurate gene tree root inferences among the methods compared, even when the model used to infer the gene tree is misspecified, that is, when LG is used for simulation and WAG for inference. In particular, even when gene trees are inferred with a misspecified substitution model, ALE gene tree rooting (mean rooted-RF = 0.258) is significantly more accurate than RANGER-DTL (mean rooted-RF = 0.322, $p < 10^{-15}$ Welch Two Sample t-test) or MAD (mean rooted-RF = 0.306, $p < 10^{-10}$ Welch Two Sample t-test) when the latter methods are provided with input gene trees obtained using the true substitution model.

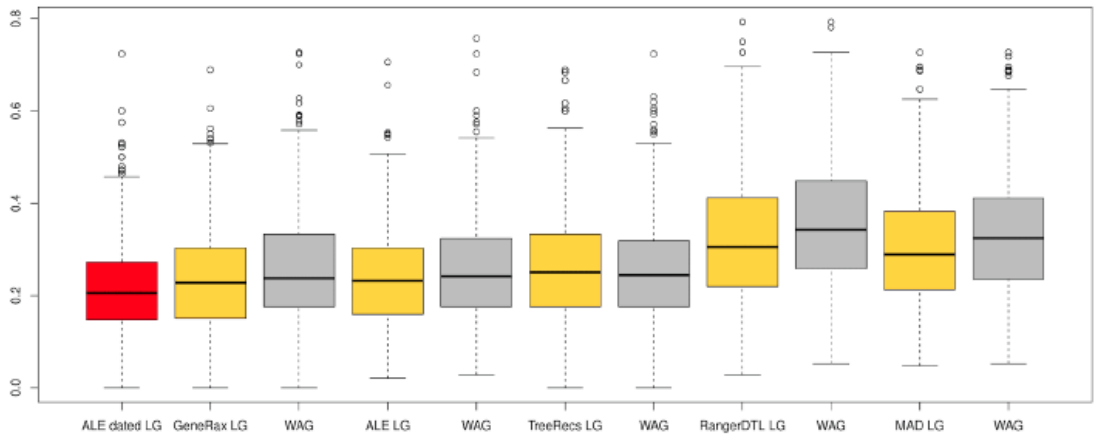


Fig. 2.6 *Accuracy of gene tree rooting methods. Species tree aware methods (GeneRax, ALE, TreeRecs) are the most accurate, even when gene trees are inferred with a misspecified (WAG) substitution model. Among species tree unaware methods, MAD outperforms the parsimony based DTL method RANGER-DTL.*

Rooting the bacterial tree using ALE

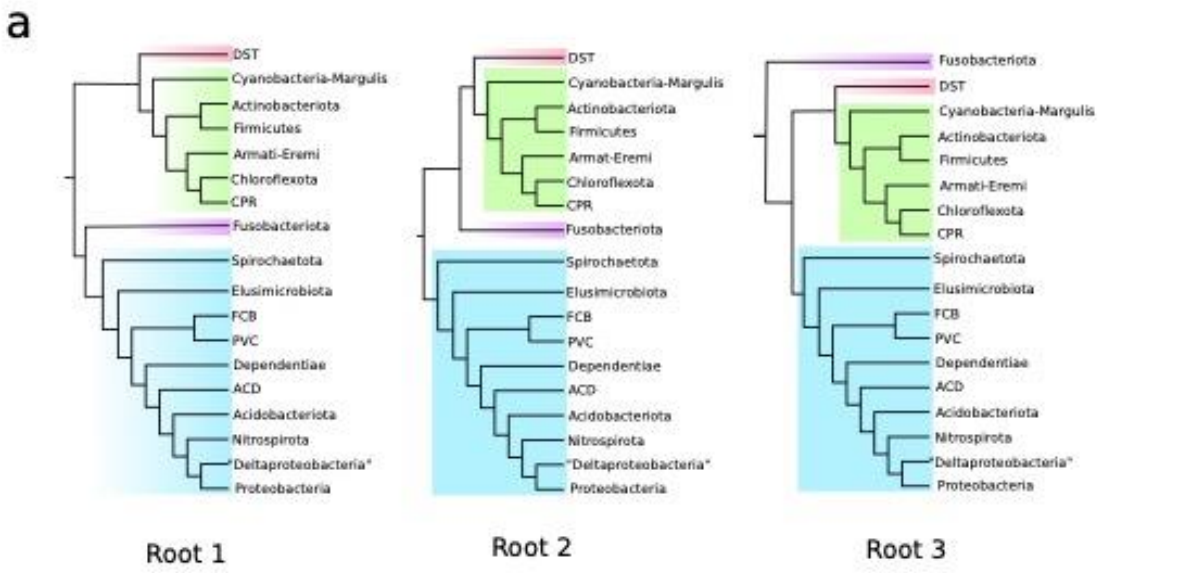
Given the ability of ALE to accurately infer rooted phylogenies, we used the method to test the support for 62 root positions on the unrooted topology derived from the GTDB by reconciling gene trees for 11,272 homologous gene families from the 265 bacterial genomes. In addition to testing root positions corresponding to published hypotheses (Table 2.5), we exhaustively tested all inner nodes of the tree above the phylum level. The ALE analysis rejected all of the roots tested ($P < 0.05$) except for three adjacent branches, lying between the two major clades of Gracilicutes and Terrabacteria (Fig. 2.7a). The position of the phylum Fusobacteriota was difficult to resolve in the tree,

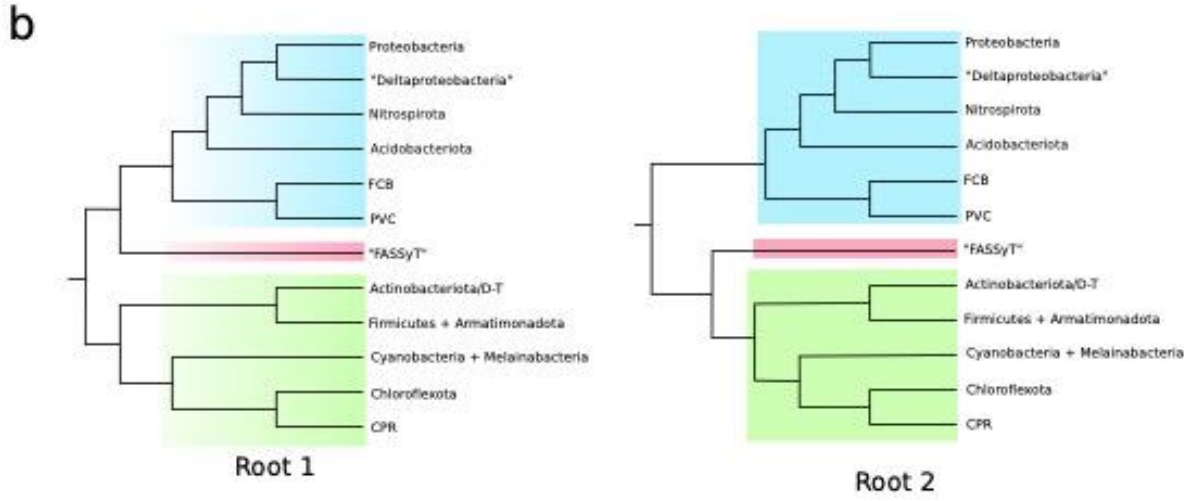
and contributed to root uncertainty. The three candidate root branches lead to (i) Terrabacteria+Deinococcus/Thermotoga/Synergistes; (ii) Gracilicutes; (iii) Fusobacteriota (Fig. 2.7a). Consistent with this being the optimal root region, alternative roots were rejected with increasing confidence as distance from the optimal region increased (Fig. 2.7c).

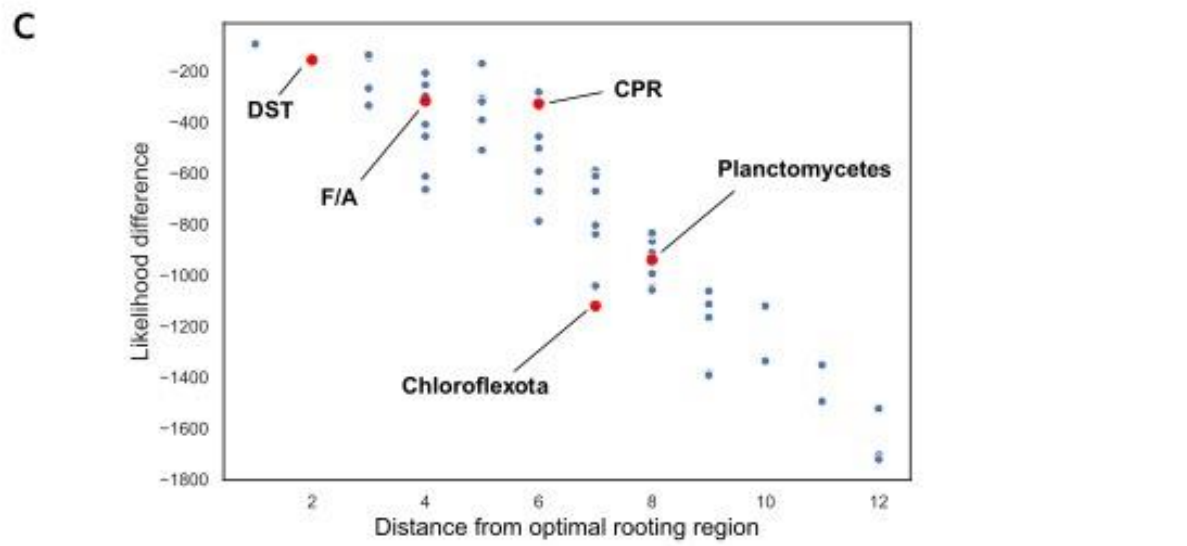
Root	*p*-value	Study
CPR basal	2e-04	Hug et al. (2016), Zhu et al. (2019)
Chloroflexota basal	1e-41	Cavalier-Smith (2006)
Between Firmicutes and Actinobacteriota	9e-05	Lake et al. (2009)
Thermotoga/Synergistota/Deinococcota basal*	0.004	
Planctomycetes basal	2e-26	Brochier and Philippe (2002)

Table 2.5 Support for published rooting hypotheses from our ALE analyses. **Our unrooted topology was incompatible with some published hypotheses, including a clade of Thermotogales and Aquificales at the root (Bocchetta et al., 2000; Battistuzzi and Hedges, 2009).*

Fig. 2.7 (below) Root positions determined by ALE for both the GTDB (a) and the GTDB-independent (b) datasets. *(a) Three rooted topologies from the GTDB dataset could not be rejected by the AU test. The root falls between two major clades of Bacteria, the Gracilicutes and the Terrabacteria, on one of three statistically equivalent adjacent branches. AU p-values are 0.476 for Root 1, 0.336 for Root 2 and 0.658 for Root 3. (b) Two rooted topologies from the GTDB-independent analysis that could not be rejected by the AU test, from ALE analysis incorporating genome completeness. AU p-values are 0.973 for Root 1 and 0.064 for Root 2. Both trees are in agreement with each other and GTDB analysis in placing the root between Terrabacteria and Gracilicutes, but disagree in the placement of the "FASSyT" taxa comprising Fusobacteriota, Aquificota, Synergistota, Spirochaetota and Thermotogota. (c) For the GTDB dataset, all tested alternative roots were rejected (Tables 2.3 and 2.4) with likelihoods decreasing as a function of distance from the root region. Previously proposed root positions, including the CPR root, are highlighted in red. D-T stands for Deinococcus-Thermus; "Deltaproteobacteria" is Desulfuromonadota, Desulfobacterota, Bdellovibrionota, and Myxococcota. FCB are the Fibrobacterota, Chlorobiota, Bacteroidota, and related lineages; PVC are the Planctomycetota, Verrucomicrobiota, Chlamydiota, and related lineages; DST are the Deinococcota, Synergistota and Thermotogota; ACD are Aquificota, Campylobacterota, and Deferribacterota; FA are Firmicutes and Actinobacteriota; FASSyT are Fusobacteriota, Aquificota, Synergystota, Spirochaetota and Thermotogota.*


a
DST
Cyanobacteria-Margulis
Actinobacteriota
Firmicutes
Armati-Eremi
Chloroflexota
CPR
Fusobacteriota
Spirochaetota
Elusimicrobiota
FCB
PVC
Dependentiae
ACD
Acidobacteriota
Nitrospirota
"Deltaproteobacteria"
Proteobacteria
Root 1
DST
Cyanobacteria-Margulis
Actinobacteriota
Firmicutes
Armat-Eremi
Chloroflexota
CPR
Fusobacteriota
Spirochaetota
Elusimicrobiota
FCB
PVC
Dependentiae
ACD
Acidobacteriota
Nitrospirota
"Deltaproteobacteria"
Proteobacteria
Root 2
Fusobacteriota
DST
Cyanobacteria-Margulis
Actinobacteriota
Firmicutes
Armati-Eremi
Chloroflexota
CPR
Spirochaetota
Elusimicrobiota
FCB
PVC
Dependentiae
ACD
Acidobacteriota
Nitrospirota
"Deltaproteobacteria"
Proteobacteria
Root 3


b
Proteobacteria
"Deltaproteobacteria"
Nitrospirota
Acidobacteriota
FCB
PVC
"FASSyT"
Actinobacteriota/D-T
Firmicutes + Armatimonadota
Cyanobacteria + Melainabacteria
Chloroflexota
CPR
Root 1
Proteobacteria
"Deltaproteobacteria"
Nitrospirota
Acidobacteriota
FCB
PVC
"FASSyT"
Actinobacteriota/D-T
Firmicutes + Armatimonadota
Cyanobacteria + Melainabacteria
Chloroflexota
CPR
Root 2


c
DST
CPR
F/A
Planctomycetes
Chloroflexota
Likelihood difference
−200
−400
−600
−800
−1000
−1200
−1400
−1600
−1800
2
4
6
8
10
12
Distance from optimal rooting region

On the non-GTDB dataset, we calculated the gene family likelihoods of 11,781 homologous gene families derived from our 342 bacterial genomes, under a set of 60 candidate root positions on different unrooted species topologies obtained from different tree construction methods (see Methods). Of the 60 number of rooted phylogenies tested, all but two were rejected (Fig. 2.7b) ($P < 0.05$). Both of these topologies had candidate root branches between the Gracilicutes and the Terrabacteria, as in the GTBD analysis. They also both recovered monophyletic FASSyT, although its position could not be resolved, lying on either side of the root. The two candidate root branches therefore lead to (i) Terrabacteria; (ii) Gracilicutes. In order to further mitigate against LBA caused by CPR, we additionally performed ALE (including the inference of the single gene trees following the same pipeline) with the CPR removed. The resulting root position was identical to that with the CPR included. CPR therefore does not seem to affect the ALE rooting results.

MCL families represent, within the limitations of the clustering approach discussed above, an unbiased view of gene family diversity for the set of genomes we analysed. We therefore base all our analyses, except those regarding the functional annotation of LBCA, on the MCL families. By contrast, our COG families are useful for functional reconstruction (see Chapter 3 for details), but are perhaps less well suited for investigating other aspects of bacterial evolution because they are constructed only from proteins that could be annotated with eggNOG-mapper. However, since gene clustering methods are not consummate and each has strengths and weaknesses, we also investigated the root signal from the COG families (see Chapter 3 for details on the construction of the COG families). This analysis resulted in a root region of four adjacent branches, comprising the root region from the MCL analysis (3 branches) plus one additional branch, in which Spirochaetota branched on the Terrabacteria side of the root (Table 2.6). This slightly expanded root region is likely due to the reduced resolution of the smaller set of COG families in comparison to the full analysis.

Root name	LLs	AU
Fusobacteriota root (398)	-7.3	0.589
Fusobacteriota on Terrabacteria side (527)	7.3	0.519
Fusobacteriota on Gracilicutes side (528)	13.6	0.432

Fusobacteriota and Spirochaetota on Terrabacteria side (520)	32.1	0.251
DST root (464)	103.7	0.008
Cyanobacteria on Gracilicutes side (517)	215.8	1e-09
Dormibacterota/Chloroflexota+CPR (510)	372.2	7e-06
CPR root (496)	425.2	2e-67
Dormibacterota/Chloroflexota (505)	630.7	5e-05
Omnitrophota/Verrucomicrobia/Planctomycetes (492)	674.4	2e-04
Fibrobacteria/Bacteroidetes/Marinisomatota (511)	1000.7	3e-71

Table 2.6 AU-test results for an ALE root analysis using 3595 COG families.

While it has not been possible to completely resolve the root, all analyses recover a root region between Gracilicutes and Terrabacteria. A similar root was previously reported (Raymann, Brochier-Armanet and Gribaldo, 2015; Adam, Borrel and Gribaldo, 2018). However, this analysis did not include the CPR, which has been recently suggested (Hug *et al.*, 2016; Zhu *et al.*, 2019) to represent the earliest diverging bacterial lineage. Further, our ALE analyses consistently recovered CPR nested within the Terrabacteria, suggesting that the CPR root is a long branch attraction artefact.

Is bacterial evolution treelike?

How much of bacterial evolution can be explained by the concept of a rooted species tree? Horizontal gene transfer (HGT) is frequent in prokaryotes, and published analyses indicate that most or all prokaryotic gene families have experienced HGT during their history (Dagan and Martin, 2007; Williams *et al.*, 2017). This implies that there is no single tree that fully describes the evolution of all bacterial genes or genomes (Doolittle, 1999; Doolittle and Bapteste, 2007). Extensive HGT is existentially challenging for concatenation, because it greatly curtails the number of genes that evolve on a single underlying tree (Dagan and Martin, 2006). Phylogenetic networks

(Doolittle and Bapteste, 2007; Alvarez-Ponce *et al.*, 2013) were the first methods to explicitly acknowledge non-vertical evolution, but can be difficult to interpret biologically. Gene tree-species tree reconciliation integrates tree and network-based approaches by modelling both the horizontal components of genome evolution (a fully reticulated network allowing all possible transfers) and the vertical trace (a common rooted species tree). This framework enables us to quantify the contributions of vertical and horizontal processes to bacterial evolutionary history.

Our analyses (Fig. 2.8) reveal that most bacterial gene families present in at least two species (9678/10518 MCL families, 92%) have undergone at least one gene transfer during their evolution; only very small families have escaped transfer entirely (Fig. 2.9). Consistent with previous analyses (Jain, Rivera and Lake, 1999; Williams *et al.*, 2017), transfer rates vary across gene functional categories, with genes functioning in defence mechanisms (such as antibiotic biosynthesis) and the production of secondary metabolites being the most frequently transferred, and those involved in translation and the cell cycle the least (Fig. 2.8(b), Table 2.2). Despite this accumulation of HGT, most gene families evolve vertically the majority of the time, in that 66% of transmission events (mean, MCL gene families) seem to evolve vertical along the species tree.

There are a number of caveats that should be considered when interpreting the results of these analyses. Our inference of HGT events is likely to be an underestimate, because our broad but sparse taxon sampling does not allow us to detect transfers or recombination within strains. Similarly, orthologous replacement may resemble vertical transmission and thus go undetected, leading to a further underestimate of transfer events. Additionally, we do not specifically consider pseudogene in this analysis. Pseudogenes are segments of non-functional DNA which resemble functional genes, often caused by loss of gene function, and may cause us to overestimate transfer events. However, while not modelled explicitly, pseudogenes should be modelled as a form of gene loss in ALE, and therefore should not have a profound effect on the analyses, unless they make up large part of the genomes analysed. We also note that the inclusion of plasmids in our dataset may also affect transfer rates as plasmids have are often involved in HGT (Gauri *et al.* 1994; Varga *et al.* 2016).

While we demonstrate a high proportion of vertical signal, which implies that a tree may be a suitable way to describe bacterial evolution, we still detect a large amount of HGT. This poses an important question as to how much transfer is necessary before a tree is no longer an apt description of bacterial evolution. Furthermore, the directionality of the transfers may have an impact on our ability to infer a species tree. If such HGTs are largely random in nature, then a tree is still likely recoverable. However, if the transfers are highly directional in nature, such directional biases may negatively impact our tree inferences. Detecting the directionality in our dataset in difficult, and we have not been able to do it here. However, while it may be difficult to detect the directionality, we may use simulations to determine the effect that both the levels and directionality of HGTs have on our ability to recover a tree. By performing simulations where the level and directionality of HTGs within the simulated gene trees would vary, we could explore the extent to which a tree is recoverable, and discern the point at which the levels HGT and directionality were so great that a tree is longer inferable. This would give us a sense of what percentage of horizontal signal could be considered so pervasive that a tree is no longer adequate to describe the data in question.

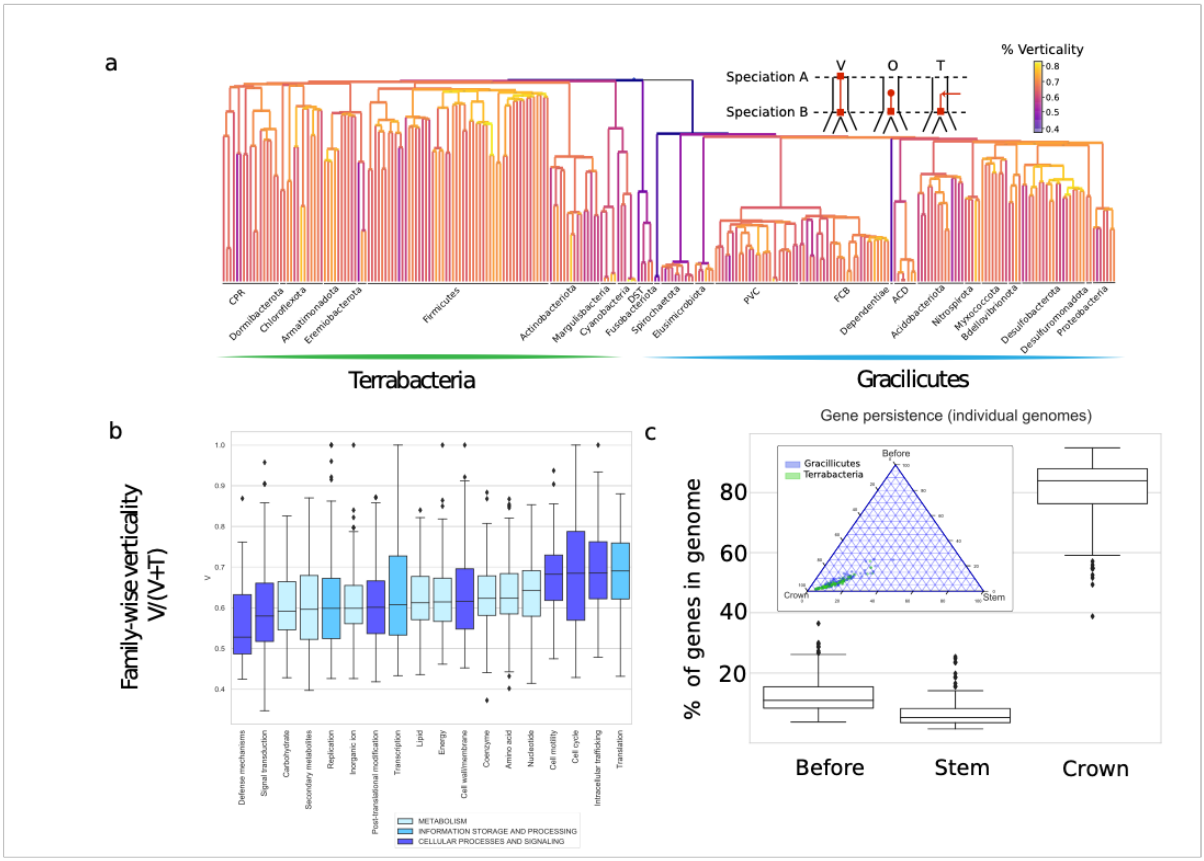


Fig. 2.8 The verticality of bacterial genome evolution. *(a) The rooted bacterial species tree (Fig. 2.7) with branches coloured according to verticality: the fraction of genes at the bottom of a branch that descend vertically from the top of that branch (see inset; V = vertical, O = origination, T = transfer into a branch; see Methods). Node heights reflect relative time order consistent with highly-supported gene transfers. (b) Verticality by COG functional category: that is, the proportion of gene tree branches that are vertical V/V+T for COG gene families. Genes involved in information processing, particularly translation (J), show the highest verticality (median 0.69), while genes involved in cell defence mechanisms (V, such as genes involved in antibiotic defence and biosynthesis) are most frequently transferred. (c) For a given genome, this combination of vertical and horizontal processes gives rise to a distribution of gene residence times, reflecting how far back in bacterial history genes are retained. Across all phyla examined, 82% of genes on sampled genomes trace back to the crown group radiation of that phylum. FCB are the Fibrobacterota, Chlorobiota, Bacteroidota, and related lineages; PVC are the Planctomycetota, Verrucomicrobiota, Chlamydiota, and related lineages; DST are the Deinococcota,*

Synergistota, and Thermotogota; ACD are Aquificota, Campylobacterota, and Deferribacterota.

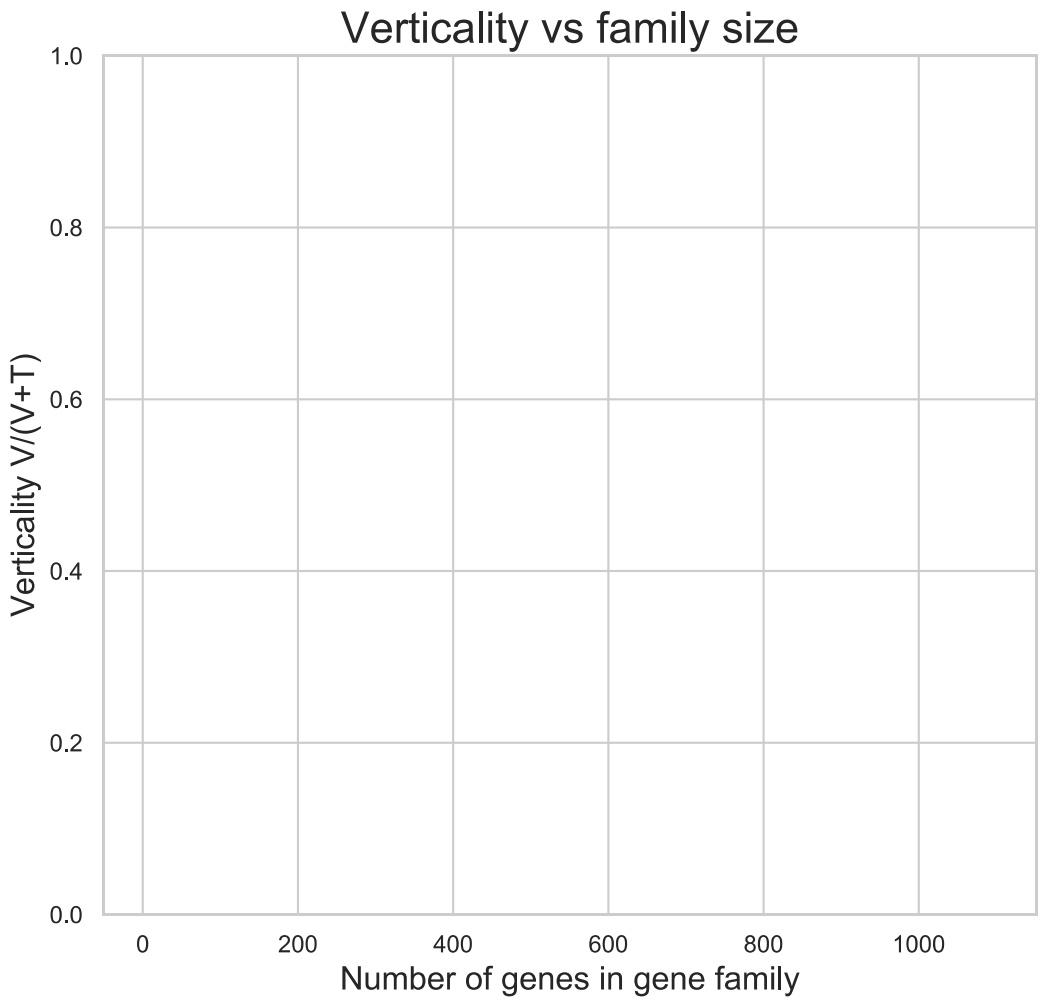


Fig. 2.9 The relationship between verticality and gene family size. *Most gene families have experienced many transfers. Verticality varies with gene functional class, but families with very low transfer rates are small; these might represent young families that have not yet had enough time to experience gene transfer.*

Mapping the branches of the gene trees onto the species tree demonstrates that the optimal tree provides an apt summary of much of bacterial evolutionary history, even for the deepest branches of the tree (Creevey *et al.*, 2004; Koonin, Wolf and Puigbò, 2009; Puigbò, Wolf and Koonin, 2010). From the gene's eye view, gene families evolve neither entirely vertically nor horizontally: core genes are occasionally transferred, and even frequently exchanged genes contribute useful vertical signal; for example, the

median number of genes that evolve vertically on a branch of the species tree is close to 998.92 (Table 2.7), far greater than the number of genes that have been concatenated at the level of all Bacteria. From the perspective of the genome, constituent genes have different ages, corresponding to the time at which they originated or were most recently acquired by gene transfer, within the resolution of our taxonomic sampling. This analysis indicates that, on average, 82% of the genes on all genomes from adequately represented phyla (5 or more genomes) trace back to the crown group radiation of that phylum, though all genomes retain a smaller proportion (10.3-26.7%) of genes that have descended vertically from the stem lineage of their phylum or even earlier (Figure 2.8(c)).

Root branch	1	2	3
Median singleton support per branch	999.09	998.47	999.21
Singleton support for branches subtending the root	98.259, 140.45	91.56, 151.09	95.25,117.16
Mean verticality	0.68136	0.68137	0.6818

Table 2.7 Singleton support (the number of genes that evolve vertically from one end of a branch to the other) on the credible set of rooted trees. *Root numbers correspond to the three root branches depicted in Fig. 2.7(a)).*

Caveats of ALE

While we believe that ALE is a robust method, and an improvement on previous analyses, there are some caveats of the method which must be considered. In its current implementation, a two-step process is used, whereby gene trees must be must be independently inferred using species-tree unaware methods, with the conditional clade probabilities being calculated for a sample of trees (e.g. posterior sample from an MCMC, or a bootstrap sample) and reconciled with the species tree to create a gene tree amalgamated from clades present in the gene tree sample. Ideally, however, the inference of gene trees would happen jointly with the reconciliation and therefore be specie-tree aware.

Another caveat relates to the manner in which we infer the clusters which are interpreted as gene trees. As we briefly discuss above, MCL clustering has no biological basis, and we must accept a trade-off between inferring clusters that are too liberal, and therefore contain distant or unrelated sequences, and those that are too conservative and break-up larger gene families. As previously discussed, we have attempted to partially address these issues by testing different inflection parameters in MLC (which controls the size of the clusters) and by performing analyses using COGs (see details above). However, more extensive analyses could be done to test the effect of gene clustering; for example, repeating the full analyses using different inflation parameters (although this would be costly and time-consuming), or by repeating the analyses using a different clustering approach, e.g. using HI Fix (Miele *et al.*, 2012). This is very an important area to explore further given the extent to which our analyses rely on the accurate creation of gene families.

While we believe ALE to be robust to taxon sampling, further exploration may be needed to determine the extent of the effects on our results. We have partially done this in the GTDB-independent dataset by not only increasing the number of taxa, but by sampling diversity differently, leading to oversampling of some phyla (e.g. the CPR) and under sampling of others (e.g. the Firmicutes) with respects to the focal dataset, and have obtained similar results in both cases. However, multiple replicated datasets would be needed to test this further. Furthermore, as we note above, the sparse taxonomic sampling could be leading to an underestimate of transfers, as we cannot detect transfers or recombinations within closely related lineages. Increasing taxon sample size could alleviate this problem, although would be computationally expensive.

A further caveat is that extreme ratios of DTLs (i.e. extreme values of D/T, D/L, T/D, T/L, L/D or L/T) within the DTL model used by ALE, particularly in small gene families, could be affecting the root positions recovered in the analysis. A possible way to explore this would be to plot all DTL rate ratios and progressively remove the most extreme gene families to see how it would affect the analyses. This would be analogous to the removal of the most compositionally heterogeneouss sites in an alignment.

Is the universal root in Bacteria?

As noted above, while the consensus view is that the root of the tree of life lies between Archaea and Bacteria, the evidence is limited and can be hard to interpret (Iwabe *et al.*, 1989; J. P. Gogarten *et al.*, 1989; Brown and Doolittle, 1995; Zhaxybayeva, Lapierre and Gogarten, 2005; Gouy, Baurain and Philippe, 2015), and alternative hypotheses in which the universal root is placed within Bacteria have been proposed (Cavalier-Smith, 2006; Skophammer *et al.*, 2007; Lake *et al.*, 2009). This may therefore raise questions about the validity of our analyses. However, a strength of the ALE method used here is that it does not need an outgroup. Thus, if Archaea do indeed branch within the Bacteria, then the root of Bacteria presented here would in fact be the universal root of all life. In this scenario, Archaea would be among the number of groups we did not sample in our dataset, due to necessarily needing a small enough dataset to be tractable. Thus, our analyses are compatible with a bacterial root of life. Nonetheless, bacterial phyla not included in our analyses were small, while the omission of Archaea would represent a much greater loss of data. We therefore conducted a preliminary analysis where we rooted the tree of life, including the Archaea, using ALE.

We used 293 species, 149 Bacteria derived from non-GTDB bacterial dataset, and 144 Archaea from (Kellner *et al.*, 2018). These were selected evenly from across the tree in order to capture the full diversity. We did not include the eukaryotes as it is widely accepted that they are derived from an endosymbiotic event between Bacteria and Archaea (Embley and Martin, 2006; Martin, Garg and Zimorski, 2015; López-García, Eme and Moreira, 2017; Roger, Muñoz-Gómez and Kamikawa, 2017; Eme *et al.*, 2018), and that the root is almost certainly within the prokaryotes. As we wish primarily to explore possibilities for the root position within the tree of life, and not to elucidate anything concerning the origins or phylogenetic position of Eukarya, we judged it not essential that they be included in the analysis, especially as the long branch leading to the eukaryotes could be problematic. The genomes for the species were downloaded from NCBI GeneBank.

We concatenated 51 orthologues and inferred a tree under the LG+C60 model in IQ-Tree. Within the Archaea, Euryarchaeota and TACK were recovered. DPANN is not monophyletic in this tree, forming successive outgroups to the other Archaea. This

may be due to LBA, as has been shown in other cases (Gouy, Baurain and Philippe, 2015; Williams *et al.*, 2017). Within Bacteria, the topology broadly follows that of the previous analyses. However, “FASSyT" is not monophyletic, with Aquificota and Thermotogota being sister to Terrabacteria/CPR, and Spirochaetota and Fusobacteriota being recovered within Gracilicutes. Synergistota, along with a few Terrabacteria, have been pulled to the base of the bacterial clade. Due to the long branches between Bacteria and Archaea, some of these topological incongruencies may be due to LBA. Interestingly, the CPR are found within Terrabacteria, and not close to the branch leading to the Archaea. This is further evidence that tree topologies where CPR are basal or near the branch to Archaea (Hug *et al.*, 2016; Parks *et al.*, 2017; Castelle and Banfield, 2018) are likely due to topological artefacts.

In order to test root positions using ALE, homologous gene families with paralogues were generated using HiFix (Miele *et al.*, 2012). The sequences were aligned in Mafft (using the auto option) and trimmed in BMGE (using BLOSUM30), and trees of these gene families were inferred in IQ-Tree under the LG+C60 model. 10 root positions were tested. All roots could be rejected except for the root between Archaea and Bacteria (Fig 2.10). As we have demonstrated that ALE seems to not be affected by LBA artefacts, it is likely that the root of life is between Archaea and Bacteria. However, it must be noted that a much more thorough analysis of the data, with extensive topology testing, use of different taxon samplings, and the testing of a much wider range of roots, would be needed to bring confidence to this result.

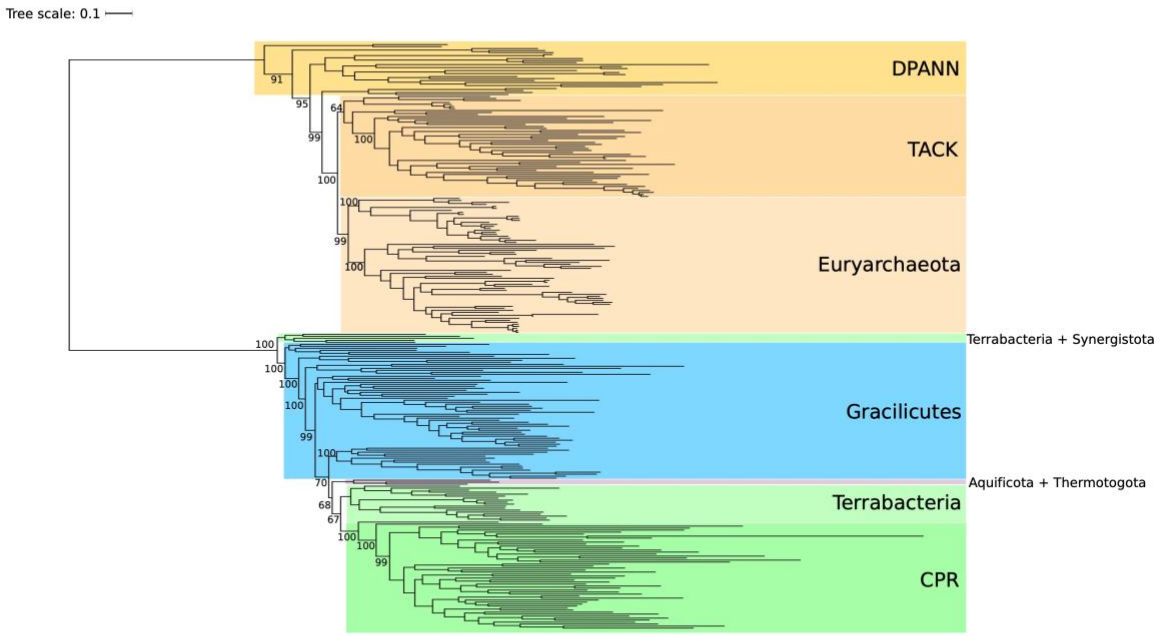


Fig. 2.10 Rooted phylogeny of Archaea and Bacteria. *The maximum likelihood phylogeny obtained under the LG+C60 model on a concatenation of 51 marker genes shared between Bacteria and Archaea. The root position between Bacteria and Archaea was the only root position tested in ALE and could not be rejected by the AU test (P < 0.05).*

2.4 Conclusions

We place the last bacterial common ancestor between two major clades, Terrabacteria and Gracilicutes, although we could not resolve the position of several smaller lineages, notably the Fusobacteriota, in relation to those major radiations. Fusobacteriota currently comprise anaerobic free-living, pathogenic and commensal diderm bacteria (Brennan and Garrett, 2019), and a clear direction for future work will

be to place them on the rooted bacterial tree, particularly if more basal members of this lineage come to light. We have found these results to be consistent across different datasets, demonstrating that, while there is still some lack of resolution in the deepest parts of the tree, there is strong and consistent signal supporting the higher level relationships we have demonstrated here, and that key features of the topology are not composition-driven LBA artefacts or caused by taxon sampling.

In contrast to recent outgroup-rooted analyses (Hug *et al.*, 2016; Parks *et al.*, 2017; Castelle and Banfield, 2018; Zhu *et al.*, 2019), we found no support for a root on the CPR branch; instead, our analysis suggests that this lineage evolved within Terrabacteria, from a common ancestor with Chloroflexota. Analyses which recover CPR in a basal position are likely artefacts, as their position changes when varying the taxon sampling, or using recoding to account for compositional heterogeneity and saturation. Furthermore, we demonstrate that outgroup rooting, at least when applied in this case where the distance to the outgroup is great, is unable to distinguish between different root positions and is highly susceptible to LBA artefacts. Other methods, including the relaxed molecular clock and MAD rooting are similarly affected by these issues. Probabilistic species tree-aware methods, such as ALE, are less susceptible to LBA artefacts and differences in taxon sampling, can utilise a greater range of data, and seem to give accurate results compared to other rooting methods. Our analyses further suggest that, despite extensive horizontal gene transfer, a phylogenetic tree is an apt representation of bacterial evolution in the sense that most bacterial gene families evolve vertically most of the time.

Chapter 3

Ancestral reconstruction of the last bacterial common ancestor

A version of this chapter forms part of a paper under revision in collaboration with Adrián A. Davín, Tara Mahendrarajah, Anja Spang, Philip Hugenholtz, Gergely J. Szöllősi, and Tom A. Williams. Gareth A. Coleman is the first author of the paper. The project was conceived by TAW, GJSz, PH, AS, GAC and AAD. Protein annotations and generation of COG families were carried out by TM, AS and TAW. COG reconciliations carried out by GJSz. Metabolic reconstruction, including metabolic maps and figures, were carried out and produced by GAC. Writing and interpretation of metabolic reconstructions were carried out by GAC, AS and TM.

Paper preprint as:
Coleman, G.A., Davín, A.A., Mahendrarajah, T., Spang, A.A., Hugenholtz, P., Szöllősi, G.J. and Williams, T.A., 2020. A rooted phylogeny resolves early bacterial evolution. *bioRxiv*.

BioRxiv preprint for the paper can be found here:
https://www.biorxiv.org/content/10.1101/2020.07.15.205187v1

Abstract

Bacteria are the most abundant and metabolically diverse cellular lifeforms on Earth and have had a profound impact on the physical environment. To understand the evolution of complex interaction between the biosphere and geosphere, we must understand the nature of the earliest bacterial cells, including the last bacterial common ancestor (LBCA). Building upon previous work, we use a rooted phylogeny inferred from the modelling of genome evolution at the level of gene duplication, transfer and loss events to reconstruct the ancestral gene content of LBCA. We infer that the LBCA possessed core carbon metabolic pathways, including glycolysis, the reverse tricarboxylic acid cycle and the pentose phosphate pathway, with the possibility of fixing carbon via either the Wood-Ljungdahl or reverse tricarboxylic acid cycles. In addition, we predict that LBCA was a free-living flagellated, rod-shaped cell featuring a double membrane with a lipopolysaccharide outer layer, bacterial phospholipids, a Type III CRISPR-Cas system, Type IV pili, and the ability to sense and respond via chemotaxis.

3.1 Introduction

Bacteria inhabit almost all known habitats and ecosystems, and employ huge diversity of physiologies to adapt to these diverse environments. As such, they perform vital roles in biogeochemical cycles and have had a profound impact on the physical Earth throughout history. To understand how such systems, both biological and biogeochemical, have evolved we must answer questions pertaining to the physiology and habitat of the earliest life, including that of the last bacterial common ancestor (LBCA). In particular, how complex or “modern” early prokaryotic cells were in comparison to present day cells, what kind of environment they lived in, and how these cells derived energy from their environment.

One of the central questions involves discerning the core carbon metabolism present in LBCA, specifically to determine if LBCA had the ability to fix carbon, and if so which pathway it would have used. Modern bacteria fix carbon using several different pathways, including the Calvin cycle, the 3-hydroxypropionate (3-HP) bicycle and variations thereof, the Wood-Ljungdahl pathway (WLP) and the reverse tricarboxylic acid (TCA) cycle, the latter two of which have been suggested to have emerged early in the history of life. The WLP, found in both methanogenic archaea and acetogenic bacteria, is thought to be one of the most ancient carbon fixation pathways on the basis of both biogeochemical and phylogenetic arguments (Fuchs, 2011; Sousa and Martin, 2014; Weiss *et al.*, 2016; Williams *et al.*, 2017; Adam, Borrel and Gribaldo, 2018). Notably, the key enzyme complex of the pathway, CODH/ACS (CO dehydrogenase/acetyl-CoA synthase), is conserved in both domains, and is predicted to have been present in both the archaeal (Williams *et al.*, 2017; Adam, Borrel and Gribaldo, 2018) and bacterial common ancestors (Adam, Borrel and Gribaldo, 2018). The reverse TCA cycle has also been suggested as a possible ancient carbon fixation pathway (Wächtershäuser, 1990; Cody *et al.*, 2001; Smith and Morowitz, 2004; Nunoura *et al.*, 2018), given the widespread presence of the TCA cycle in modern Bacteria, and that it may function in both the oxidative and reductive direction. Some combinations of pathways have been suggested, notably a coupling of both the TCA cycle and the 3-HP bicycle (Marakushev and Belonogova, 2011, 2013).

In addition to the central carbon pathways, there are many questions regarding LBCA's structural morphology and its ability to respond to the environment it inhabited. These include the nature of the cell envelope and whether it possessed all the key components found in modern Bacteria, the degree to which it could sense environmental stimuli, and whether or not it was a motile cell. Modern Bacteria typically have cell membranes comprising a bilayer composed of G3P phospholipids with fatty acids attached via ester bonds, which contrast with the G1P phospholipids with isoprenoids attached via ether bonds found in Archaea (Lombard, López-García and Moreira, 2012b). This "lipid divide" (Koga, 2011) has been touted as a hallmark difference between Archaea and Bacteria, implying that the earliest Bacteria would have had bacterial-type lipids, although recent evidence has challenged this assumption (Sinninghe Damsté *et al.*, 2002; Weijers *et al.*, 2006; Damsté, Rijpstra, *et al.*, 2007; Goldfine, 2010; Villanueva, Schouten and Damsté, 2017; Caforio *et al.*, 2018; Coleman, Pancost and Williams, 2019). In addition, modern bacteria possess peptidoglycan cell walls, and many have an additional outer membrane. This latter trait is thought to be a derived character within Bacteria (Lake, 2009; Gupta, 2011; Tocheva, Ortega and Jensen, 2016), though recent work suggests that the outer membrane arose early in bacterial evolution and may have been present in LBCA (Sutcliffe, 2010; Antunes *et al.*, 2016; Megrian *et al.*, 2020). Furthermore, many scenarios describing the evolution of the earliest prokaryotic communities envisage early cells as non-motile living on substrates, for example within the pores of hydrothermal vents (Martin and Russell, 2007; Lane and Martin, 2012; Sousa *et al.*, 2013; Sousa and Martin, 2014), while other evidence points to the early appearance of the flagellum (Liu and Ochman, 2007a, 2007b), and other machinery for motility and sensory systems (Melville and Craig, 2013).

To answer these questions, we must be able to predict the suite of genes LBCA possessed. The DTL method implemented in ALE, as described in Chapter 2, can count the proportion of sampled reconciliations in which a given gene family is present in a given node, from which a probability of the presence of that gene can be calculated. This allows us to predict the gene content, and therefore reconstruct the metabolic capabilities, of any given node in the tree. The use of this method allows the incorporation of more data about evolutionary history due to use of gene tree topologies, and can distinguish between duplication, transfer or loss with respect to

changing genome size (Szöllősi *et al.*, 2013; Szöllősi, Tannier, *et al.*, 2015). As the method takes HGT into account, it circumnavigates the 'genome of Eden' problem, where ancestral genomes of unrealistic size are predicted due to transfer events (Dagan and Martin, 2007). It must be noted that we can only infer the presence of genes that are in extant genomes, and that the extinction of gene families will leave any reconstructions incomplete. This means that, while we can model the probability of gene family extinction to correct estimated genome size, and infer what modern features ancient life exhibited and the evolutionary timing of appearances of these features, we cannot infer the full potential metabolic capabilities of any ancestral organism. This method has been used in previous research to reconstruct the metabolic capabilities of LACA, inferring it to be an anaerobe which may have used the WLP to fix carbon (Williams *et al.*, 2017). In this study, we use the ALE approach to attempt to reconstruct the gene content and metabolic capabilities of LBCA.

3.2 Methods

Taxon sampling

All analyses and metabolic reconstruction in this chapter were carried out based on the analysis of the GTDB dataset and associated gene family reconciliations detailed in Chapter 2.

Protein and protein family functional annotation

Protein sequences from all genomes in our GTDB taxon sample were annotated using a variety of databases. Functional annotations were obtained using hmmsearch v3.1b2 (settings: -E 1e-5)(Finn, Clements and Eddy, 2011) against KOs from the KEGG Automatic Annotation Server (KAAS) (Aramaki *et al.*, 2020). Additionally, all proteins were scanned for protein domains using InterProScan (v5.31-70.0; settings: --iprlookup --goterms) (Jones *et al.*, 2014). Multiple hits corresponding to the individual domains of a protein are reported using a custom script. For the functional annotation of the 4256 COG families investigated in our ancestral reconstructions, we assigned KOs using a majority rule: i.e. we assigned the KO that was reported in > 50% of the sequences comprising each of the COG families yielding a COG-to-KO mapping file.

Subsequently, we mapped COG descriptions, COG Process/Class, Category description, kegg id, kegg description, and kegg pathway to the COG-to-KO mapping file. COG descriptions were collected from the root annotations downloaded at EggNOG (v5.0.0)(Huerta-Cepas *et al.*, 2019). COG functional category and Process/Class descriptions were derived from EggNOG (v4.0) (Tatusov *et al.*, 2003; Huerta-Cepas *et al.*, 2016). KO pathways were manually curated based on an KO-to-pathway mapping file, and were subsequently mapped to the respective KO.

COG gene families for ancestral gene content reconstruction

We built a set of gene families based on the COG (Tatusov *et al.*, 2003) database for ancestral functional inference. To do so, we annotated each genome in the dataset using eggNOG-mapper v2 (Huerta-Cepas *et al.*, 2017), then clustered proteins into families based on their COG annotations. For proteins annotated with more than one COG category (8% of proteins), we included the protein in both COG families. This resulted in 4256 COG families, of which 3723 had 4 or more sequences. COG families are ideal for ancestral reconstruction because they comprise all of the sequences on extant genomes that can be annotated with a given unambiguous function from the COG ontology. In addition, the hierarchical nature of the COG classification (comprising gene family annotations nested within 23 broader functional categories) enabled us to explicitly model the different evolutionary ages of gene functional classes as part of the analysis, by using category-specific root origination priors (see below).

Root gene mapping approach

To estimate root presence posterior probabilities (PPs) for each gene family for each of the three supported roots, we first estimated the root origination prior (O_R) by maximum likelihood, finding the O_R value that maximises the total reconciliation likelihood summed over all gene families. We then used the global ML O_R value to calculate the root presence posterior probabilities for each family; that is, the probability that one or more copies of a given gene family were present at the root, given the ML O_R value. These indicated that families with different functions varied widely in terms of root presence probability, in agreement with established theory (Jain, Rivera and Lake, 1999); for example, proteins involved in translation (J) had the highest root presence probabilities among the functional classes investigated (Table

3.1). We therefore estimated root origination rates independently for each of the 23 COG functional categories, and used these rates to estimate the posterior probability of presence at the root node for each gene family. Initial gene content and metabolic inferences at particular nodes were based on gene families with a PP of >0.95 at that node. This approach is conservative and can result in a range of PP values for different proteins within a metabolic pathway. Therefore, we manually investigated the PPs of key pathways identified from the initial PP cut-off and inferred the presence of specific pathways or functional modules if most its components were found with PP >0.5.

COG category	No. at root (flat prior)	% at root (flat prior)	No. at root (ML prior, PP >= 0.95)	% at root (ML prior, PP >= 0.95)	No. at root (ML prior, PP >= 0.8)	% at root (ML prior, PP >= 0.8)	No. of families in category	ML O_R	COG category description
J	3	1.69	125	70.62	157	88.7	177	6739.434	Translation ribosomal structure and biogenesis
F	1	0.98	57	55.88	89	87.25	102	5294.183	Nucleotide transport and metabolism
L	4	2.23	42	23.46	100	55.87	179	1468.033	Replication recombination and repair
H	3	1.82	26	15.76	85	51.52	165	1393.54	Coenzyme transport and metabolism
-	0	0	0	0	1	50	2	1290.933	
N	1	1.43	18	25.71	38	54.29	70	1270.758	Cell motility
E	2	0.88	45	19.91	116	51.33	226	1146.195	Amino acid transport and metabolism
D	2	4.44	7	15.56	16	35.56	45	850.0477	Cell cycle
P	4	2.01	16	8.04	64	32.16	199	650.1995	Inorganic ion transport and metabolism
M	1	0.71	15	10.64	49	34.75	141	625.6504	Cell wall/membrane/envelope biogenesis
C	4	1.67	22	9.21	57	23.85	239	581.1914	Energy production and conversion
G	3	1.61	11	5.91	43	23.12	186	513.1411	Carbohydrate transport and metabolism
I	0	0	1	1.3	12	15.58	77	457.9165	Lipid transport and metabolism

U	1	1.2	3	3.61	12	14.46	83	364.7495	Intracellular trafficking
K	1	0.76	3	2.29	10	7.63	131	260.9843	Transcription
O	0	0	1	0.81	9	7.32	123	256.3592	Post-translational modification
S	6	0.44	9	0.66	69	5.02	1374	163.0099	Function unknown
B	0	0	0	0	1	20	5	126.1288	Chromatin structure and dynamics
T	0	0	1	1.14	3	3.41	88	125.0658	Signal transduction mechanisms
V	0	0	0	0	0	0	32	76.72498	Defense mechanisms
Q	0	0	0	0	0	0	73	1.196214	Secondary metabolites
A	0	0	0	0	0	0	4	1	RNA processing and modification
Z	0	0	0	0	0	0	2	1	Cytoskeleton

Table 3.1 *Estimated root origination rates and root presences by COG functional category, including COGs recoveredas a percentage of the number of gene families in the average extant bacterial genome (indicated by "% at root").*

Impact of root branch on LBCA gene content

The credible set of root branches from the ALE analysis comprised three adjacent branches at the centre of the tree (Chapter 2, Figure 2.7(a)). The difference between these three root positions relates to the placement of Fusobacteriota, either as the root branch or as the most basal split on either the Gracilicutes or Terrabacteria+DST "sides" of the rooted tree. We therefore estimated root PPs for COG families on all three branches; Supplementary Table 1 provides root PPs under all three roots and indicates when genes were present in 1, 2, or all 3 candidate root positions.

Metabolic comparisons

Results from the PP analysis were used as the framework for metabolic comparisons and reconstruction of the proteome of LBCA and to explore occurrence of gene families across the tree (subsequent nodes are dealt with in Chapter 4). First, the occurrence of an individual COG family across each taxon was counted in R (v3.6.3) (Supplementary Table 1). This binary presence/absence matrix was combined with

the PP values for Nodes corresponding to the CPR, Chloroflexota+CPR, Chloroflexota, Terrabacteria, DST+Terrabacteria, Gracilicutes-Spirochaetota, Gracilicutes+Spirochaetota, Root 1, Root 2, and Root 3, filtered with a cutoff of PP>0.50. The combined count table was summarised using the ddply function of the plyr package (v1.8.4), which was used to summarise the counts across each phylogenetic cluster, node, and root. Data is visualised in a heatmap generated using the ggplot function with geom_tile and facet_grid of the ggplot2 package (v3.2.0). Heatmap categories for pathways were scaled based on the number of COG families, results were plotted using the grid.draw function of the grid package (v3.6.3). Heatmaps were manually merged with a representative tree in Adobe Illustrator (v22.0.1).

3.3 Results and Discussion

Estimating the size of LBCA's genome

For the following discussion regarding metabolic reconstructions, we refer to the three branches in the root region as Root 1, Root 2, and Root 3 respectively, as shown in Figure 2.7(a) (Chapter 2). Based on the root placement and estimated rates of gene family extinction (Williams *et al.*, 2017), we predict that LBCA encoded 1292.6-2142.9 COG family members, the majority of which (median estimates 65-69.5%; 95% CI 57-82%) survived to be sampled in at least one present day genome. Based on the relationship between COG family members and genome size for extant Bacteria (Pearson's r = 0.96, $P = 8 \times 10^{-153}$), we estimate the genome size of LBCA to be 2.69Mb +/- 0.4Mb (standard error) for Root 1 (Fusobacteriota with Terrabacteria; 2.59Mb +/- 0.41Mb for Root 2 (Fusobacteriota with Gracilicutes), and 1.6 +/- 0.5Mb for Root 3 (Fusobacteriota basal).

Information processing, cell division and signaling

In what follows, we provide probability ranges across the three branches of the root region for the presence of key genes. One caveat of our analyses is that the method has limited power to distinguish between ancestral presence in LBCA as opposed to origin on an early descendant branch followed by gene transfer; this may contribute to

the mapping of some combinations of gene families to LBCA that are likely to be ancient but, on physiological grounds, are unlikely to have coexisted in a single cell (see below).

A large number of genes that can be mapped back to LBCA in all three roots with PPs>0.5 are involved in informational processing and storage machineries such as translation, transcription and replication, with many being highly supported (PP>0.95). This includes the majority of ribosomal proteins, tRNA synthetases and genes involved in their biosynthesis. We also recovered DNA polymerase I, III and IV, DNA topoisomerase type I, and DNA ligase. Additionally many exonucleases, endonucleases and ribonucleases are recovered, as well genes for base excision repair, nucleotide excision repair, mismatch repair and homologous recombination (Supplementary Table 1).

Furthermore, we detected many genes for cellular processes and signalling, including cell division, signal transduction, membrane transport, intracellular trafficking, chemotaxis and cellular mobility. For example, the PP for the presence of the cell division proteins FtsZ, FtsQ and FtsA (K03531, K03589, and K03590) were >0.9 in each root. A small number of proteins involved in signal transduction, i.e. two-component regulatory systems, had a PP>0.95, with many more genes recovered with a PP>0.5. Additionally, we recover genes for components of 16 ABC transporters across all three roots, with 23 recovered in both Roots 1 and 2 (PP >0.5). We also recover evidence for the bacterial secretion system, with four proteins of Sec system having a PP>0.8 across all three roots. Two of these, SecY (K03076) and SecD/F (K12257) had a PP>0.95 in roots 1 and 2. Additionally, the GspD (K02453) and GspJ (K02459) subunits of secretion system II were recovered with a PP>0.8 in all three roots, or with a PP >0.95 in Root 1 and 2, respectively (Supplementary Table 1).

Cell envelope and motility

One interesting aspect of prokaryotic evolution is the so-called ‘lipid divide’ (Koga, 2011) Typically, Archaea have G1P phospholipids with ether bonds and isoprenoid chains, which are often membrane spanning (Lombard, López-García and Moreira, 2012b). Bacteria, on the other hand, possess G3P phospholipids, typically with ester bonds and fatty-acid chains, that form bilayers (Lombard, López-García and Moreira,

2012b). While several recent studies indicate the existence of lipids with mixed characteristics, their provenance and the mechanisms by which they are synthesised are currently unclear (Sinninghe Damsté *et al.*, 2002; Weijers *et al.*, 2006; Damsté, Rijpstra, *et al.*, 2007; Goldfine, 2010). Nonetheless, such discoveries have led to the questioning of the lipid divide. Genes encoding components of archaeal lipids have been found to be widespread in Bacteria (Villanueva, Schouten and Damsté, 2017; Coleman, Pancost and Williams, 2019) and genes encoding components for bacterial lipids are found in Archaea, although to a lesser extent (Coleman, Pancost and Williams, 2019). Furthermore, a recent study demonstrated that *E. coli* engineered to have a stable hybrid heterochiral lipid membrane do not experience any change in growth rate (Caforio *et al.*, 2018). While, phylogenetic analyses suggest that the archaeal pathway may predate the bacterial one (Coleman, Pancost and Williams, 2019), there was no significant support for the presence of archaeal lipid biosynthesis genes in LBCA. In fact, genes coding for the enzymes that determine the phospholipid stereochemistry did not have PPs above a threshold of 0.5 (Table 3.2), though a gene coding for glycerol-3-phosphate (*glpA*, K00111) might have been present in Root 1 (PP=0.49), in agreement with some previous research (Yokobori *et al.*, 2016). We do, however, recover glycerol kinase (GlpK), which can synthesise G3P form glycerol (K00864, PP=0.9/0.87/0.73). Furthermore, our analyses suggest the presence of PlsY (K08591, PP=0.94/0.92/0.82) and PlsX (K03621, PP=0.71/0.68/0.38), which attach the first fatty acid chain to G3P in many Bacteria (some species alternatively use PlsB, which we do not recover PP>0.5 in any root). We also recover a putative PlsC (K15781, PP=0.86/0.83/0.64), which attaches the second fatty-acid side-chain. Our inferences therefore suggest that LBCA had bacterial phospholipid membranes, while being unable to synthesise archaeal lipids.

FAM	kegg_id	kegg_description	Root 1	Root 2	Root 3
COG0344	**K08591**	**acyl_phosphate:glycerol-3-phosphate_acyltransferase_[EC:2.3.1.275]**	**0.94**	**0.92**	**0.82**
COG0554	**K00864**	**glycerol_kinase_[EC:2.7.1.30]**	**0.9**	**0.87**	**0.73**
COG2376	K05878	phosphoenolpyruvate---glycerone_phosphotransferase_subunit_DhaK_[EC:2.7.1.121]	0.89	0.86	0.7
COG0560	**K15781**	**putative_phosphoserine_phosphatase_/_1-acylglycerol-3-phosphate_O-acyltransferase_[EC:3.1.3.3_2.3.1.51]**	**0.86**	**0.83**	**0.64**

COG1502	K06131	cardiolipin_synthase_A/B_[EC:2.7.8.-]	0.79	0.75	0.51
COG4589	K00981	phosphatidate_cytidylyltransferase_[EC:2.7.7.41]	0.78	0.75	0.5
COG3412	K05881	phosphoenolpyruvate---glycerone_phosphotransferase_subunit_DhaM_[EC:2.7.1.121]	0.77	0.74	0.48
COG1597	K07029	diacylglycerol_kinase_(ATP)_[EC:2.7.1.107]	0.73	0.7	0.4
COG0416	**K03621**	**phosphate_acyltransferase_[EC:2.3.1.274]**	**0.71**	**0.68**	**0.38**
COG0575	K00981	phosphatidate_cytidylyltransferase_[EC:2.7.7.41]	0.69	0.66	0.34
COG0584	K01126	glycerophosphoryl_diester_phosphodiesterase_[EC:3.1.4.46]	0.66	0.63	0.31
COG1267	K01095	phosphatidylglycerophosphatase_A_[EC:3.1.3.27]	0.65	0.61	0.29
COG4302	K03736	ethanolamine_ammonia-lyase_small_subunit_[EC:4.3.1.7]	0.57	0.53	0.22
COG0578	**K00111**	**glycerol-3-phosphate_dehydrogenase_[EC:1.1.5.3]**	**0.49**	**0.45**	**0.15**
COG5379	K13622	S-adenosylmethionine-diacylglycerol_3-amino-3-carboxypropyl_transferase	0.48	0.45	0.15
COG3675	K16818	phospholipase_A1_[EC:3.1.1.32]	0.4	0.37	0.09
COG4303	K03735	ethanolamine_ammonia-lyase_large_subunit_[EC:4.3.1.7]	0.39	0.37	0.08
COG0371	**K00096**	**glycerol-1-phosphate_dehydrogenase_[NAD(P)+]_[EC:1.1.1.261]**	**0.38**	**0.36**	**0.08**
COG0240	**K00057**	**glycerol-3-phosphate_dehydrogenase_(NAD(P)+)_[EC:1.1.1.94]**	**0.32**	**0.31**	**0.05**
COG3075	**K00112**	**glycerol-3-phosphate_dehydrogenase_subunit_B_[EC:1.1.5.3]**	**0.3**	**0.3**	**0.04**
COG1368	K19005	lipoteichoic_acid_synthase_[EC:2.7.8.20]	0.29	0.29	0.04
COG4819	K04019	ethanolamine_utilization_protein_EutA	0.28	0.27	0.03
COG0644	K17830	digeranylgeranylglycerophospholipid_reductase_[EC:1.3.1.101_1.3.7.11]	0.27	0.27	0.03
COG1887	K09809	CDP-glycerol_glycerophosphotransferase_[EC:2.7.8.12]	0.27	0.26	0.03
COG0688	K01613	phosphatidylserine_decarboxylase_[EC:4.1.1.65]	0.24	0.24	0.02
COG0558	K08744	cardiolipin_synthase_(CMP-forming)_[EC:2.7.8.41]	0.19	0.19	0.01
COG2829	K01058	phospholipase_A1/A2_[EC:3.1.1.32_3.1.1.4]	0.16	0.15	0
COG1075	K01046	triacylglycerol_lipase_[EC:3.1.1.3]	0.09	0.09	0
COG2937	**K00631**	**glycerol-3-phosphate_O-acyltransferase_[EC:2.3.1.15]**	**0.08**	**0.08**	**0**

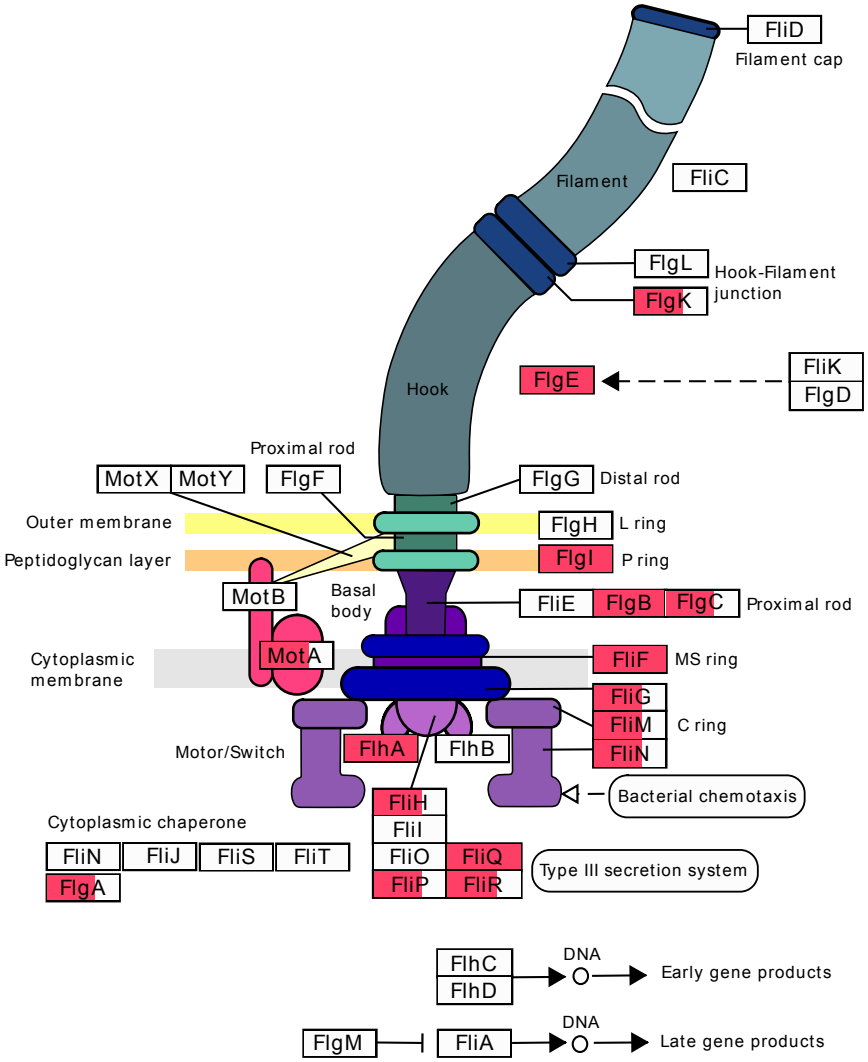
COG1646	K17104	phosphoglycerol_geranylgeranyltransferase_[EC:2.5.1.41]	0.04	0.04	0
COG4909	K06120	glycerol_dehydratase_large_subunit_[EC:4.2.1.30]	0	0	0
COG5153	K17900	lipase_ATG15_[EC:3.1.1.3]	0	0	0
COG5153	K17900	lipase_ATG15_[EC:3.1.1.3]	0	0	0

Table 3.2 *Posterior probabilities for the presence of glycerolipids in the last bacterial common ancestor (LBCA). Key genes whose Kegg IDs are given in the text are highlighted in bold. Annotations and PP values for KOs can be found in Supplementary Table 2 and all KOs in Supplementary Table 1.*

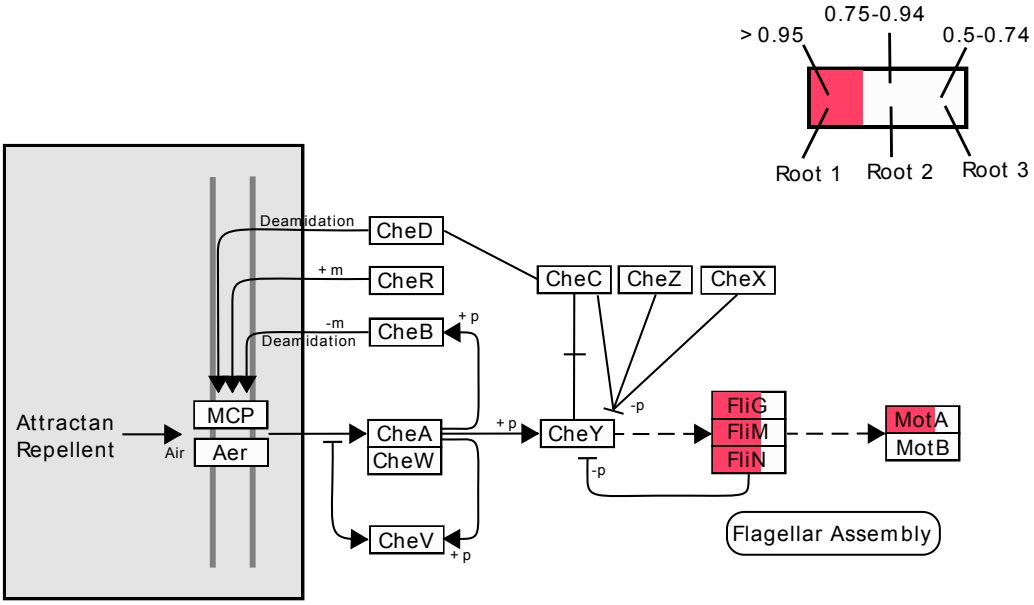
Our analyses also suggest that LBCA possessed a fully functioning flagellum (Fig. 3.1), in agreement with the previous research (Liu and Ochman, 2007b). More than half of the genes involved in flagellar construction are found with a PP>0.5 in all three roots, with 35/46 present in Roots 1 and 2 (Supplementary Tables 1 and 2), including components of the C, Ms and P rings, the hook and hook-filament junction, as well as the Type III secretion system. We also recover three flagellar genes unique to diderm bacteria, flgH (PP=0.92/0.9/0.77), flgI (PP=1/0.99/0.99), which code for the L and P ring respectively, as we as flgA (PP=0.96/0.95/0.88). The corresponding proteins anchor flagella in diderm membranes, indicating that LBCA had a typical gram-negative flagellum and a double-membrane. In addition to a flagellum, we find 15 proteins for the construction of pili (PP>0.5), including 5 that seem implicated in the synthesis of a Type IV pilus. The key protein PliQ (K02666), which anchors the pilus in the outer membrane, is recovered with PP=0.86/0.82/0.62 across the different roots.

Fig. 3.1 (below): *Components of the flagellum (a) and chemotaxis (b) inferred in the last bacterial common ancestor (LBCA). The reconstruction is based on genes that could be mapped to a given node with PP>0.5. White indicates PP<0.5. The presence of a gene within a pathway is indicated as shown in the key. Annotations and PP values for KOs can be found in Supplementary Table 2 and all KOs in Supplementary Table 1.*

a


FliD
Filament cap
Filament
FliC
FlgL
Hook-Filament junction
FlgK
Hook
FlgE
FliK
FlgD
Proximal rod
MotX
MotY
FlgF
FlgG
Distal rod
Outer membrane
FlgH
L ring
Peptidoglycan layer
FlgI
P ring
MotB
Basal body
FliE
FlgB
FlgC
Proximal rod
Cytoplasmic membrane
MotA
FliF
MS ring
FliG
FliM
FliN
C ring
Motor/Switch
FlhA
FlhB
Bacterial chemotaxis
FliH
FliI
FliO
FliQ
FliP
FliR
Type III secretion system
Cytoplasmic chaperone
FliN
FliJ
FliS
FliT
FlgA
FlhC
FlhD
DNA
Early gene products
FlgM
FliA
DNA
Late gene products

b


0.75-0.94
> 0.95
0.5-0.74
Root 1
Root 2
Root 3
Deamidation
CheD
+ m
CheR
CheC
CheZ
CheX
-m
Deamidation
CheB
+ p
Attractan
Repellent
Air
MCP
Aer
CheA
CheW
+ p
CheY
-p
FliG
FliM
FliN
MotA
MotB
-p
CheV
+ p
Flagellar Assembly

Bacteria have classically been divided into Gram-positive monoderms, with a single cell membrane and a thick peptidoglycan wall, and Gram-negative diderms with a thin peptidoglycan wall between two cell membranes (Megrian *et al.*, 2020), although many Bacteria have been shown to exhibit various atypical cell envelopes with a mixture of different characteristics (Sutcliffe, 2010). It has often been suggested that having a double-membrane is a derived state (Lake, 2009; Gupta, 2011; Tocheva *et al.*, 2011; Megrian *et al.*, 2020), though more recent work has raised the possibility for a diderm ancestor (Antunes *et al.*, 2016; Megrian *et al.*, 2020). The latter would be consistent with our phylogenetic analyses that resolve diderm Bacteria on either side of the possible roots (Fig. 3.2). In agreement with this, we found many genes for lipopolysaccharide biosynthesis proteins, 24 of which had a PP>0.5 (8 with PP>0.9, see Table 3.3), including the key proteins LpxC (K02535, PP=0.99/0.99/0.98) and KdsA (K01627, PP=0.92/0.9/0.78). In addition to lipopolysaccharides, two proteins used in transport across the outer membrane are recovered with PP>0.5, BamA (K07277, PP=0.65/0.61/0.29) and OmpH (K06142, PP=0.94/0.93/0.82). The inferred presence of these proteins in LBCA lends further support to the hypothesis that LBCA was a diderm, featuring an outer membrane with a full complement of lipopolysaccharides. We additionally recovered various genes encoding proteins for the construction of the cell wall and cell envelope, including 14 proteins predicted to be involved in peptidoglycan biosynthesis that had a PP>0.5 in roots 1 and 2. Moderate support for the presence of *mreB* (0.9/0.88/0.73, root branches 1-3 as depicted in Fig. 1(b)), *mreC* (0.82/0.79/0.57) and *mreD* (0.86/0.83/0.63) at the root suggests that LBCA possessed rod-shaped cells.

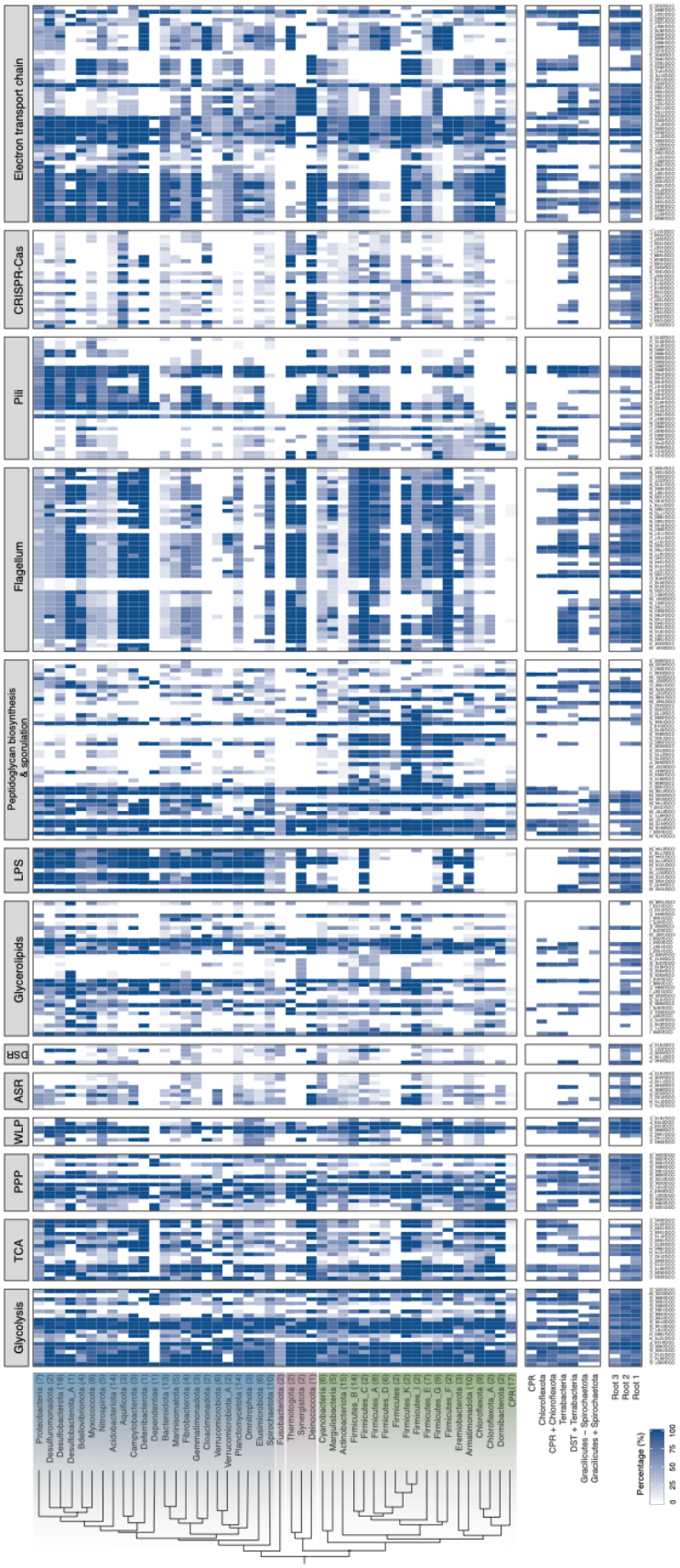
FAM	kegg_id	kegg_description	Root 1 PPs	Root 2 PPs	Root 3 PPs
COG0774	K02535	UDP-3-O-[3-hydroxymyristoyl]_N-acetylglucosamine_deacetylase_[EC:3.5.1.108]	0.99	0.99	0.98
COG0279	K03271	D-sedoheptulose_7-phosphate_isomerase_[EC:5.3.1.28]	0.99	0.99	0.98
COG0241	K03273	D-glycero-D-manno-heptose_1,7-bisphosphate_phosphatase_[EC:3.1.3.82_3.1.3.83]	0.99	0.98	0.96
COG4370	K00748	lipid-A-disaccharide_synthase_[EC:2.4.1.182]	0.96	0.95	0.88
COG1560	K02517	Kdo2-lipid_IVA_lauroyltransferase/acyltransferase_[EC:2.3.1.241_2.3.1.-]	0.94	0.93	0.83

COG4591	K09808	lipoprotein-releasing_system_permease_protein	0.94	0.92	0.82
COG1137	K06861	lipopolysaccharide_export_system_ATP-binding_protein_[EC:3.6.3.-]	0.92	0.9	0.79
COG2877	K01627	2-dehydro-3-deoxyphosphooctonate_aldolase_(KDO_8-P_synthase)_[EC:2.5.1.55]	0.92	0.9	0.78
COG1043	K00677	UDP-N-acetylglucosamine_acyltransferase_[EC:2.3.1.129]	0.9	0.87	0.73
COG1682	K09690	lipopolysaccharide_transport_system_permease_protein	0.89	0.86	0.7
COG1044	K02536	UDP-3-O-[3-hydroxymyristoyl]_glucosamine_N-acyltransferase_[EC:2.3.1.191]	0.88	0.85	0.68
COG1663	K00912	tetraacyldisaccharide_4'-kinase_[EC:2.7.1.130]	0.87	0.84	0.66
COG1682	K09690	lipopolysaccharide_transport_system_permease_protein	0.86	0.83	0.65
COG0763	K00748	lipid-A-disaccharide_synthase_[EC:2.4.1.182]	0.84	0.81	0.6
COG2605	K07031	D-glycero-alpha-D-manno-heptose-7-phosphate_kinase_[EC:2.7.1.168]	0.83	0.8	0.59
COG2870	K03272	D-beta-D-heptose_7-phosphate_kinase_/_D-beta-D-heptose_1-phosphate_adenosyltransferase_[EC:2.7.1.167_2.7.7.70]	0.82	0.79	0.57
COG2121	K09778	Kdo2-lipid_IVA_3'_secondary_acyltransferase_[EC:2.3.1.-]	0.82	0.79	0.57
COG0794	K06041	arabinose-5-phosphate_isomerase_[EC:5.3.1.13]	0.81	0.77	0.54
COG1134	K09691	lipopolysaccharide_transport_system_ATP-binding_protein	0.7	0.67	0.36
COG1212	K00979	3-deoxy-manno-octulosonate_cytidylyltransferase_(CMP-KDO_synthetase)_[EC:2.7.7.38]	0.68	0.65	0.33
COG1134	K09691	lipopolysaccharide_transport_system_ATP-binding_protein	0.66	0.63	0.31
COG0615	K03272	D-beta-D-heptose_7-phosphate_kinase_/_D-beta-D-heptose_1-phosphate_adenosyltransferase_[EC:2.7.1.167_2.7.7.70]	0.61	0.57	0.25
COG1519	K02527	3-deoxy-D-manno-octulosonic-acid_transferase_[EC:2.4.99.12_2.4.99.13_2.4.99.14_2.4.99.15]	0.51	0.47	0.17
COG2956	K19804	lipopolysaccharide_assembly_protein_B	0.5	0.46	0.16
COG0795	K11720	lipopolysaccharide_export_system_permease_protein	0.45	0.42	0.13
COG4785	K05803	lipoprotein_NlpI	0.42	0.4	0.11
COG2980	K03643	LPS-assembly_lipoprotein	0.29	0.29	0.04
COG1368	K19005	lipoteichoic_acid_synthase_[EC:2.7.8.20]	0.29	0.29	0.04
COG4261	K02517	Kdo2-lipid_IVA_lauroyltransferase/acyltransferase_[EC:2.3.1.241_2.3.1.-]	0.27	0.27	0.03
COG5375	K11719	lipopolysaccharide_export_system_protein_LptC	0.18	0.17	0

COG3642	K11211	3-deoxy-D-manno-octulosonic_acid_kinase_[EC:2.7.1.166]	0.17	0.16	0
COG1778	K03270	3-deoxy-D-manno-octulosonate_8-phosphate_phosphatase_(KDO_8-P_phosphatase)_[EC:3.1.3.45]	0.1	0.1	0
COG2908	K03269	UDP-2,3-diacylglucosamine_hydrolase_[EC:3.6.1.54]	0.09	0.09	0
COG5416	K08992	lipopolysaccharide_assembly_protein_A	0.06	0.05	0
COG3528	K09953	lipid_A_3-O-deacylase_[EC:3.1.1.-]	0.05	0.05	0
COG1934	K09774	lipopolysaccharide_export_system_protein_LptA	0.04	0.04	0
COG3475	K07271	lipopolysaccharide_cholinephosphotransferase_[EC:2.7.8.-]	0.03	0.03	0
COG3117	K11719	lipopolysaccharide_export_system_protein_LptC	0.03	0.03	0
COG3127	K09808	lipoprotein-releasing_system_permease_protein	0	0	0

Table 3.3 *Posterior probabilities for the presence of genes involved in lipopolysaccharide biosynthesis in LBCA. Annotations and PP values for KOs can be found in Supplementary Table 2 and all KOs in Supplementary Table 1.*

Figure 3.2 (below): *Distribution of COG families from key metabolic pathways inferred to the last bacterial common ancestor (LBCA). The occurrence of COG families in the taxa sampled in this study are represented as percentage presence across phylogenetic clusters (phylum) based on a presence/absence table. COG families inferred to the given nodes and the tree possible root positions (see Methods) are represented by corresponding PP values (PP>0.5). TCA=Tricarboxylic Acid Cycle, PPP=Pentose phosphate pathway, ASR=Assimilatory sulphate reduction, DSR=Dissimilatory suphfate reduction, LPS=Lipopolysaccharide. Annotations and PP values for all KOs can be found in Supplementary Table 2. Full heatmap can be found in Appendix B, Fig. 1.*


Glycolysis
TCA
PPP
WLP
ASR
DSR
Glycerolipids
LPS
Peptidoglycan biosynthesis & sporulation
Flagellum
Pili
CRISPR-Cas
Electron transport chain
Proteobacteria (7)
Desulfuromonadota (2)
Desulfobacterota (18)
Desulfobacterota_A (1)
Bdellovibrionota (4)
Myxococcota (8)
Nitrospirota (5)
Acidobacteriota (14)
Aquificota (3)
Campylobacterota (3)
Deferribacterota (1)
Dependentiae (1)
Bacteroidota (13)
Marinisomatota (5)
Fibrobacterota (4)
Gemmatimonadota (3)
Cloacimonadota (2)
Verrucomicrobiota (6)
Verrucomicrobiota_A (2)
Planctomycetota (14)
Omnitrophota (3)
Elusimicrobiota (6)
Spirochaetota (10)
Fusobacteriota (2)
Thermotogota (2)
Synergistota (2)
Deinococcota (1)
Cyanobacteria (6)
Margulisbacteria (5)
Actinobacteriota (15)
Firmicutes_B (14)
Firmicutes_C (2)
Firmicutes_A (8)
Firmicutes_D (6)
Firmicutes (2)
Firmicutes_K (2)
Firmicutes_I (2)
Firmicutes_E (7)
Firmicutes_G (9)
Firmicutes_F (2)
Eremiobacterota (3)
Armatimonadota (10)
Chloroflexota (9)
Chloroflexota_A (2)
Dormibacterota (2)
CPR (17)
CPR
Chloroflexota
CPR + Chloroflexota
Terrabacteria
DST + Terrabacteria
Gracilicutes – Spirochaetota
Gracilicutes + Spirochaetota
Root 3
Root 2
Root 1
Percentage (%)
0 25 50 75 100

Carbon metabolism, autotrophy and respiratory complexes

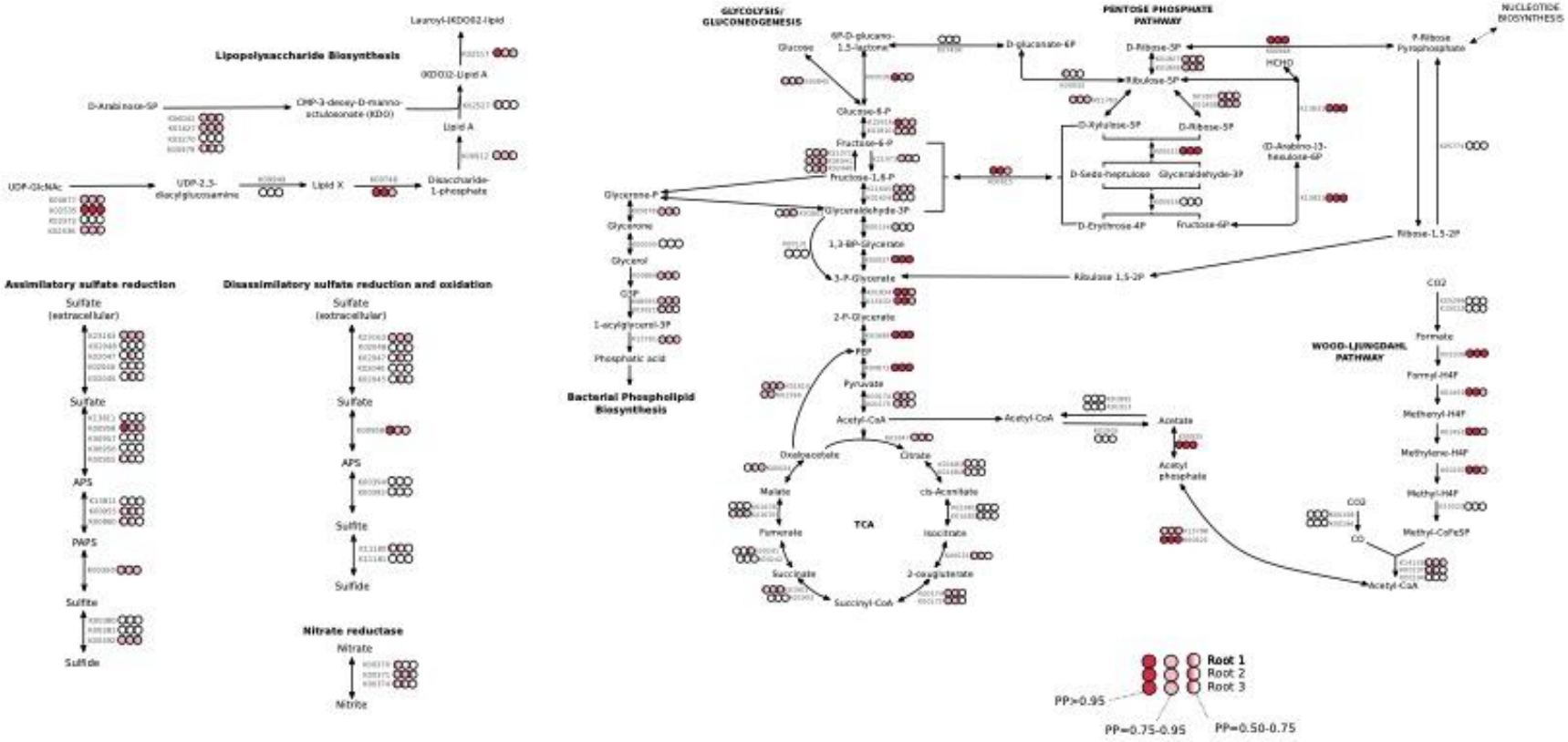
The largest category of genes mapped to LBCA encode proteins involved in metabolism and transport of amino acids, coenzymes, nucleotides, inorganic ions and carbohydrates including pathways involved in central carbohydrate metabolism. We recovered components of several core pathways for carbohydrate metabolism with high posterior support, namely glycolysis, the tricarboxylic acid (TCA) cycle, and the pentose phosphate pathway (Fig. 3.3, Supplementary Tables 1 and 2).

There are several carbon fixation pathways in extant autotrophic Bacteria including the Wood-Ljungdahl Pathway (WLP), the reverse TCA cycle, the Calvin cycle, and the 3-hydroxypropionate bicycle as well as distinct variants for the different pathways (Fuchs, 2011). While the large subunit of the key enzyme of the Calvin cycle, i.e. ribulose-bisphosphate carboxylase (RubisCO) and RubisCO-like proteins (RLP) are widespread in microbes and represent one of the most abundant protein families in the biosphere, we found only low to moderate support for the presence of a large subunit RubisCO-encoding gene in in LBCA (PP-range in the three roots: 0.24-0.59), in agreement with hypotheses in which the Calvin cycle and perhaps the carboxylation function of RubisCO/RLP evolved late (Erb and Zarzycki, 2018).

In contrast, the reverse TCA cycle with the hallmark enzyme ATP citrate lyase, has been suggested as a possible ancient carbon fixation pathway (Wächtershäuser, 1990; Cody *et al.*, 2001; Smith and Morowitz, 2004; Nunoura *et al.*, 2018). While our analyses do not support the presence of ATP citrate lyase in LBCA, we do identify other enzymes of the TCA, including a citrate synthase as well as subunits of an oxoacid:ferredoxin oxidoreductase, which may function in the TCA in both the oxidative and reductive direction (Fig. 3.3, Supplementary Tables 1 and 2). For example, it has recently been shown that the citrate synthase, originally thought to operate only in the oxidative direction, can in fact catalyse the reverse reaction and allows to fix carbon in the facultatively chemolithoautotrophic thermophile *Thermosulfidibacter takaii* ABI70S6 (Nunoura *et al.*, 2018). It has also been suggested that ATP citrate lyase may have emerged at a later stage from the domains of citrate synthase and succinyl-CoA synthase (Kanao *et al.*, 2001; Nunoura *et al.*, 2018), which may further suggest that the TCA could operate in the reductive direction without ATP citrate lyase. Therefore, our analyses do not exclude the possibility that LBCA was

capable of using the reverse TCA cycle to fix carbon. Similarly, we find support for components of the reductive glycine pathway, which shares the methyl-branch of the WLP (Sanchez-Andrea *et al.* 2020), although as with the reverse TCA, we do not find the key enzymes which specifies the direction of the pathway. We therefore cannot exclude the possibility that the reductive glycine pathway was being used to fix carbon.

Fig. 3.3 (below): *Metabolic map of the central metabolic pathways inferred in the last bacterial common ancestor (LBCA). The reconstruction is based on genes that could be mapped to a given node with PP >0.5. The presence of a gene within a pathway is indicated as shown in the key. Annotations and PP values for KOs can be found in Supplementary Table 2 and all KOs in Supplementary Table 1.*


Lipopolysaccharide Biosynthesis
Lauroyl-(KDO)2-Lipid
(KDO)2-Lipid A
Lipid A
D-Arabinose-5P
CMP-3-deoxy-D-manno-octulosonate (KDO)
UDP-GlcNAc
UDP-2,3-diacylglucosamine
Lipid X
Disaccharide-1-phosphate
Assimilatory sulfate reduction
Sulfate (extracellular)
Sulfate
APS
PAPS
Sulfite
Sulfide
Dissimilatory sulfate reduction and oxidation
Sulfate (extracellular)
Sulfate
APS
Sulfite
Sulfide
Nitrate reductase
Nitrate
Nitrite
GLYCOLYSIS/ GLUCONEOGENESIS
PENTOSE PHOSPHATE PATHWAY
NUCLEOTIDE BIOSYNTHESIS
Glucose
6P-D-glucono-1,5-lactone
D-glucanate-6P
D-Ribose-5P
HCHO
P-Ribose Pyrophosphate
Ribulose-5P
Glucose-6-P
D-Xylulose-5P
D-Ribose-5P
Fructose-6-P
(D-Arabino-3)-hexulose-6P
Fructose-1,6-P
D-Sedo-heptulose
Glyceraldehyde-3P
Glycerone-P
Glyceraldehyde-3P
D-Erythrose-4P
Fructose-6P
Ribose-1,5-2P
Glycerone
1,3-BP-Glycerate
Ribulose 1,5-2P
Glycerol
3-P-Glycerate
G3P
2-P-Glycerate
1-acylglycerol-3P
PEP
Phosphatic acid
Pyruvate
Bacterial Phospholipid Biosynthesis
Acetyl-CoA
Acetyl-CoA
Acetate
Acetyl phosphate
Oxaloacetate
Citrate
Malate
cis-Aconitate
TCA
Fumarate
Isocitrate
Succinate
2-oxoglutarate
Succinyl-CoA
WOOD-LJUNGDAHL PATHWAY
CO2
Formate
Formyl-H4F
Methenyl-H4F
Methylene-H4F
Methyl-H4F
CO2
CO
Methyl-CoFeSP
Acetyl-CoA
Root 1
Root 2
Root 3
PP>0.95
PP=0.75-0.95
PP=0.50-0.75

Additionally, the WLP is generally thought to represent an ancient carbon fixation pathway on the basis of both biogeochemical and phylogenetic arguments (Fuchs, 2011; Sousa and Martin, 2014; Williams *et al.*, 2017; Adam, Borrel and Gribaldo, 2018) and previous phylogenetic work has suggested its presence in both the archaeal (Williams *et al.*, 2017; Adam, Borrel and Gribaldo, 2018) and bacterial (Adam, Borrel and Gribaldo, 2018) common ancestors. While components of the methyl branch of the pathway were mapped to the root with PP support >0.95 for all roots, PPs for the subunits of the hallmark enzyme of the WLP, the carbon monoxide dehydrogenase/acetyl-CoA synthase (CODH/ACS) were had only moderate root support (PP=0.5-0.75 for two subunits) and low support (PP <0.5) for other subunits. Considering the methyl-branch of the WLP is also involved in alternative metabolisms including formate and folate transformations (Stover, 2009; Brosnan and Brosnan, 2016), it remains unclear whether the lack of strong support for a CODH/ACS in LBCA indicates that the WLP was absent in the bacterial ancestor or simply reflects the difficulty of mapping genes to the root with high statistical support. CODH/ACS subunits are not as widely distributed in extant Bacteria (Fig. 3.3, Supplementary Tables 1 and 2) as in Archaea, so that their presence in LBCA would require extensive subsequent loss or HGT throughout the diversification of Bacteria or suggests the later acquisition of this enzyme complex mediating the carbonyl-branch of the WLP.

On the other hand, our analyses provided strong support for the presence of phosphate acetyltransferase and acetate kinase, enzymes synthesizing acetate from acetyl-CoA (K13788, PP=0.86/0.9/0.74; K00925, PP=0.997/0.997/0.98)(Schuchmann and Müller, 2014). Furthermore, we find all six subunits of an Na^+-translocating ferredoxin:NAD^+ oxidoreductase (Rnf) complex, comprised of key genes *rnfA* (K03617, PP=0.99/0.99/0.95), *rnfB* (K03616, PP=0.84/0.95/0.70), *rnfC* (K03615, PP=0.77/0.89/0.59), *rnfD* (K03614, PP=0.89/0.94/0.78), *rnfE* (K03613, PP=0.89/0.96/0.77), and *rnfG* (K03612, PP=0.56/0.74/0.34) in LBCA. The Rnf complex is a membrane-bound respiratory enzyme that couples the oxidation of reduced ferredoxin (Fd_{red}) to the reduction of NAD^+ through a flavin-based electron transport chain, which is concomitantly coupled with the translocation of Na^+ ions, generating a transmembrane Na^+ motive force (Biegel and Muller, 2010). In the anaerobic acetogen, *Acetobacterium woodii*, this chemiosmotic gradient is used to drive ATP synthesis via a Na^+-dependent F_1F_0 ATP synthase (Biegel *et al.*, 2011).

Evidence has shown that the Rnf complex can function reversibly, where it catalyzes the reduction of oxidised ferredoxin with NADH using a chemiosmotic gradient (H^+/Na^+) generated via ATP hydrolysis (Hess, Schuchmann and Müller, 2013; Westphal *et al.*, 2018). Electron bifurcation through the redox coupling of NADH and ferredoxin allows for the production of high-energy intermediates from low-potential electron donors, which can be used to reduce CO_2 in the WLP (Buckel and Thauer, 2013; Müller, Chowdhury and Basen, 2018). Taken together and in agreement with previous work suggesting the antiquity of the WLP (Fuchs, 2011; Sousa and Martin, 2014; Williams *et al.*, 2017; Adam, Borrel and Gribaldo, 2018), this leaves open the possibility for the presence of the WLP in LBCA and its capability of facultative acetogenic growth (Schuchmann and Müller, 2014).

Our ancestral reconstructions indicated that LBCA encoded both membrane-bound F- and V-type ATP synthases. For the F-type ATP synthase, we recover all subunits PP>0.5 for roots 1 and 2, with the exception of subunit b. The alpha, beta, a and c subunits are recovered in all roots with PP >0.9. For the V-type ATP synthase, all subunits are found with PP >0.5 in all roots, except for subunits K and G/H (see Supplementary Tables 1 and 2). Further, we identify the three subunits of a respiratory nitrate reductase (Nar) (Sparacino-Watkins, Stolz and Basu, 2014) with moderate PP values across the three root positions. NarH (K00371) has the highest support across the all three root positions (PP=0.84/0.81/0.6), with NarG (K00370) being recovered with moderate support in Roots 1 and 2 (PP=0.62/0.57/0.26) and NarI (K00374) only being recovered in Root 1 with moderate to low support (0.5/0.48/0.18). This suggests that LBCA might have had the ability for anaerobic respiration of nitrate.

Other components of the electron transport chain are patchily distributed across the different roots, with only a few subunits of each complex being found with PP >0.5 in all given roots (see Supplementary Tables 1 and 2). For instance, while difficult to annotate accurately in the absence of genome content and gene cluster information, we found subunits related to hydrogenases (Vignais, Billoud and Meyer, 2001) (e.g. K18023, PP=0.96/0.97/0.89; K15830, PP=0.99/0.98/0.93), including a protein family comprising large subunits of [Ni-Fe]-hydrogenases (K00333, PP=0.80/0.92/0.58), in agreement with the hypothesis that hydrogen was a primordial electron donor (Lane, Allen and Martin, 2010; Greening *et al.*, 2016). However, we also find the subunits

NuoG (K00336), CytB (K00412), and heme-copper type oxidase (CoxA) (K02274) in all three roots with PP >0.8 (Supplementary Table 1). Support for terminal oxidases are unexpected given that the atmosphere of the early Earth is predicted to be anoxic (Arndt and Nisbet, 2012). Interestingly, these genes were also recovered in a study that inferred the gene set present in the last universal common ancestor (Weiss *et al.*, 2016). One possible explanation for these results is that aerobes are overrepresented among sequenced prokaryotic genomes; the wide distribution of these enzymes across the tips of the tree could then increase their probability to be mapped to the root in comparative analyses. Similarly, the relative patchy distribution of the key enzymes of the WLP across the tips of the bacterial tree could result in relatively low PPs (see above).

Nonetheless, the support for hydrogen as an electron donor, along with the methyl branch of the WLP and the presence of the Rnf complex, point to an anaerobic acetogenic LBCA with carbon dioxide as the electron acceptor. However, more in depth analyses of gene families, and an expanded taxon sampling would be needed to strengthen these results.

Defence mechanisms, CRISPR

Interestingly, we find several CRISPR-associated (Cas) proteins inferred to the root, suggesting the presence of a putative CRISPR-based prokaryotic immune system in LBCA. Highly-supported families (PP>0.95) belong to the Class 1 CRISPR-Cas systems, including the universal and essential Cas protein, Cas1 (K15342, PP=0.96/0.93/0.89), and three Cas proteins belonging to the Type III effector complex, Cas5 (K19139, PP=0.96/0.90/0.86), Cas7 (K19140, PP=0.96/0.90/0.85), and SS (K19138, PP=0.98/0.93/0.89). An additional nine Cas proteins belonging to Type I and III systems, were inferred to be present in LBCA with PP>0.80 (see Supplementary Table 2). The presence of eight additional Cas proteins in the root was supported with PP>0.50. Here we find low support for Cas10 (K19076, PP=0.46/0.26/0), the signature cleavage protein of Type III systems. With the exception of Cas10, we recover all other essential and dispensable elements of a complete Type III system with moderate to high support in our reconstructions suggesting the likely presence of this CRISPR system in LBCA (Supplementary Table 2).

The CRISPR-Cas proteins identified in this analysis were absent in the root of the CPR (Node 496) and rarely recovered across the taxa evaluated in this study (see Supplementary Table 2), suggesting absence of CRISPR system components in members of the CPR clade. These findings are congruent with previous evidence showing that the CPR lack CRISPR-Cas systems possibly due to their host-associated or obligate-symbiont lifestyle (Burstein *et al.*, 2016). Recent metagenomic analyses have uncovered two highly compact Class 2 CRISPR-Cas systems in uncultivated Bacteria, CRISPR-CasX and CRISPR-CasY, the latter of which was encoded in select CPR bacterial genomes (Burstein *et al.*, 2017). We find no support for the key Cas proteins (CasX and CasY) of these novel CRISPR systems in our gene family reconstructions, suggesting a loss of the canonical CRISPR-Cas loci in this lineage and the later acquisitions of these novel systems in certain members of the CPR.

The majority of the CRISPR-Cas proteins recovered in our analysis belong to Class 1 systems, which exhibit greater architectural complexity and diversity in their effector modules compared to their Class 2 counterparts. For this reason, the Class 1 CRISPRs are rarely used for genome modification despite representing up to 90% of CRISPR-Cas systems (Makarova *et al.*, 2015). It is postulated that this multiplex nature of Class 1 effector modules, specifically those of Type III systems, likely arose through a series of duplications and fusions of ancestral RNA recognition motifs (RRM) (Koonin and Makarova, 2019). While the origin, organization, and composition of the CRISPR ancestor remains enigmatic, recent evidence has shown that a built-in signalling pathway in Type III systems comprised of nucleotide-binding (CRISPR-Associated Rossman Fold, CARF) and RNase (Higher-Eukaryote and Prokaryote Nucleotide-binding, HEPN) domains may be a key determinant between programmed cell death or induced dormancy and a targeted immune response (Kazlauskiene *et al.*, 2017; Niewoehner *et al.*, 2017; Koonin and Makarova, 2019).

3.4 Conclusions

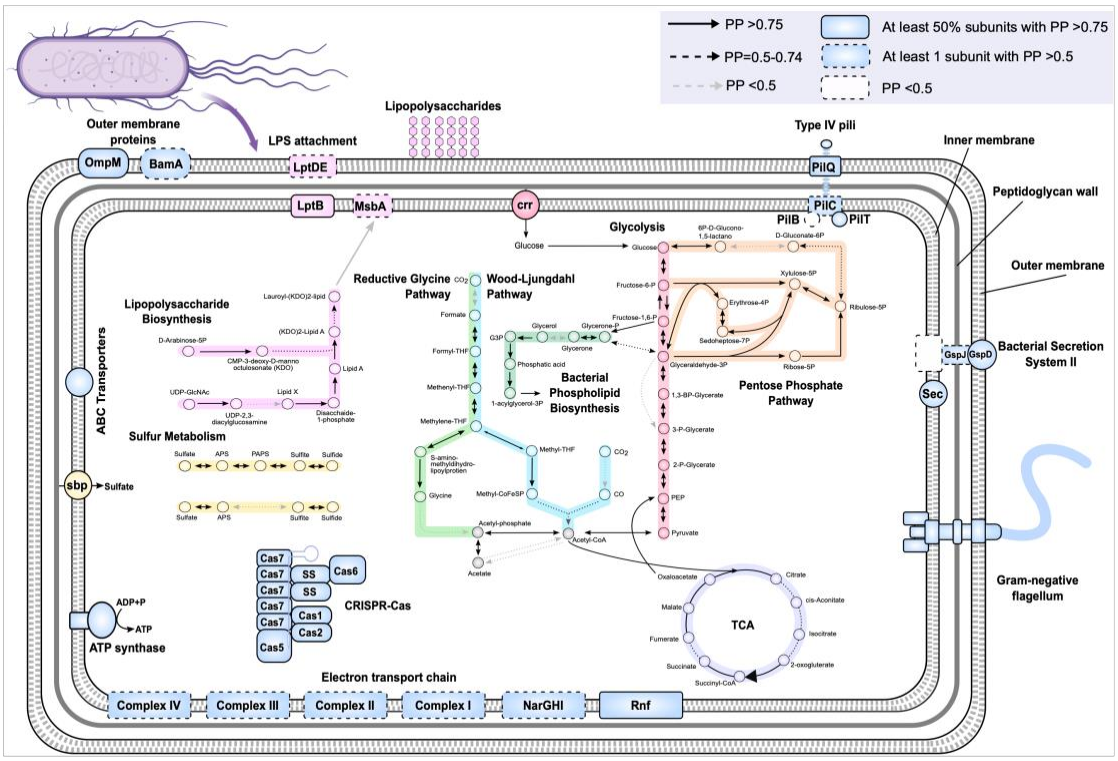
Our inferred ancestral gene set for LBCA includes most of the components of the modern bacterial transcription, translation and DNA replication systems. It also

includes an FtsZ-based cell division machinery and pathways for signal transduction, membrane transport and secretion (Fig. 3.4). Further, we identified proteins involved in bacterial phospholipid biosynthesis, suggesting that LBCA had bacterial-type ester-lipid membranes (Fig. 3.4). We also identified most of the proteins required to synthesise appendages such as flagella and pili as well as to enable quorum sensing, suggesting that LBCA was motile; which is in agreement with the previous suggestion that flagella were present in LBCA (Liu and Ochman, 2007b). Since bacterial genes are typically maintained by strong positive selection (Sela, Wolf and Koonin, 2016), these findings imply that LBCA lived in an environment in which dispersal, chemotaxis and surface attachment were advantageous. We also obtained high root posterior probabilities for proteins mediating outer cell envelope biosynthesis including for LPS, from which we infer that LBCA possessed a double membrane with an LPS layer. Consistent with this inference, we obtained high posterior probabilities for the flagellar subunits FlgH, FlgI and FgA in LBCA, which anchor flagella in diderm membranes (Antunes *et al.*, 2016), and for the Type IV pilus subunit PilQ, which among extant bacteria is specific to diderms (Antunes *et al.*, 2016; Megrian *et al.*, 2020). Altogether, this is consistent with hypotheses (Cavalier-Smith, 2006) in which LBCA was a diderm (Antunes *et al.*, 2016; Megrian *et al.*, 2020), and argues against scenarios in which the Gram-negative double membrane originated by endosymbiosis between monoderms (single-membraned bacteria (Lake, 2009)) or via the arrest of sporulation (Tocheva, Ortega and Jensen, 2016) in a spore-forming monoderm ancestor. Subsequently, diderm-to-monoderm transitions may have occurred on multiple occasions within Bacteria (Antunes *et al.*, 2016; Megrian *et al.*, 2020).

We recovered components of several core pathways for carbohydrate metabolism with high posterior support, including glycolysis, the TCA cycle, and the pentose phosphate pathway (Fig. 3.4). We identified several enzymes of the TCA cycle, although the directionality of the enzymes is difficult to assess (Nunoura *et al.*, 2018). Furthermore, we identified several enzymes of the methyl-branch of the WLP, for acetate biosynthesis as well as components of a putative RNF complex, which together may indicate that LBCA was capable of acetogenic growth (Schuchmann and Müller, 2014) (Fig. 3.4). However, the key enzyme of the WLP, the Carbon monoxide dehydrogenase/acetyl-CoA synthase complex (Adam, Borrel and Gribaldo, 2018), had only moderate to low support for its subunits. Thus, while our analyses provide strong

support for the antiquity of components of the WLP, acetogenesis, the TCA cycle and several other core metabolic pathways, they do not confidently establish the combination of pathways employed by LBCA as distinct from other organisms present at the same time.

Figure 3.4 (below): Ancestral reconstruction of the last bacterial common ancestor (LBCA). *The reconstruction is based on genes that could be mapped to at least one branch within the root region with PP>0.5 (Supplementary Table 2). The presence of a gene within a pathway is indicated as shown in the key. Our analyses suggest that LBCA was a rod-shaped, motile, flagellated double-membraned cell. We recover strong support for central carbon pathways, including glycolysis, the tricarboxylic acid cycle (TCA) and the pentose phosphate pathway. We did not find unequivocal evidence for the presence of a carbon fixation pathway, although we found moderate support for components of both the Wood-Ljungdahl pathway and the reverse TCA cycle. Though not depicted here, our analyses suggest that the machinery for transcription, translation, tRNA and amino acid biosynthesis, homologous recombination, nucleotide excision and repair, and quorum sensing was also present in LBCA (Supplementary Table 1).*



Finally, our reconstruction also indicated high posterior support for several elements of an adaptive immunity CRISPR-Cas system (Makarova *et al.*, 2011; Koonin and Makarova, 2019), including the universally conserved Cas endonuclease, Cas1, which is essential for spacer acquisition and insertion into CRISPR cassettes (Nuñez *et al.*, 2014; Makarova *et al.*, 2015). Interestingly, highly supported CRISPR components in LBCA belong primarily to Class 1 systems, specifically Type I and Type III, which exhibit greater modular diversity than their Class 2 counterparts, (Koonin and Makarova, 2019). We recovered a near complete prototypical Type III CRISPR system, providing strong support for its presence in LBCA. Among other roles, CRISPR systems are crucial in antiviral defence and activate in response to viral exposure (Barrangou *et al.*, 2007); therefore these findings are consistent with hypotheses suggesting that LBCA already co-evolved with parasitic replicators such as bacteriophages (Koonin, 2014; Krupovic, Dolja and Koonin, 2019).

Chapter 4

Genomic evolution of Bacteria through time

A version of this chapter forms part of a paper under revision in collaboration with Adrián A. Davín, Tara Mahendrarajah, Anja Spang, Philip Hugenholtz, Gergely J. Szöllősi, and Tom A. Williams. Gareth A. Coleman is the first author of the paper. The project was conceived by TAW, GJSz, PH, AS, GAC and AAD. Relative time and verticality analyses were carried out by AAD, GJSz, TW and GAC, with interpretation and writing carried out by all authors. Metabolic reconstructions, including all interpretation and writing thereof, were carried out by GAC.

Paper preprint as:
Coleman, G.A., Davín, A.A., Mahendrarajah, T., Spang, A.A., Hugenholtz, P., Szöllősi, G.J. and Williams, T.A., 2020. A rooted phylogeny resolves early bacterial evolution. *bioRxiv*.

BioRxiv preprint for the paper can be found here:
https://www.biorxiv.org/content/10.1101/2020.07.15.205187v1

Abstract

Given the huge genetic and metabolic diversity, abundance, and ubiquity of Bacteria in the modern environment, understanding how diverse physiologies and morphologies evolved in the early stages of bacterial evolution is central to our understanding of the history of life. Having previously used methods of modelling genome evolution using gene duplication, transfer and loss events to reconstruct the ancestral gene content of the last bacterial common ancestor (LBCA), we now use these methods to explore how LBCA gave rise to the major clades of extant Bacteria. We infer the early radiation of Terrabacteria, particularly CPR, while Cyanobacteria and Gracilicutes emerged later in bacterial diversification. We find that early bacterial evolution was likely dominated by autotrophic, possibly acetogenic, motile diderm cells, living alongside highly reduced and metabolically streamlined cells, followed by the later rise of photosynthetic and aerobic prokaryotes, and the eventually the Eukaryotes.

4.1 Introduction

Bacteria have a profound effect on the physical environment and have been interacting with the geosphere for billions of years. Inferring the order of events within bacterial diversification is important for understanding how the metabolisms which shape these interactions have evolved through time. Transfers contain information about relative divergence times because donor lineages must be as old as the recipient lineages (Cédric Chauve *et al.*, 2017; Davín *et al.*, 2018), allowing patterns of gene transfers to infer the relative order of divergences within the tree. Such an approach has been used to infer the early radiation of methanogenic Archaea, and the relatively late radiation of the highly genomically reduced DPANN (Davín *et al.*, 2018). Applying such analyses to Bacteria will help us understand the relative timing of different lineages. For example, although we have shown that CPR are not the earliest diverging clade within Bacteria (Chapter 2), determining the relative timing of their radiation is important in elucidating the role played by such organisms in the early stages of bacterial diversification. Similarly, inferring the relative timings of the appearance of Cyanobacteria and Alphaproteboacteria not only have implications for our understanding of the evolution of diverse metabolisms (notably the timing of the evolution of oxygenic photosynthesis in Cyanobacteria relative to other lineages in the tree) but also the origin of Eukaryotes. Eukaryotic cells are thought to have arisen via a symbiosis between an archaeal host cell and a bacterial endosymbiont which later evolved into the mitochondrion (Embley and Martin 2006; Martin et al. 2015; Eme et al. 2017; Roger et al. 2017), with plastids in photosynthetic eukaryotes derived from Cyanobacteria. The relative emergence of Cyanobacteria and Alphaproteobacteria therefore place constraints on the timing of the emergence of eukaryotic cells.

Determining the evolution and development of the core carbon pathways and energy metabolisms of the deepest nodes within the bacterial tree is an important step in understanding the evolution of microbial communities and their effects on the environment. Specifically, the evolution of carbon fixation, and possible changes from autotrophy to heterotrophy within the bacterial tree, are important aspects to explore. Much evidence has pointed to the ancient origin and use of the Wood-Ljungdahl Pathway (WPL) (Fuchs, 2011; Sousa and Martin, 2014; Weiss *et al.*, 2016; Williams *et*

al., 2017; Adam, Borrel and Gribaldo, 2018), and the reverse tricarboxylic acid (TCA) cycle (Wächtershäuser, 1990; Cody *et al.*, 2001; Smith and Morowitz, 2004; Nunoura *et al.*, 2018) as carbon fixation pathways. In Chapter 3, we found possible support for both pathways in the last bacterial common ancestor (LBCA), although the evidence is difficult to interpret. It is therefore important to determine whether subsequent nodes possess similar abilities to LBCA, and at which points along the tree these metabolic capabilities change.

Reconstructing the gene content at different nodes can also help us understand the evolution of cell morphology, in particular the architecture of the cell envelope and the evolution of structure involved in cell motility, and test hypotheses on the origins of these characteristics. As explained in the previous chapter, the “lipid divide” (Koga, 2011) refers to the difference in type of membrane phospholipids produced by Archaea and Bacteria, and has been touted as a hallmark difference between the two domains, although recent evidence has challenged these assumptions and found the divide to be less clear cut than previously thought (Sinninghe Damsté *et al.*, 2002; Weijers *et al.*, 2006; Damsté, Rijpstra, *et al.*, 2007; Goldfine, 2010; Villanueva, Schouten and Damsté, 2017; Caforio *et al.*, 2018; Coleman, Pancost and Williams, 2019). The presence in many Bacteria of genes involved in the biosynthesis of archaeal-like lipids (Villanueva, Schouten and Damsté, 2017; Coleman, Pancost and Williams, 2019) has raised questions pertaining to the relative antiquity of both pathways, specifically which phospholipids were being produced by early Bacteria (Coleman, Pancost and Williams, 2019). We found some evidence for the production of bacterial phospholipids in LBCA (Chapter 3), but it would be pertinent to discern whether other early nodes with the bacterial tree were producing bacterial- or archaeal-type phospholipids. Many modern Bacteria also possess an outer membrane, the origin of which remains debated. Monoderm-first scenarios have been advocated whereby the outer membrane evolved stepwise via gene insertions (Gupta, 2011), by endosymbiosis between monoderms (single-membraned bacteria (Lake, 2009)), or via the arrest of sporulation (Tocheva, Ortega and Jensen, 2016) in a spore-forming monoderm ancestor. In Chapter 3, we also find that LBCA possesses an outer membrane, which supports Diderm-first hypothesis regarding the evolution of diderm envelope architecture (Cavalier-Smith, 2002; Megrian *et al.*, 2020). To strengthen evidence for the Diderm-first scenario, we

must explore the presence of the outer membrane in deep nodes, particularly with the Terrabacteria, where many modern lineages have lost the outer membrane.

To answer these questions, we build on results from Chapter 3 by using patterns of gene transfer to infer the relative timing of different events within bacterial evolution, and expanding our metabolic reconstructions to other deep nodes within the bacterial tree.

4.2 Methods

Inference of relative divergence times of bacterial clades

We parsed the transfers inferred using ALEml_undated and discarded those with posterior probability <0.05. We used bootstrapping to estimate constraint support in the following way: for each of the three candidate species trees, we sampled the gene families 100 times with replacement and, for each replicate, converted detected transfers to constraints and performed a MaxTiC analysis (Cedric Chauve *et al.*, 2017; Davín *et al.*, 2018). A total of 8743, 8629, and 9079 constraints were recovered in at least 95/100 replicates for the 3 possible roots respectively, and we used this subset of highly supported constraints in our final analysis. We generated 1000 time orders compatible with those constraints for every root. We then ranked all interior nodes on the tree, with the root node having rank 0 and the most recent speciation node having rank 263.

Metabolic reconstruction

Metabolic reconstructions were carried out as detailed in Chapter 3. Briefly, protein sequences form the GTDB dataset were annotated using both KEGG and COG databases. Sets of COG families were produced, and those with more than 4 sequences were reconciled with the rooted species trees using a specific origination prior for each family based on its COG category (see Chapter 3 for details). As in Chapter 3, initial metabolic infers were based on gene families with a posterior presence probability (PP) of >0.95, but as this was found to be too conservative for most metabolic pathways, we investigated the PPs of key pathways identified from the initial PP cut-off and inferred the presence of a pathway of protein complex or pathway

if the majority of its components were found with PPs >0.5. We explored eight nodes (Fig. 4.1), namely the ancestors of Terrabacteria and Gracilicutes respectively, the common ancestors of Firmicutes and Actinobacteriota, and Chloroflexota and CPR respectively, and all the ancestors of the aforementioned individual phyla. We do not explore any nodes involving Fusobacteriota, or any of the FASSyT/DST taxa (see Chapter 2) and as a result the ambiguity within the root region has little effect on the genes recovered at the nodes surveyed.

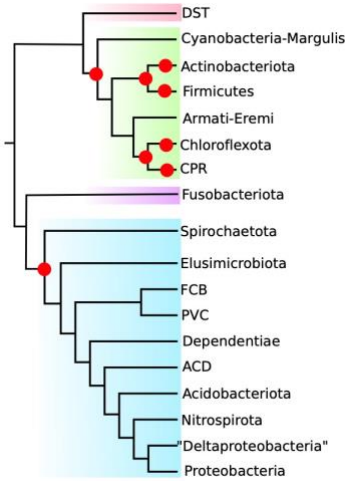


Fig. 4.1 *Nodes for which we inferred ancestral gene content. Note we do not explore any nodes involving Fusobacteriota or FASSyT/DST due to uncertainty of root position. FCB are the Fibrobacterota, Chlorobiota, Bacteroides, and related lineages; PVC are the Planctomycetes, Verrucomicrobia, Chlamydiae, and related lineages; DST are the Deinococcota, Synergistota, and Thermatogota; ACD are Aquificota, Campylobacterota, and Deferribacterota.*

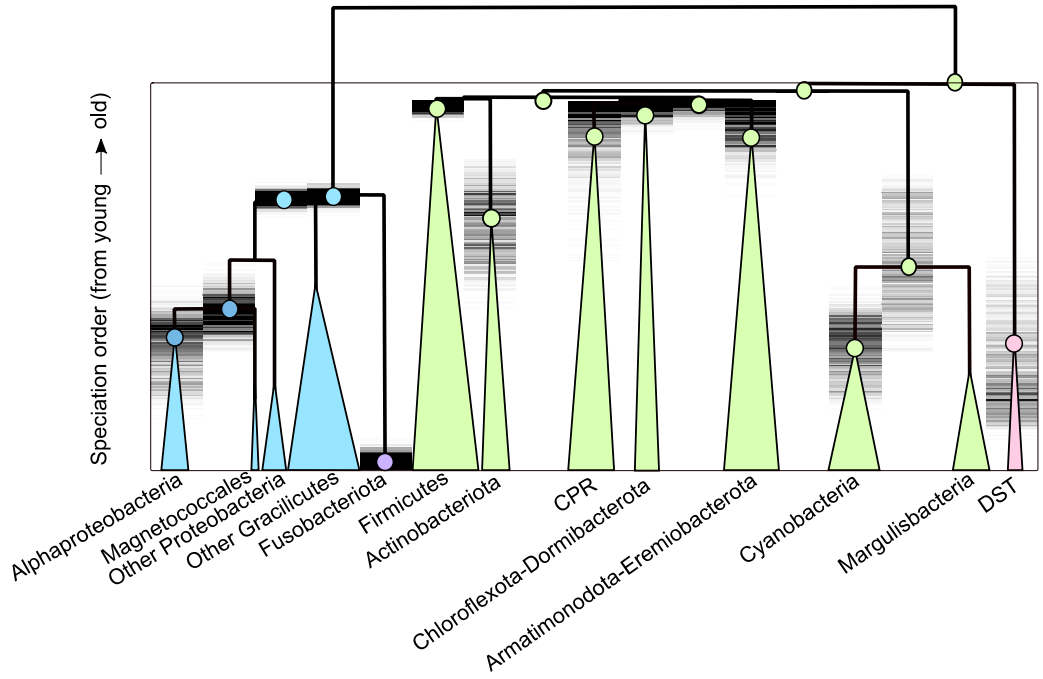
4.3 Results and Discussion

Relative dating of the bacterial tree

Transfers contain information about the relative timing of divergences because donors must be at least as old as their recipients (Cedric Chauve *et al.*, 2017; Davín *et al.*, 2018). Since inferred transfer events are uncertain, we used only high-confidence relative age constraints recovered in at least 95/100 bootstrap replicates (see Methods) to establish the relative ages of bacterial clades (Fig. 4.2). These analyses suggest that several groups within Terrabacteria are older than the entire Gracilicutes

radiation, including the crown groups of the CPR (97.4% of sampled time orders) and Firmicutes (100% of sampled time orders). Despite being a derived lineage within Terrabacteria, the analysis suggests that the radiation of CPR was one of the earliest events during the diversification of Bacteria (Fig. 4.2). The early radiation of CPR, taken together with evidence for the basal position of the analogous DPANN superphylum within Archaea (Castelle *et al.*, 2015; Williams *et al.*, 2017), implies that highly reduced, metabolic minimalist prokaryotes have been a part of microbial communities throughout the history of cellular life (Beam *et al.*, 2020), although it should be noted that relative dating of Archaea using the same method showed crown-group DPANN to be a relatively late occurrence within archaeal evolution (Davín *et al.*, 2018). By contrast, the emergence of the Alphaproteobacteria and the photosynthetic Cyanobacteria were relatively late events during bacterial evolution: the divergence between Alphaproteobacteria and *Magnetococcales* was the 153rd of 264 internal divergences (median rank), while the divergence of photosynthetic Cyanobacteria from their closest relatives had a median rank of 172nd (Fig. 4.2). These divergences confirm that the mitochondrial and plastid endosymbioses, and therefore the origin of eukaryotic cells, occurred during the later stages of bacterial diversification (Parfrey *et al.*, 2011; Eme *et al.*, 2014b; Knoll, 2014; Nicholas J. Butterfield, 2015; Betts *et al.*, 2018).

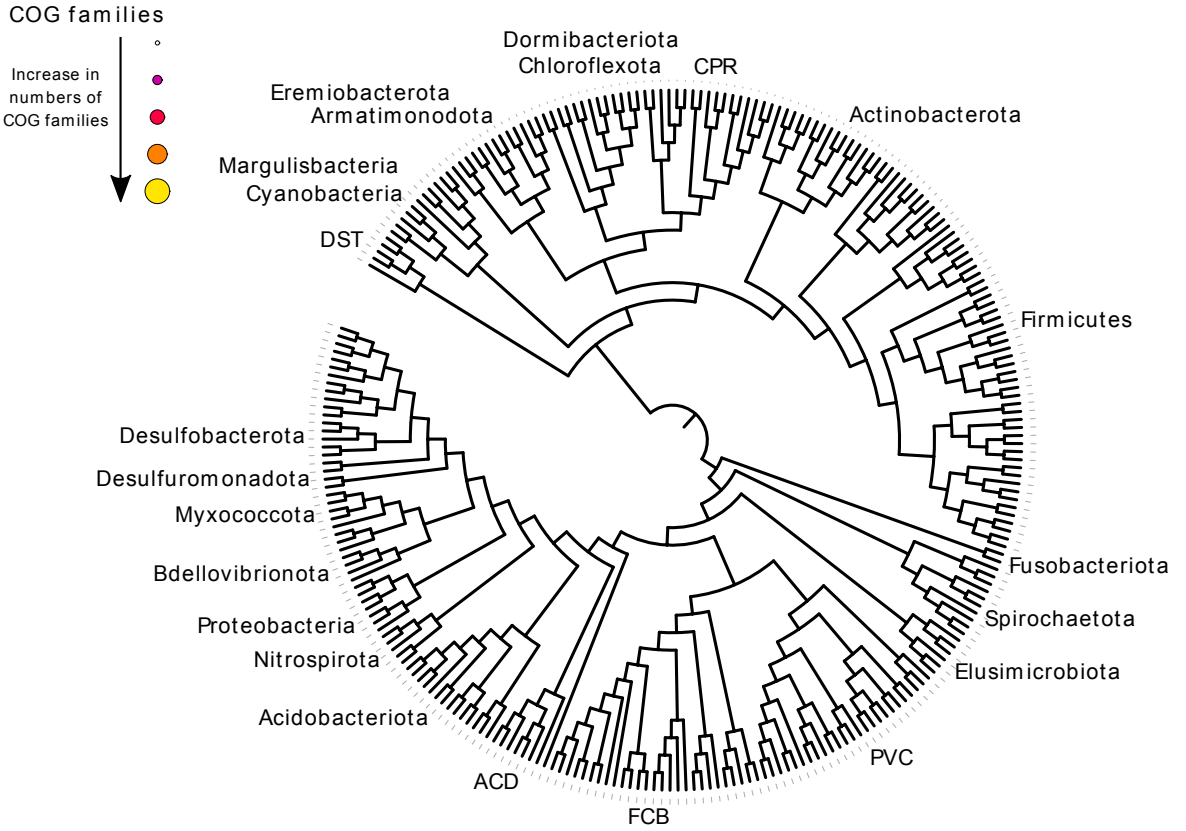
Fig. 4.2 (below) Relative ages of bacterial clades. *We used the relative time information provided by directional (donor-to-recipient) patterns of gene transfer to infer the relative ages of bacterial clades. Node numbers in the cladogram (left) correspond to rows in the speciation plot (right). Uncertainties represent the range of sampled time orders that are consistent with high-confidence constraints implied by gene transfers. Following the divergence between Terrabacteria and Gracilicutes, the earliest radiations of extant groups were among Terrabacteria, including CPR and Firmicutes. Note that gene transfers indicate the order of branching events, but provide no information about the absolute time intervals separating events. The geological record provides evidence for oxygenic photosynthesis prior to 3.2 Gya (Betts et al., 2018), suggesting that divergence 5 - and by extension all earlier divergences - occurred prior to this date. DST are the Deinococcota, Synergistota, and Thermotogota.*

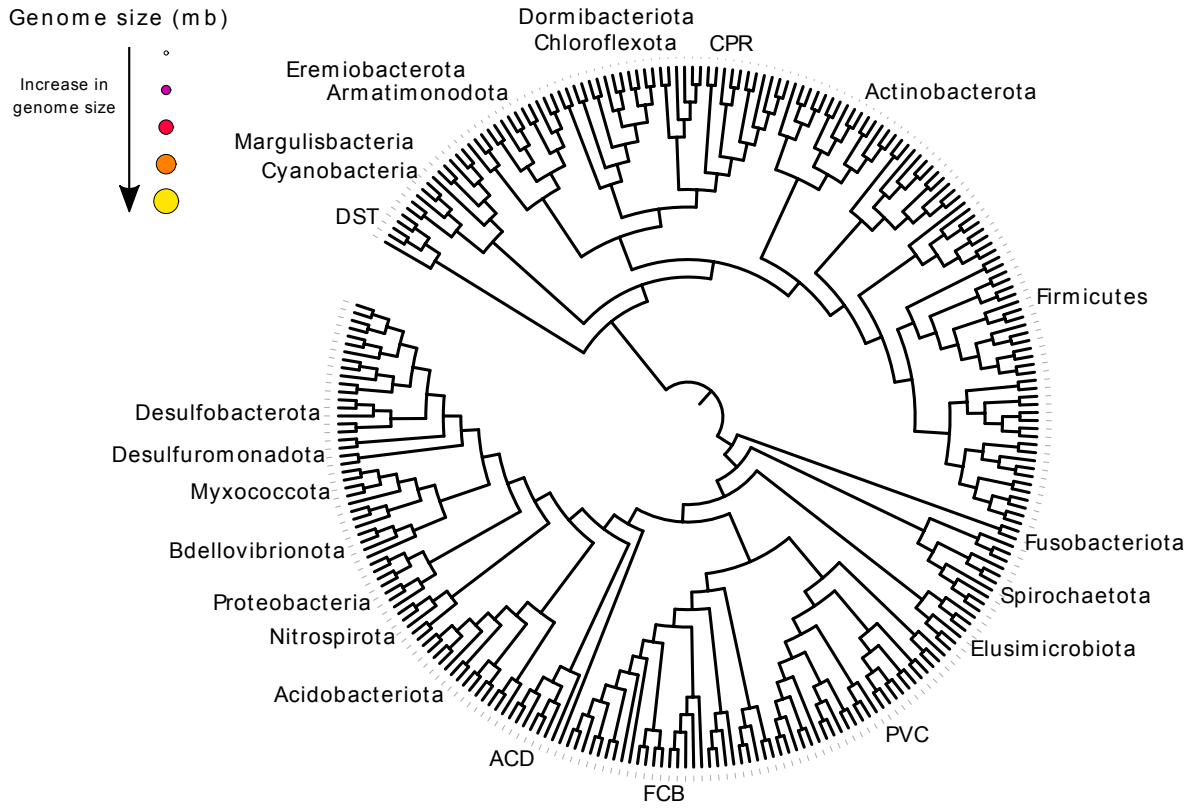


Genome evolution and size through time

Under all three roots, the trend in genome size evolution from LBCA to modern taxa is an ongoing moderate increase through time in estimated COG family complements and genome sizes. Genome reduction of 0.47-0.56Mb on the CPR stem lineage after divergence from their common ancestor with Chloroflexota is the most significant departure from this trend (Fig. 4.3). COG families lost on the CPR stem include components of the electron transport chain, carbon metabolism, flagellar biosynthesis and motor switch proteins, amino acid biosynthesis, the Clp protease subunit ClpX and RNA polymerase sigma factor-54, in agreement with previous findings (Castelle *et al.*, 2018) (Supplementary Table 3).

Fig. 4.3 (below) Evolution of COG family repertoires and inferred genome size over the bacterial tree. *(a) The inferred number of COG family members and (b) inferred genome size at each internal node of the tree. Genome sizes were predicted from the relationship between COG family members and genome size among extant Bacteria (LOESS regression). Circle diameter and colour are proportional to family number or genome size. FCB are the Fibrobacterota, Chlorobiota, Bacteroides, and related lineages; PVC are the Planctomycetes, Verrucomicrobia, Chlamydiae, and related lineages; DST are the Deinococcous, Synergistota, and Thermatogota; ACD are Aquificota, Campylobacterota, and Deferribacterota. The figure depicts inferences for root 1 (as shown in Chapter 2, Fig. 2.7(a)); the data for all three roots are provided in Supplementary Table 4.*


COG families
Increase in numbers of COG families
Dormibacteriota
Chloroflexota
CPR
Eremiobacterota
Armatimonodota
Actinobacterota
Margulisbacteria
Cyanobacteria
DST
Firmicutes
Desulfobacterota
Desulfuromonadota
Myxococcota
Fusobacteriota
Bdellovibrionota
Spirochaetota
Proteobacteria
Elusimicrobiota
Nitrospirota
Acidobacteriota
PVC
ACD
FCB


Genome size (mb)
Increase in genome size
Dormibacteriota
Chloroflexota
CPR
Eremiobacterota
Armatimonodota
Actinobacterota
Margulisbacteria
Cyanobacteria
DST
Firmicutes
Desulfobacterota
Desulfuromonadota
Myxococcota
Fusobacteriota
Bdellovibrionota
Spirochaetota
Proteobacteria
Elusimicrobiota
Nitrospirota
Acidobacteriota
PVC
ACD
FCB

Evolution of carbon metabolism after LBCA

Our reconstructions of the central metabolic pathways are similar to those inferred in LBCA. We find moderate to strong support (PP=0.5-1) for the presence of glycolysis in all nodes, attesting to its ancient origins (Romano and Conway, 1996). Support for the tricarboxylic acid (TCA) cycle and pentose phosphate pathway are less strong, with very patchy distributions of proteins across different nodes (Fig. 4.4). The ancestor of CPR in particular seems to lack the majority of the components of these pathways, with only two proteins (both in the same step of the pathway) and one protein from the TCA and pentose phosphate pathway recovered respectively.

When looking for evidence of carbon fixation, components of various pathways were patchy. Evidence for carbon fixation is largely absent in the ancestor of Actinobacteriota. Although some autotrophic examples within Actinobacteriota are known (Norris *et al.*, 2011), the large majority are heterotrophic (Lechevalier and Lechevalier, 1965; Barka *et al.*, 2016), and therefore the ancestor of Actinobacteriota was also likely heterotrophic. Of the several possible carbon fixation pathways, there is no evidence for the presence of any variants of the 3-hydroxypropionate bicycle in any of the nodes. The apparent presence of components of the Calvin Cycle (also known as the reductive pentose phosphate pathway) most likely represent the pentose phosphate pathway. The enzyme RuBisCO is unique to the Calvin Cycle, and is not found in any nodes surveyed. We have some moderate support of both subunits of oxoacid:ferredoxin oxidoreductase in all nodes (Fig. 4.4), which may function in the TCA in both the oxidative and reductive direction. However, as noted above, other components of the TCA cycle are patchily distributed, and we do not recover the hallmark enzyme ATP citrate lyase in any nodes. The support for the reverse TCA cycle is highest in the ancestor of Terrabacteria, with six steps of the pathway recovered with PP>0.5 (Fig. 4.4), although still lacking ATP citrate lyase. We therefore have little evidence to suggest the presence of the reverse TCA cycle in any of these nodes.

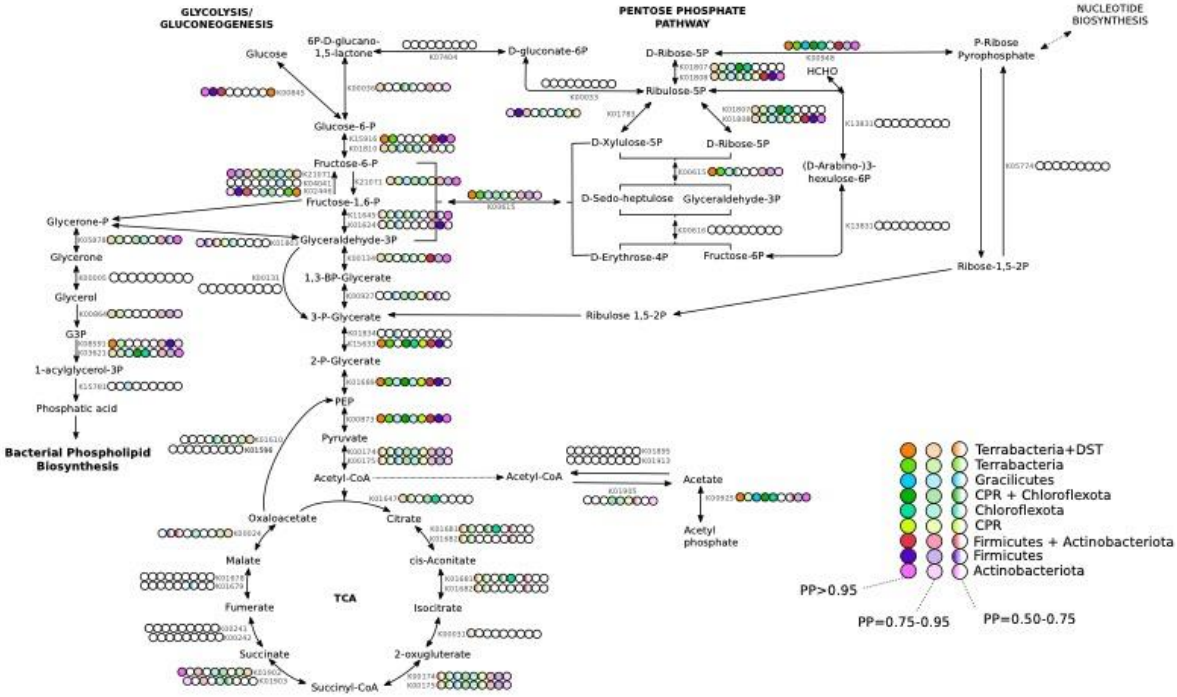


Fig 4.4 *Metabolic map of the central carbohydrate pathways (glycolysis, tricarboxylic acid cycle and pentose phosphate cycle) inferred in the ancestors of Terrabacteria and DTS, Terrabacteria, Gracilicutes, Chloroflexota and CPR, Chloroflexota, CPR, Firmicutes and Actinobacteriota, Firmicutes, and Actinobacteriota respectively. The reconstruction is based on genes that could be mapped to a given node with PP >0.5. The presence of a gene within a pathway is indicated as shown in the key. Annotations and PP values for all KOs can be found in Supplementary Table 1.*

Acetogenesis in early Bacteria

The components of the methyl branch of the WLP are mapped with moderate to strong support (PP= 0.5-0.95) across all the nodes, except the common ancestor of CPR and Chloroflexota and both their respective ancestors, which lack some of the steps (Fig. 4.5). However, as with LBCA we do not find most of the components of the hallmark enzyme of the WLP, the carbon monoxide dehydrogenase/acetyl-CoA synthase (CODH/ACS), finding only very moderate support for a single subunit (K14138, PP= 0.66) in the ancestor of Gracilicutes. As the methyl-branch of the WLP may be used in other metabolic pathways (Stover, 2009; Brosnan and Brosnan, 2016), it is difficult to assign the presence of WLP to any node based solely on the presence of the methyl branch. CODH/ACS is not especially common among either Gracilicutes or Terrabacteria, although it does have wide phylogenetic spread, indicating either

multiple transfers of these genes, or a presence in the ancestors of Gracilicutes and Terrabacteria followed by multiple losses.

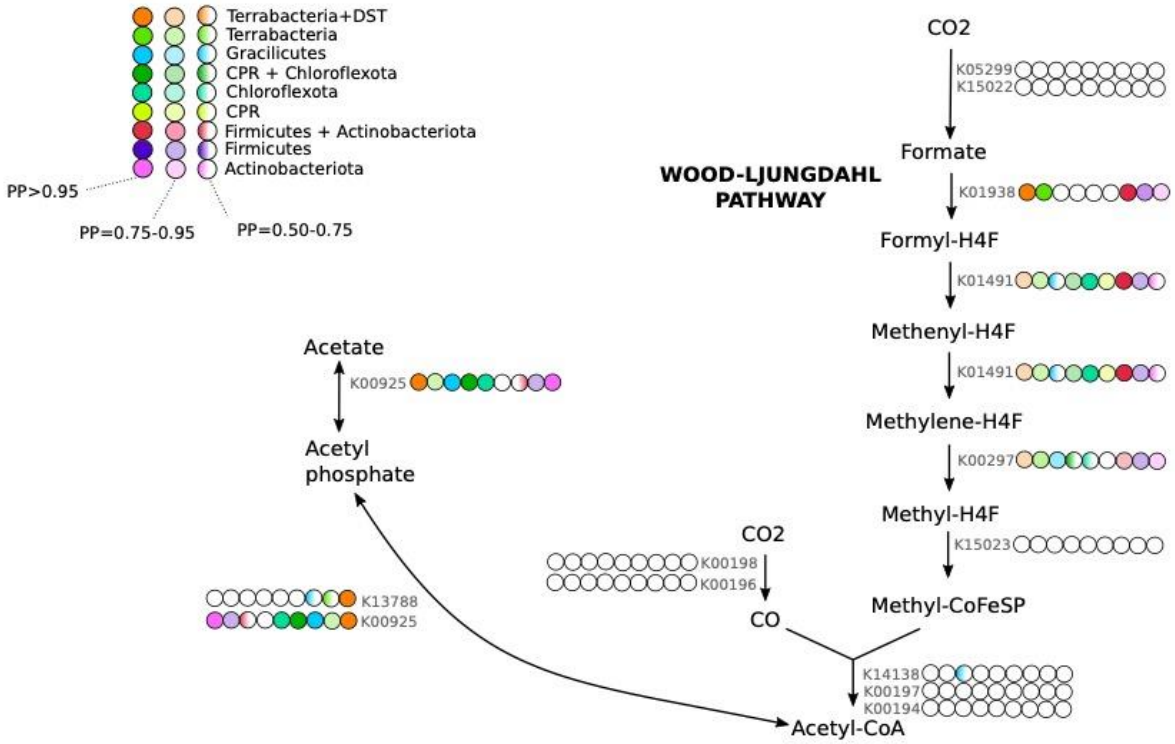


Fig. 4.5 *Metabolic map of acetogenesis and the Wood-Ljungdahl Pathway inferred in the ancestors of Terrabacteria and DTS, Terrabacteria, Gracilicutes, Chloroflexota and CPR, Chloroflexota, CPR, Firmicutes and Actinobacteriota, Firmicutes, and Actinobacteriota respectively. The reconstruction is based on genes that could be mapped to a given node with PP >0.5. The presence of a gene within a pathway is indicated as shown in the key. Annotations and PP values for all KOs can be found in Supplementary Table 1.*

Acetate kinase was recovered with strong support (PP>0.75) in the Gracilicutes and Terrabacteria ancestor, as well as the common ancestor of Chloroflexota and CPR and the ancestor of Chloroflexota. Support was more moderate (PP=0.5-0.75) in the common ancestor of Firmicutes and Actinobacteriota and their respective ancestors, and was absent in the ancestor of CPR (Fig. 4.5). We recovered phosphate acyltransferase only with moderate (PP=0.5-0.75) support in the ancestors of Gracilicutes and Terrabacteria, and not recovered in other nodes (Fig. 4.5).

As it was found to be present in LBCA (Chapter 3), we further investigated the presence of the Na^+-translocating ferredoxin:NAD^+ oxidoreductase (Rnf) complex, which comprises six subunits coded by the genes *rnfA* (K03617), *rnfB* (K03616), *rnfC* (K03615), *rnfD* (K03614), *rnfE* (K03613), and *rnfG* (K03612) respectively. We recover five of these components (all except *rnfG*) with PP>0.5 in the ancestor of Gracilicutes (Supplementary Table 1). Given the role that the Rnf complex may play in the WLP in acetogens (Buckel and Thauer, 2013; Müller, Chowdhury and Basen, 2018), as well as evidence for the antiquity of the pathway (Chapter 3 of this thesis, (Fuchs, 2011; Sousa and Martin, 2014; Williams *et al.*, 2017; Adam, Borrel and Gribaldo, 2018)), we have some evidence to suggest that the ancestor of Gracilicutes, like LBCA, possessed the WLP and may have been facultatively acetogenic. Support is less strong in other nodes (Supplementary Table 1). In the ancestral Terrabacteria, we recover only *rnfA* (PP=0.96). Four of the six genes are recovered in the ancestor of Firmicutes, providing some evidence that it was acetogenic. Only two were recovered in the ancestor of Actinobacteriota, indicating that it was unlikely to have been able to use this pathway. Indeed, as discussed above, the last ancestor of Actinobacteriota was most likely heterotrophic. None of these genes are recovered in the common ancestor of Chloroflexota and CPR or their respective ancestors, so we cannot say that acetogenesis was present in these nodes. It is possible that the ancestor of Terrabacteria was acetogenic, with this pathway being subsequently lost in most lineages, although maintained in Firmicutes. However, the evidence is only moderate.

Energy generation and respiratory complexes

As in LBCA, our analyses indicate the presence of a F-type ATP synthase, including components of both the F_0 regions (which is embedded into the cell membrane) and the F_1 region respectively, in all nodes. Several of the subunits, e.g. alpha, gamma, a and c subunits, are recovered in the majority of the nodes with high support (PP>0.9) (Supplementary Table 1). Other subunits are more patchily distributed. A V-type ATP synthase is also recovered in the ancestor of Terrabacteria, Gracilicutes, the common ancestor of Firmicutes and Actinobacteriota and their respective ancestors. The subunits a (K02117), b (K02118) and d (K02120, stalk) of the V_1 region are recovered in all these nodes with moderate to high support (PP=0.5-1) (Supplementary Table 1). Of the V_0 region, which is embedded in the cell membrane, subunits e and i are widely recovered. No subunits for V-type ATP synthases are found in the ancestor of

Chloroflexota, and only the a subunit is found in the ancestor of CPR and the common ancestor of both Chloroflexota and CPR respectively. The V-type ATP synthase was most likely present in most early lineages, and lost on the branch to Chloroflexota and CPR.

As with our reconstruction of LBCA, we find a patchy distribution of other components of the electron transport chain, with a small number of subunits for each complex found with PP >0.5, except in the ancestor of CPR where we find no components of the electron transport chain, other than the aforementioned ATP synthases (Supplementary Table 1). Unlike our reconstruction of LBCA, we do not find strong support for the presence of respiratory nitrate reductase (Nar), so we cannot determine the ability for anaerobic respiration of nitrate in any of the subsequent nodes. We recover some subunits related to hydrogenase in the Gracilicutes ancestor (K18023, PP=0.92; K00333, PP=0.73), and in the Terrabacteria ancestor (K18023, PP=0.92) suggesting hydrogen as the election doner, in line with our reconstruction of LBCA, and with previous studies (Lane, Allen and Martin, 2010; Greening *et al.*, 2016)).

Again, similarly to LBCA, we find support for the terminal oxidases NuoG (K00336), CytB (K00412) in all nodes except the ancestors of Firmicutes and CPR. NuoG has strong support in the ancestor of Terrabacteria and Gracilicutes (PP=0.85 and 0.91 respectively). Support is more moderate in other nodes (PP=0.5-0.75). CytB has moderate to strong support (PP>0.7), but support is weaker in the ancestor of Gracilicutes (PP=0.53). The nodes with the strongest support for the presence of terminal oxidases are the common ancestor of Chloroflexota and CPR, and the ancestor of Chloroflexota. Both nodes not only recover CytB with strong support (PP=0.96 and 1 respectively) and NuoG with moderate support (PP=0.58 and 0.67 respectively) but also have several heme-copper type oxidases, notably coxA (K02274), coxB (K02275) and coxC (K02276), with moderate to high supports (PP=0.6-1). It is possible that aerobic respiration was acquired on the branch to Chloroflexota and CPR before being lost in CPR. It should also be noted that, other than in the common ancestor of Chloroflexota and CPR and ancestor of Chloroflexota, we do not find evidence for any heme-copper type oxidases, and that other components for complexes which comprise these terminal oxidases are patchy in their distribution or are absent, so we cannot confidently map terminal oxidases to any of

these nodes. Furthermore, all of these clades predate the radiation of photosynthetic Cyanobacteria in our relative dating analysis, including in the case of Chloroflexota and CPR, which may indicate they predate the oxygenation of the atmosphere, although this is contentious (Betts *et al.*, 2018).

As was the case with LBCA, it is difficult to determine which electron acceptor these early Bacteria were using. However, based on the presence of components of the acetogenesis pathway, carbon dioxide was the most likely candidate, implying that the ancestors of most clades were anaerobic. This is congruent with models of the atmosphere of the early Earth (Arndt and Nisbet, 2012). However, more in depth exploration of the histories of these genes would be needed to bring greater clarity to this issue, and other electron acceptors, including oxygen, cannot be confidently rejected by our analyses.

Evolution of membrane phospholipids in Bacteria

We inferred the presence of bacterial-type G3P membrane in LBCA (Chapter 3) and expect similar inferences in subsequent nodes, given the widespread use of G3P lipids in modern Bacteria. We recover moderate to high support for G3P dehydrogenase (GpsA, K00057) in all nodes we surveyed (Table 4.1), with the exception of the ancestor of CPR and the ancestor of Gracilicutes. We additionally recover glycerol kinase in the Terrabacteria ancestor (K00864, PP=0.84). Similarly to GpsA, PlsX (K03621, which attaches the first fatty to G3P along with PlsY), was recovered with moderate to strong support (0.78-0.98) across all nodes except the ancestor of CPR. Other proteins involved in phospholipid biosynthesis are more sparsely distributed across the tree (Table 4.1). PlsY (K08591) was found with high support in the Terrabacteria ancestor (PP=0.93), as well as in both Firmicutes and Actinobacteriota and their common ancestor (PP=0.94, 0.98 and 0.81 respectively). It was not recovered in the Gracilicute ancestor, or in the ancestor of CPR and Chloroflexota and descendent nodes. No PlsC is recovered, except a putative PlsC in the Gracilicutes ancestor (K15781, PP=0.75). While the distribution is sparse, given the presence for this pathway in LBCA (Chapter 3) and the distribution of these genes in modern taxa, it is probable that the ability the produce bacterial G3P lipids was present in all the deep nodes of the tree, and subsequently lost in CPR. In contrast to the above, we do not find any strong evidence for the presence of archaeal type lipids in any nodes,

despite evidence for the presence in modern representatives of several bacterial phyla (Villanueva, Schouten and Damsté, 2017; Coleman, Pancost and Williams, 2019), and their possible ancient presence within Bacteria (Coleman, Pancost and Williams, 2019).

FAM	kegg_id	kegg_description	Terra+DST	Terra	Graci	Chlo+CPR	Chlo	CPR	F+A	F	A
COG0344	**K08591**	**acyl_phosphate:glycerol-3-phosphate_acyltransferase_[EC:2.3.1.275]**	**0.97**	**0.93**	**0.42**	**0.46**	**0.49**	**0.13**	**0.94**	**0.98**	**0.81**
COG0554	**K00864**	**glycerol_kinase_[EC:2.7.1.30]**	**0.88**	**0.84**	**0.21**	**0.15**	**0.25**	**0.03**	**0.8**	**0.78**	**0.75**
COG1368	K05878	phosphoenolpyruvate---glycerone_phosphotransferase_subunit_DhaK_[EC:2.7.1.121]	0.86	0.86	0.12	0.85	0.93	0.22	0.93	0.71	1
COG4589	K00981	phosphatidate_cytidylyltransferase_[EC:2.7.7.41]	0.82	0.42	0.38	0.63	0.72	0.26	0.05	0.05	0.01
COG0558	K06131	cardiolipin_synthase_A/B_[EC:2.7.8.-]	0.79	0.72	0.54	0.38	0.7	0.22	0.18	0.13	0.14
COG0416	**K03621**	**phosphate_acyltransferase_[EC:2.3.1.274]**	**0.78**	**0.8**	**0.82**	**0.9**	**1**	**0.25**	**0.98**	**0.83**	**0.97**
COG1597	K07029	diacylglycerol_kinase_(ATP)_[EC:2.7.1.107]	0.68	0.85	0.33	0.93	0.95	0.44	0.91	0.92	0.94
COG2376	K05881	phosphoenolpyruvate---glycerone_phosphotransferase_subunit_DhaM_[EC:2.7.1.121]	0.58	0.4	0.34	0.12	0.19	0.02	0.41	0.48	0.16
COG0575	K00981	phosphatidate_cytidylyltransferase_[EC:2.7.7.41]	0.57	0.88	0.7	0.49	0.44	0.86	0.98	0.99	0.9
COG0644	K01126	glycerophosphoryl_diester_phosphodiesterase_[EC:3.1.4.46]	0.49	0.21	0.41	0.26	0.49	0.1	0.11	0.22	0.23
COG0240	**K00057**	**glycerol-3-phosphate_dehydrogenase_(NAD(P)+)_[EC:1.1.1.94]**	**0.41**	**0.85**	**0.15**	**0.78**	**0.67**	**0.27**	**0.98**	**0.96**	**1**
COG1502	K03736	ethanolamine_ammonia-lyase_small_subunit_[EC:4.3.1.7]	0.35	0.19	0.17	0.02	0.01	0.01	0.14	0.15	0.01
COG0371	K17830	digeranylgeranylglycerophospholipid_reductase_[EC:1.3.1.101_1.3.7.11]	0.31	0.4	0.04	0.22	0.24	0.1	0.5	0.46	0.82
COG4303	K01095	phosphatidylglycerophosphatase_A_[EC:3.1.3.27]	0.29	0.06	0.62	0.01	0.03	0.08	0.04	0.12	0.02
COG0560	**K15781**	**putative_phosphoserine_phosphatase_/_1-acylglycerol-3-phosphate_O-acyltransferase_[EC:3.1.3.3_2.3.1.51]**	**0.26**	**0.23**	**0.75**	**0.12**	**0.27**	**0.09**	**0.27**	**0.14**	**0.46**
COG0584	**K00096**	**glycerol-1-phosphate_dehydrogenase_[NAD(P)+]_[EC:1.1.1.261]**	**0.23**	**0.19**	**0.05**	**0.07**	**0.05**	**0.02**	**0.24**	**0.25**	**0.31**
COG4302	K04019	ethanolamine_utilization_protein_EutA	0.21	0.05	0	0	0	0	0.03	0.06	0

COG1646	K13622	S-adenosylmethionine-diacylglycerol_3-amino-3-carboxypropyl_transferase	0.19	0.18	0.09	0.01	0.04	0	0	0	0.01
COG0578	**K00111**	**glycerol-3-phosphate_dehydrogenase_[EC:1.1.5.3]**	**0.17**	**0.07**	**0.37**	**0.06**	**0.2**	**0.01**	**0.06**	**0.02**	**0.25**
COG3675	K16818	phospholipase_A1_[EC:3.1.1.32]	0.14	0.12	0.08	0.03	0.01	0.01	0.01	0.02	0.01
COG3075	**K00112**	**glycerol-3-phosphate_dehydrogenase_subunit_B_[EC:1.1.5.3]**	**0.12**	**0.1**	**0.12**	**0.35**	**0.86**	**0.03**	**0.02**	**0.02**	**0**
COG4819	K08744	cardiolipin_synthase_(CMP-forming)_[EC:2.7.8.41]	0.1	0.15	0.21	0.42	0.66	0.31	0.21	0.27	0.39
COG1267	K03735	ethanolamine_ammonia-lyase_large_subunit_[EC:4.3.1.7]	0.06	0.02	0.12	0	0	0	0	0	0
COG3412	K01046	triacylglycerol_lipase_[EC:3.1.1.3]	0.05	0.11	0.05	0.06	0.06	0.03	0.04	0.06	0.06
COG2829	K01058	phospholipase_A1/A2_[EC:3.1.1.32_3.1.1.4]	0.04	0.04	0	0	0	0	0	0	0
COG0688	K01613	phosphatidylserine_decarboxylase_[EC:4.1.1.65]	0.03	0.04	0.18	0.05	0.13	0.01	0.14	0.31	0.07
COG2937	**K00631**	**glycerol-3-phosphate_O-acyltransferase_[EC:2.3.1.15]**	**0.01**	**0.02**	**0.02**	**0**	**0**	**0**	**0**	**0**	**0**
COG1075	K19005	lipoteichoic_acid_synthase_[EC:2.7.8.20]	0.01	0	0.23	0	0	0.01	0	0	0
COG4909	K06120	glycerol_dehydratase_large_subunit_[EC:4.2.1.30]	0	0.01	0	0	0	0	0.03	0.11	0.01
COG5153	K17900	lipase_ATG15_[EC:3.1.1.3]	0	0	0	0	0	0	0	0	0
COG5153	K17900	lipase_ATG15_[EC:3.1.1.3]	0	0	0	0	0	0	0	0	0
COG5379	K17104	phosphoglycerol_geranylgeranyltransferase_[EC:2.5.1.41]	0	0	0	0	0	0	0.01	0.01	0

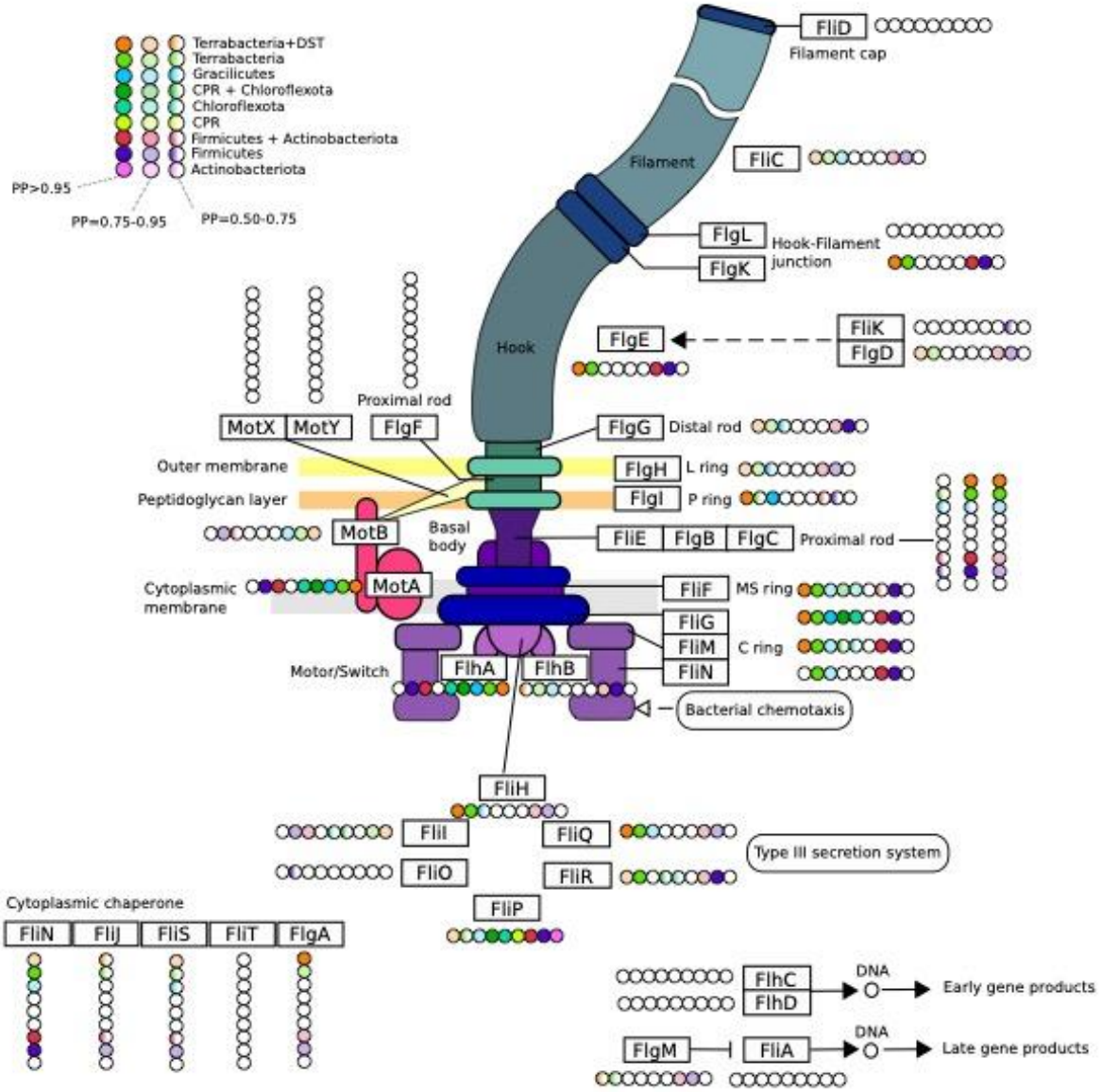
Table 4.1 *Posterior probabilities for the presence of glycerolipids in the ancestors of Terrabacteria and DTS (Terra+DTS), Terrabacteria (Terra), Gracilicutes (Graci), Chloroflexota and CPR (Chlo+CPR), Chloroflexota (Chlo), CPR, Firmicutes and Actinobacteriota (F+A), Firmicutes (F) and Actinobacteriota (A) respectively. Key genes whose Kegg IDs are given in the text are highlighted in bold. Annotations and PP values for all KOs can be found in Supplementary Table 1.*

Evolution of bacterial motility

The flagellum is very widespread across the tree of Bacteria (heatmap in Chapter 3, Fig. 3.2), and is thought to have evolved early in bacterial evolution (Liu and Ochman, 2007a, 2007b; Miyata *et al.*, 2020). Supporting this, we find evidence for the flagellum

in LBCA (Chapter 3). We recover many proteins involved in the flagellar construction in Terrabacteria and Gracilicutes ancestors respectively (heatmap in Chapter 3, Fig 3.2; Fig. 4.6), including components of the C, Ms and P rings, the proximal and distal rods and the filament, as well as the Type III secretion system (Fig. 4.6). We similarly infer a fully functioning flagellum in the ancestor of Firmicutes and Actinobacteriota, and the Firmicute ancestor. However, all flagellar genes are lost in the ancestor of Actinobacteriota. Most modern members of Actinobacteriota are non-flagellated, although some do possess flagella, most likely acquired via HGT, or analogous flagella-like structures (Barka *et al.*, 2016). We infer a few flagellar genes in the common ancestor of CPR and Chloroflexota, and the Chloroflexota ancestor, although it is unclear whether any of these would have had a fully functioning flagellum (Fig. 4.6). Modern members of Chloroflexota are not thought to be flagellated (Miyata *et al.*, 2020), although there is evidence to suggest that at least some members possess gram-positive type flagella (Mehrshad *et al.*, 2018), or archaeal flagella (known as archaella) via HGT from an archaeon (Hug *et al.*, 2013). It is unclear whether these taxa acquired flagellar genes via HGT or inherited them vertically. No flagellar genes are recovered in the ancestor of CPR, reflecting their loss in that clade (Castelle *et al.*, 2018). In addition to the flagellum, we find evidence for the presence of type IV pili across the deep nodes of the tree, although somewhat patchily distributed (heatmap in Chapter 3, Fig. 3.2). In the common ancestor of CPR and Chloroflexota and the ancestors of CPR and Chloroflexota respectively, we infer a small number of proteins for the construction of pili, but none for the Type IV pilus specifically (Supplementary Table 1).

Fig. 4.6 *Components of the flagellum inferred in the ancestors of Terrabacteria and DTS, Terrabacteria, Gracilicutes, Chloroflexota and CPR, Chloroflexota, CPR, Firmicutes and Actinobacteriota, Firmicutes, and Actinobacteriota respectively. The reconstruction is based on genes that could be mapped to a given node with PP >0.5. White indicates PP <0.5. The presence of a gene within a pathway is indicated as shown in the key. Annotations and PP values for all KOs can be found in Supplementary Table 1.*



Evolution of the outer membrane

Proteins which anchor the flagellum (i.e. FlgI, K02394; FlgH, K02393; and FlgA, K02386) and the type IV pilus (i.e. PilQ, K02666) into the outer membrane were recovered in the ancestors of Gracilicutes, Terrabacteria, the common ancestor of Firmicutes and Actinobacteriota and the ancestor of Firmicutes respectively, indicating the presence of an outer-membrane at these nodes (heatmap in Chapter 3, Fig. 3.2; Table 4.2). Congruent with this evidence, we recover a number of other components involved in the construction of an outer membrane in each of these nodes. We recover many proteins involved in lipopolysaccharide (LPS) synthesis in the ancestor of both Gracilicutes and Terrabacteria (heatmap in Chapter 3, Fig. 3.2; Table 4.2), although

support is stronger in the former. Indeed, LPS genes are found in all Gracilicutes within our sample, while the distribution is more patchy in Terrabacteria (heatmap in Chapter 3, Fig. 3.2). Within Terrabacteria, they are found in the Cyanobacteria, two classes of Firmicutes, Armatimonadota, and some in Eremiobacterota.

FAM	kegg_id	kegg_description	Terra+DST	Terra	Graci	Chlo+CPR	Chlo	CPR	F+A	F	A
COG0763	K00748	lipid-A-disaccharide_synthase_[EC:2.4.1.182]	0.99	0.81	0.01	0.05	0.03	0.01	0.28	0.39	0.04
COG0279	K03271	D-sedoheptulose_7-phosphate_isomerase_[EC:5.3.1.28]	0.98	0.83	1	0.32	0.43	0.19	0.15	0.06	0.11
COG0774	K02535	UDP-3-O-[3-hydroxymyristoyl]_N-acetylglucosamine_deacetylase_[EC:3.5.1.108]	0.98	0.99	0.98	0.06	0.01	0.02	0.91	0.95	0.04
COG1922	K05946	N-acetylglucosaminyldiphosphoundecaprenol_N-acetyl-beta-D-mannosaminyltransferase_[EC:2.4.1.187]	0.98	0.93	0.98	0.97	0.95	1	0.91	0.96	0.5
COG1137	K06861	lipopolysaccharide_export_system_ATP-binding_protein_[EC:3.6.3.-]	0.93	0.62	0.92	0.04	0	0.01	0.46	0.53	0.03
COG0241	K03273	D-glycero-D-manno-heptose_1,7-bisphosphate_phosphatase_[EC:3.1.3.82_3.1.3.83]	0.88	0.86	0.94	0.79	0.78	0.64	0.28	0.2	0.17
COG1043	K00677	UDP-N-acetylglucosamine_acyltransferase_[EC:2.3.1.129]	0.83	0.98	0.9	0.24	0.06	0.06	0.98	0.99	0.12
COG2605	K07031	D-glycero-alpha-D-manno-heptose-7-phosphate_kinase_[EC:2.7.1.168]	0.82	0.69	0.38	0.55	0.57	0.39	0.19	0.03	0.15
COG4370	K00748	lipid-A-disaccharide_synthase_[EC:2.4.1.182]	0.79	0.99	0.52	0.03	0	0.02	0.7	0.71	0.01
COG1682	K09690	lipopolysaccharide_transport_system_permease_protein	0.77	0.75	0.35	0.74	0.66	0.72	0.8	0.21	0.94
COG1044	K02536	UDP-3-O-[3-hydroxymyristoyl]_glucosamine_N-acyltransferase_[EC:2.3.1.191]	0.76	0.74	0.93	0.15	0.08	0.06	0.45	0.67	0.26
COG1682	K09690	lipopolysaccharide_transport_system_permease_protein	0.73	0.76	0.37	0.79	0.71	0.75	0.79	0.2	0.92
COG1295	K07058	membrane_protein	0.68	0.71	0.06	0.93	0.97	0.12	0.67	0.27	0.72
COG0615	K03272	D-beta-D-heptose_7-phosphate_kinase_/_D-beta-D-heptose_1-phosphate_adenosyltransferase_[EC:2.7.1.167_2.7.7.70]	0.62	0.33	0.63	0.19	0.22	0.07	0.11	0.1	0.07
COG1134	K09691	lipopolysaccharide_transport_system_ATP-binding_protein	0.58	0.6	0.2	0.68	0.44	0.77	0.67	0.11	0.84
COG0795	K11720	lipopolysaccharide_export_system_permease_protein	0.57	0.65	0.77	0.1	0.06	0.07	0.19	0.33	0.09
COG0794	K06041	arabinose-5-phosphate_isomerase_[EC:5.3.1.13]	0.57	0.36	0.83	0.04	0.02	0.05	0.4	0.59	0.24

COG1134	K09691	lipopolysaccharide_transport_system_ATP-binding_protein	0.54	0.6	0.12	0.56	0.38	0.69	0.7	0.18	0.83
COG0615	K03272	D-beta-D-heptose_7-phosphate_kinase_/_D-beta-D-heptose_1-phosphate_adenosyltransferase_[EC:2.7.1.167_2.7.7.70]	0.44	0.46	0.65	0.28	0.24	0.28	0.13	0.11	0.05
COG2076	K12962	undecaprenyl_phosphate-alpha-L-ara4N_flippase_subunit_ArnE	0.36	0.33	0.1	0.15	0.04	0.22	0.13	0.12	0.12
COG2870	K03272	D-beta-D-heptose_7-phosphate_kinase_/_D-beta-D-heptose_1-phosphate_adenosyltransferase_[EC:2.7.1.167_2.7.7.70]	0.33	0.65	0.76	0.29	0.24	0.36	0.05	0.08	0.02
COG1560	K02517	Kdo2-lipid_IVA_lauroyltransferase/acyltransferase_[EC:2.3.1.241_2.3.1.-]	0.31	0.24	0.95	0.33	0.95	0.09	0.21	0.1	0.34
COG1519	K02527	3-deoxy-D-manno-octulosonic-acid_transferase_[EC:2.4.99.12_2.4.99.13_2.4.99.14_2.4.99.15]	0.26	0.28	0.43	0	0.02	0	0.43	0.58	0.02
COG1212	K00979	3-deoxy-manno-octulosonate_cytidylyltransferase_(CMP-KDO_synthetase)_[EC:2.7.7.38]	0.19	0.18	0.68	0.03	0.03	0.02	0.17	0.28	0.03
COG2121	K09778	Kdo2-lipid_IVA_3'_secondary_acyltransferase_[EC:2.3.1.-]	0.18	0.16	0.83	0.06	0.08	0.03	0.06	0.15	0
COG1663	K00912	tetraacyldisaccharide_4'-kinase_[EC:2.7.1.130]	0.18	0.05	0.81	0.03	0.03	0	0	0.07	0.02
COG2877	K01627	2-dehydro-3-deoxyphosphooctonate_aldolase_(KDO_8-P_synthase)_[EC:2.5.1.55]	0.17	0.08	0.97	0.01	0.03	0	0.03	0.12	0.01
COG3765	K05789	chain_length_determinant_protein_(polysaccharide_antigen_chain_regulator)	0.17	0.1	0.01	0	0	0	0.01	0	0.01
COG1778	K03270	3-deoxy-D-manno-octulosonate_8-phosphate_phosphatase_(KDO_8-P_phosphatase)_[EC:3.1.3.45]	0.05	0.07	0.12	0.02	0.02	0.02	0.06	0.08	0.02
COG4261	K02517	Kdo2-lipid_IVA_lauroyltransferase/acyltransferase_[EC:2.3.1.241_2.3.1.-]	0.05	0.03	0.16	0	0.01	0.01	0.01	0.01	0.01
COG3117	K11719	lipopolysaccharide_export_system_protein_LptC	0.03	0.01	0	0.01	0	0	0.01	0.05	0
COG3642	K11211	3-deoxy-D-manno-octulosonic_acid_kinase_[EC:2.7.1.166]	0.03	0.04	0.04	0.01	0.02	0.02	0.03	0.02	0.03
COG2908	K03269	UDP-2,3-diacylglucosamine_hydrolase_[EC:3.6.1.54]	0.03	0.01	0.03	0.01	0.01	0.01	0.01	0.01	0.01
COG1442	-	-	0.02	0.01	0	0	0	0.01	0.02	0.02	0
COG5416	K08992	lipopolysaccharide_assembly_protein_A	0.01	0.01	0	0.05	0.14	0	0.13	0.68	0.02
COG3528	K09953	lipid_A_3-O-deacylase_[EC:3.1.1.-]	0.01	0.01	0	0	0	0	0.01	0	0.01

COG2956	K19804	lipopolysaccharide_assembly_protein_B	0	0	0.48	0	0	0.01	0	0.01	0.01
COG3475	K07271	lipopolysaccharide_cholinephosphotransferase_[EC:2.7.8.-]	0	0	0	0	0	0	0	0	0
COG1934	K09774	lipopolysaccharide_export_system_protein_LptA	0	0.01	0.07	0.02	0.02	0.02	0.15	0.46	0.03
COG5375	K11719	lipopolysaccharide_export_system_protein_LptC	0	0.02	0.01	0.01	0.01	0	0.02	0.03	0.02

Table 4.2 *Posterior probabilities for the presence of genes involved in lipopolysaccharide biosynthesis in the ancestors of Terrabacteria and DTS (Terra+DTS), Terrabacteria (Terra), Gracilicutes (Graci), Chloroflexota and CPR (Chlo+CPR), Chloroflexota (Chlo), CPR, Firmicutes and Actinobacteriota (F+A), Firmicutes (F) and Actinobacteriota (A) respectively. Annotations and PP values for all KOs can be found in Supplementary Table 1.*

Proteins for LPS synthesis are recovered in the common ancestor of Firmicutes and Actinobacteriota, as well as in the ancestor of Firmicutes. This is congruent with recent research demonstrating that Firmicutes were ancestrally diderms, with multiple subsequent lineage specific losses (Antunes *et al.*, 2016; Megrian *et al.*, 2020). Proteins for lipopolysaccharide are absent in the ancestor Actinobacteriota, reflecting the loss of the outer membrane in that clade (Barka *et al.*, 2016). Subunits of the key enzymes, Lpx (LpxA, K00677; LpxC, K02535; and LpxD, K02536) are recovered with moderate to high supports in the all nodes (excluding Actinobacteriota, and CPR+Chloroflexota and daughter nodes), with Kds (KdsD, K06041; KdsA, K01627; and KdsC, K00979) being more patchily distributed (Table 4.2). The outer membrane transport protein OmpH (K06142) is also recovered with moderate to high support (>0.70) in these nodes. Proteins for LPS synthesis and other outer membrane components are not found in any Actinobacteriota, Chloroflexota or CPR in our dataset (heatmap in Chapter 3, Fig. 3.2), and therefore are not recovered in the respective ancestor nodes.

All the evidence taken together, ancestors of Gracilicutes and Terrabacteria respectively were likely diderms, a characteristic inherited from LBCA. This is contrary to hypotheses which argue for a monoderm first scenario (Lake, 2009; Gupta, 2011; Tocheva *et al.*, 2011). Within Terrabacteria, the outer membrane was lost

independently in various lineages (Megrian *et al.*, 2020), including the Actinobacteriota, several classes of Firmicutes (Antunes *et al.*, 2016), and the branch leading to CPR and Chloroflexota.

Terrabacteria and terrestrialisation

In Chapter 2, we recover support for Terrabacteria, a clade first described by Battistuzzi et al. (2004). They name the group due to its inclusion of many phyla that are partially or fully terrestrial, and hypothesise that the emergence of Terrabacteria may represent an early colonisation of the terrestrial environment. They further expand on this, suggesting the widespread presence of genes related to sporulation and presence of a monderm cell envelope with a thick peptidoglycan layer confer resistance to various environmental stresses related to terrestrial habitats, such as desiccation, ultraviolet radiation, and high salt concentration (Battistuzzi and Hedges, 2009). We recover a small number of sporulation genes in the ancestors of both Terrabacteria and Gracilicutes (called “Hydrobacteria” by Battistuzzi and Hedges), yet it is difficult to assess whether either would have had the ability to form spores, although it seems unlikely given their sparse distribution. A number of sporulation genes are found in the ancestor of Firmicutes, which contains many spore forming members. Additionally, as already explained above, and in Chapter 3, we find evidence for the presence of a diderm cell envelope architecture in the ancestor of Terrabacteria and its most basally branching clades. It is therefore difficult to assess the validity of Battistuzzi and Hedges’ hypothesis, but given our results, it seems unlikely.

Terrestrialisation has also been suggested to have evolved via expanding genome sizes, mediated by acquiring the *dnaE2* gene in terrestrial lineages (Wu *et al.*, 2014). Wu et al. claim that increased genomes size would facilitate greater adaptive ability to terrestrial environments, which are more heterogeneous than marine environments. They demonstrate that within clades, terrestrial species have larger genomes than marine species, with genomes decreasing in clades which re-enter the marine realm. In our analyses, we do not recover *dnaE2* in any of the nodes surveyed. We do see larger genome sizes within the ancestral nodes of Terrabacteria compared to Gracilicutes, and larger genomes sizes in the ancestors of predominantly terrestrial phyla, such as Firmicutes, when compared to predominantly marine phyla, such as

Cyanobacteria (Fig. 4.3). However, the differences are not dramatic, and further study with better taxon sampling would be needed to investigate this further.

4.4 Conclusion

Patterns of gene transfer show that the Terrabacteria radiated earlier than the Gracilicutes, with all clades within Terrabacteria predating the entire Gracilicute radiation except for Cyanobacteria and Actinobacteriota. CPR seems to represent an early radiation, demonstrating that, though not in a basal position in the tree, they have formed a part of bacterial communities for most of evolutionary history. In contrast, Cyanobacteria and Alphaproteobacteria both emerged relatively late during bacterial evolution, which further implies the emergence of eukaryotic cells at a later stage of the diversification of life (Parfrey *et al.*, 2011; Eme *et al.*, 2014b; Knoll, 2014; Nicholas J. Butterfield, 2015; Betts *et al.*, 2018), rather than being contemporaneous or pre-dating the emergence of most bacterial lineages (Kurland, Collins and Penny, 2006).

We recover evidence for central carbohydrate metabolic pathways across the deepest nodes in the bacterial tree, with glycolysis being strongly supported in all cases, and other pathways, such as the pentose phosphate pathway and the TCA cycle being more moderate in their support and inconsistent in their distribution. With regards to carbon fixation, we do not find evidence of the Calvin Cycle, or any of the three variants of the 3-hydroxypropionate bicycle in any nodes. Although we identify some components of the TCA cycle in various nodes, the key enzyme of the reverse TCA cycle, ATP citrate lyase, is not recovered in any of them. Given that the directionality of the enzymes is difficult to assess (Nunoura *et al.*, 2018), we cannot conclusively say whether any of the nodes assessed possessed the reverse TCA cycle, although we cannot rule out the possibility. The presence of components of the methyl-branch of the WLP, as well as components of a putative Rnf complex, together may indicate that, as in LBCA, both the ancestor of Gracilicutes and Terrabacteria were capable of acetogenic growth (Schuchmann and Müller, 2014). This pathway seems to have been subsequently lost in Actinobacteriota and in the ancestor of Chloroflexota and CPR, as although these lineages maintain components of the methyl branch, they lack the Rnf

complex. However, we fail to recover any components of the key enzyme of the pathway, CODH/ACS (Adam, Borrel and Gribaldo, 2018) in any of the nodes, except for a single subunit in the ancestor of Gracilicutes, making it difficult to conclusively establish presence of the WLP or acetogenesis. As in LBCA, it has proven difficult to conclusively establish the central energy metabolism of the earliest Bacteria.

We find evidence of proteins for G3P phospholipids in all the surveyed nodes, except for CPR, although the components of the pathway are patchily distributed. We also find no evidence of the presence of archaeal G1P lipids. Early bacterial lineages therefore likely possessed G3P lipids, as in modern Bacteria, with the ability to produce membrane phospholipids being lost in the CPR (Castelle *et al.*, 2018). We further identified most of the proteins required to synthesise motile appendages such as flagella and pili in the ancestors of Gracilicutes and Terrabacteria. This implies that the earliest lineages of Bacteria were motile, and likely lived in environments in which dispersal, chemotaxis and surface attachment would have been advantageous. The flagellum was lost in Actinobacteriota and the common ancestor of Chloroflexota and CPR, modern members of which do not typically have flagella (Barka *et al.*, 2016; Castelle *et al.*, 2018; Miyata *et al.*, 2020).

We recover proteins involved in the construction of the outer cell membrane, including for LPS biosynthesis, in the ancestors of Gracilicutes and Terrabacteria, from which we infer that both possessed a double membrane with an LPS layer. Consistent with this inference, we infer the presence of the flagellar subunits FlgH, FlgI and FgA in, which anchor flagella in diderm membranes (Antunes *et al.*, 2016), and for the Type IV pilus subunit PilQ, which among extant bacteria is specific to diderms (Antunes *et al.*, 2016; Megrian *et al.*, 2020). We additionally recover these proteins in the common ancestor of Firmicutes and Actinobacteriota, with the loss of these pathways in Actinobacteriota and the common ancestor of Chloroflexota and CPR. Taken together, these results are consistent with hypotheses (Cavalier-Smith, 2006) in which the earliest bacterial lineages were diderms (Antunes *et al.*, 2016; Megrian *et al.*, 2020), and argue against Monoderm-first scenarios (Lake 2009; Gupta 2011; Tocheva, Ortega and Jensen 2016). Subsequent diderm-to-monoderm transitions may have occurred on multiple occasions within Bacteria (Antunes *et al.*, 2016; Megrian *et al.*, 2020). Furthermore, the presence of flagellar in the deepest nodes further supports

hypotheses that the earliest Bacteria were motile organism (Liu and Ochman, 2007a, 2007b), and not non-motile cells living on mineral substrates (Martin and Russell, 2007; Lane and Martin, 2012; Sousa *et al.*, 2013; Sousa and Martin, 2014), with the evolution of alternative means of motility (Miyata *et al.*, 2020) or loss of motility in later lineages.

We do not recover evidence of terrestrial adaptation in the ancestral nodes within Terrabacteria, as had been previously suggested (Battistuzzi, Feijao and Hedges, 2004; Battistuzzi and Hedges, 2009). However, the hypothesised relationship between genomes size and terrestrialisation (Wu *et al.*, 2014) is less clear, and needs additional exploration. It seems unlikely that terrestrialisation happened deep within the terrabacterial tree, but rather occurred independently in various lineages, as is likely the case in Gracilicutes.

More extensive analyses would need to be done on individual gene families to better answer some of the outstanding questions, especially relating to carbon fixation and the evolution of terminal oxidases. However, the early evolution of Bacteria was seemly dominated by autotrophic, tentatively anaerobic acetogenic, motile diderm cells on the one hand, and highly reduced, metabolically minimalist cells on the other, before the later rise of photosynthetic and aerobic Bacteria, and the eventually appearance of eukaryotic cells.

Chapter 5

Investigating the Origins of Membrane Phospholipid Biosynthesis Genes Using Outgroup-Free Rooting

This chapter has been published as a Coleman *et al.* (2019) in Genome Biology and Evolution in collaboration with Richard D. Pancost and Tom A. Williams. Gareth A. Coleman is the first author of this paper. The project was conceived by GAC, RDP and TAW. GAC and TAW designed and performed the analyses. GAC, RDP and TAW interpreted the results and wrote the manuscript.

Paper published as:
Coleman, G.A., Pancost, R.D. and Williams, T.A., 2019. Investigating the origins of membrane phospholipid biosynthesis genes using outgroup-free rooting. *Genome biology and evolution*, *11*(3), pp.883-898.

Paper can be found here:
https://academic.oup.com/gbe/article/11/3/883/5310093

Abstract

One of the key differences between Bacteria and Archaea is their canonical membrane phospholipids, which are synthesised by distinct biosynthetic pathways with nonhomologous enzymes. This "lipid divide" has important implications for the early evolution of cells and the type of membrane phospholipids present in the last universal common ancestor. One of the main challenges in studies of membrane evolution is that the key biosynthetic genes are ancient and their evolutionary histories are poorly resolved. This poses major challenges for traditional rooting methods because the only available outgroups are distantly related. Here, we address this issue by using the best available substitution models for single-gene trees, by expanding our analyses to the diversity of uncultivated prokaryotes recently revealed by environmental genomics, and by using two complementary approaches to rooting that do not depend on outgroups. Consistent with some previous analyses, our rooted gene trees support extensive interdomain horizontal transfer of membrane phospholipid biosynthetic genes, primarily from Archaea to Bacteria. They also suggest that the capacity to make archaeal-type membrane phospholipids was already present in the last universal common ancestor.

5.1 Introduction

Archaea and Bacteria form the two primary domains of life (Williams *et al.*, 2013). Although similarities in their fundamental genetics and biochemistry, and evidence of homology in a near-universally conserved core of genes (Weiss *et al.*, 2016) strongly suggest that Archaea and Bacteria descend from a universal common ancestor (LUCA), they also differ in ways that have important implications for the early evolution of cellular life. These differences include DNA replication (Kelman and Kelman, 2014), transcription (Bell and Jackson, 1998), DNA packaging (Reeve, Sandman and Daniels, 1997), and cell wall compositions (Kandler, 1995). One striking difference is in the phospholipid composition of the cell membranes (Fig. 5.1), which is particularly important for understanding the origin of cellular life. Canonically, Archaea have isoprenoid chains attached to a glycerol-1-phosphate (G1P) backbone via ether bonds and can have either membrane spanning or bilayer-forming phospholipids (Lombard, López-García and Moreira, 2012b). Most Bacteria, as well as eukaryotes, classically have acyl (fatty-acid) chains attached to a glycerol-3-phosphate (G3P) backbone via ester bonds and form bilayers (Lombard, López-García and Moreira, 2012b), although a number of exceptions have been documented (Sinninghe Damsté *et al.*, 2002; Weijers *et al.*, 2006; Damsté, Sinninghe Damsté, *et al.*, 2007; Goldfine, 2010). Archaeal and bacterial phospholipids are synthesised by non-homologous enzymes via different biosynthetic pathways (Fig. 5.1). This so-called “lipid divide” (Koga, 2011) raises some important questions regarding the early evolution of cellular life, including the nature of the membrane phospholipids present in LUCA and the number of times cell membranes have evolved.

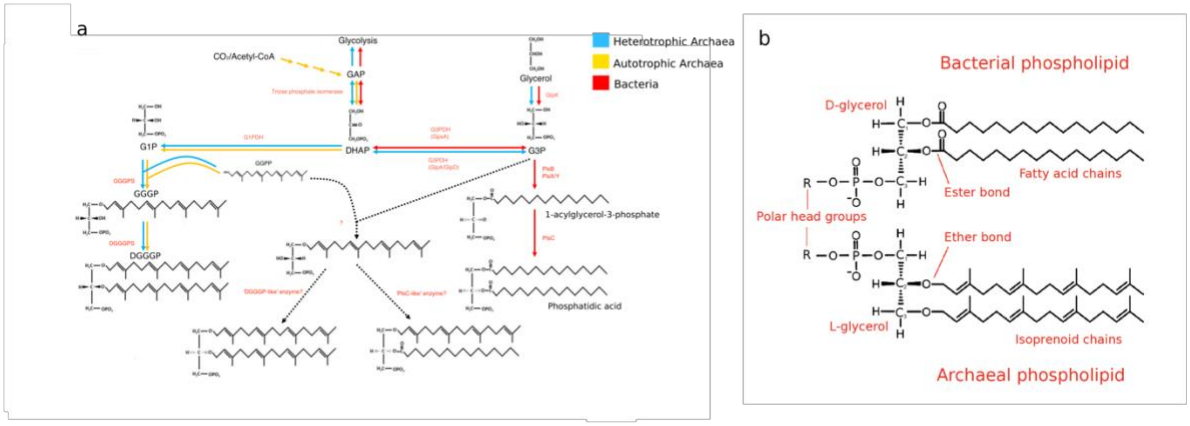


Fig. 5.1 (above) *(a) The canonical ether/ester biosynthetic pathways in Archaea and Bacteria and how they relate to glycerol metabolism. Based on Figure 1 from Villanueva et al. (2017). Archaeal pathways in blue and yellow (blue = heterotrophic Archaea and yellow = autotrophic Archaea), bacterial pathway in red. Hypothetical biosynthetic pathway, as suggested by Villanueva et al. (2017), in dashed lines. (b) Composition of bacterial and archaeal phospholipids. In Archaea, glycerol-1-phosphate (G1P) is synthesised from dihydroxyacetone phosphate (DHAP) using the enzyme glycerol-1-phosphate dehydrogenase (G1PDH). The first and second isoprenoid chains (GGGPs) are added by geranylgeranylglyceryl phosphate synthase (GGGPS) and digeranylgeranylglyceryl phosphate synthase (DGGGPS), respectively. In Bacteria, glycerol-3-phosphate (G3P) is synthesised by glycerol-3-phosphate dehydrogenase (G3PDH) from DHAP. There are two forms of this enzyme, GpsA and GlpA/GlpD, encoded by the gps and glp genes, respectively. G3P may also be produced from glycerol by glycerol kinase (GlpK). In certain Bacteria, such as Gammaproteobacteria, the first fatty-acid chain is added by a version of glycerol-3-phosphate acyltransferase called PlsB. Other Bacteria, including most gram-positive bacteria, use a system which includes another glycerol-3-phosphate acyltransferase, PlsY, in conjunction with the enzyme PlsX (Parsons and Rock, 2013; Yao and Rock, 2013). The second fatty-acid chain is attached by 1-acylglycerol-3-phosphate O-acyltransferase (PlsC).*

The observation that phospholipid biosynthesis in Bacteria and Archaea is non-homologous has motivated various hypotheses on the nature of LUCA's membrane. The likely presence of some genes for lipid biosynthesis (Lombard and Moreira, 2011; Lombard, López-García and Moreira, 2012b; Koga, 2014; Weiss *et al.*, 2016) and, in particular, a membrane-bound ATPase (Sojo, Pomiankowski and Lane, 2014; Weiss *et al.*, 2016) in reconstructions of LUCA's genome implies that LUCA possessed a membrane, although its properties may have been somewhat different to those of modern, ion-tight prokaryote cell membranes (Lombard, López-García and Moreira, 2012b; Koga, 2014; Sojo, Pomiankowski and Lane, 2014). It has also been suggested that LUCA may have had a heterochiral membrane (Wächtershäuser, 2003), with later independent transitions to homochirality in Bacteria and Archaea, driven by increased membrane stability. However, the available experimental evidence, including the

recent engineering of an *Escherichia coli* cell with a heterochiral membrane (Caforio *et al.*, 2018), suggests that homochiral membranes are not necessarily more stable than heterochiral ones (Fan *et al.*, 1995; Shimada and Yamagishi, 2011; Caforio *et al.*, 2018), requiring some other explanation for the loss of ancestral heterochirality.

Despite the importance of the lipid divide for our understanding of early cellular evolution, membrane phospholipid stereochemistry of the glycerol moiety has been directly determined for a surprisingly limited range of Bacteria and Archaea. Since the initial full structural characterization of archaeol by (Kates, 1978), most subsequent studies of ether membrane lipids have assumed their stereochemistry while focusing on other aspects of their structure. Those studies that have determined the glycerol stereochemistry of membrane lipids (i.e., (Sinninghe Damsté *et al.*, 2002; Weijers *et al.*, 2006) are largely consistent with the idea that it is a conserved difference between Bacteria and Archaea. Nonetheless, there is evidence that some Bacteria can make G1P-linked ether lipids. For example, the model bacterium *Bacillus subtilis* has been shown to possess homologs of archaeal glycerol-1-phosphate dehydrogenase (G1PDH) and geranylgeranylglyceryl phosphate synthase (GGGPS) (Guldan, Sterner and Babinger, 2008; Guldan *et al.*, 2011). These enzymes allow *B. subtilis* to synthesise a typically archaeal ether link between G1P and HepPP, resulting in a lipid with archaeal characteristics, although there is no evidence that these archaeal-like lipids are used to make phospholipids or are incorporated into the *B. subtilis* membrane.

Apart from stereochemistry, other characteristics of membrane phospholipids appear to be more variable, showing a mixture of archaeal and bacterial features. For example, the plasmalogens of animals and anaerobic Bacteria include an ether bond (Goldfine, 2010). Branched glycerol dialkyl glycerol tetra-ether lipids found in the environment have bacterial stereochemistry and branched rather than isoprenoidal alkyl chains, but they also contain ether bonds and span the membrane, as observed for canonical archaeal lipids (Schouten *et al.*, 2000; Weijers *et al.*, 2006). These branched glycerol dialkyl glycerol tetra-ethers are particularly abundant in peat bogs and were thought to be produced by Bacteria as adaptations to low pH environments (Weijers *et al.*, 2006; Damsté, Sinninghe Damsté, *et al.*, 2007), but are now known to occur in a wide range of soils and aquatic settings (Schouten, Hopmans and Sinninghe Damsté, 2013). The

enzymes responsible for their synthesis are currently unknown. On the other side of the "lipid divide," some Archaea have been shown to produce membrane lipids with fatty-acid chains and ester bonds (Gattinger, Schloter and Munch, 2002). The biosynthetic pathways for all of these mixed-type membrane lipids remain unclear. However, given the frequency with which prokaryotes undergo horizontal gene transfer (Garcia-Vallvé, Romeu and Palau, 2000), one possibility is that these mixed biochemical properties reflect biosynthetic pathways of mixed bacterial and archaeal origin.

A number of previous studies have investigated the evolutionary origins of phospholipid biosynthesis genes in Bacteria and Archaea using phylogenetic approaches, in order to test hypotheses about the nature of membranes in the earliest cellular life-forms (Peretó, López-García and Moreira, 2004; Koga, 2014; Yokobori *et al.*, 2016; Villanueva, Schouten and Damsté, 2017). In this study, we build upon that work by performing comprehensive phylogenetic analyses for the core phospholipid biosynthesis genes in Bacteria and Archaea: the enzymes that establish membrane lipid stereochemistry and attach the two carbon chains to the glycerol phosphate backbone (Fig. 5.1), as the histories of these enzymes are key to understanding the evolution of membrane biosynthesis and stereochemistry. Our analyses take advantage of the wealth of new genome data from environmental prokaryotes that has become available recently, and we employ new approaches for rooting single-gene trees in order to circumvent some of the difficulties inherent in traditional outgroup rooting for anciently diverged genes. Our results agree with previous work in suggesting that LUCA likely possessed a cell membrane. Our rooted gene trees indicate that transfers of lipid biosynthetic genes from Archaea to Bacteria have occurred more frequently in evolution, particularly during the early diversification of the two domains.

5.2 Materials and Methods

Sequence Selection

For Archaea, we selected 43 archaeal genomes, sampled evenly across the archaeal tree. We took corresponding archaeal G1PDH, geranylgeranylglyceryl phosphate

synthase (GGGPS), and digeranylgeranylglyceryl phosphate synthase (DGGGPS) amino acid sequences from the data set of Villanueva et al. (2017) and performed BlastP searches to find these sequences in genomes not included in that data set. For Bacteria, we selected 64 bacterial genomes, sampled so as to represent the known genomic diversity of bacterial phyla (Hug *et al.*, 2016). We used GpsA, GlpA/GlpD, and GlpK sequences from Yokobori et al. (2016) and performed BlastP searches to find those sequences in bacterial species not in their data set. For PlsC and PlsY, we took the corresponding sequences from Villanueva et al. (2017) and performed BlastP searches to find these sequences in the remaining genomes. For PlsB and PlsX, we searched for the respective terms in the gene database on the NCBI website, and upon finding well-verified occurrences, performed BlastP searches to find the corresponding amino acid sequences in the remaining genomes. We then used BLASTp to look for bacterial orthologues of the archaeal enzymes and vice versa. We selected sequences that had an E-value of less the 10e-7 and at least 50% coverage. Accession numbers for sequences used are provided in Supplementary Table 5.

Phylogenetics

The sequences were aligned in mafft (Katoh *et al.*, 2002) using the –auto option and trimmed in BMGE (Criscuolo and Gribaldo, 2010) using the BLOSUM30 model, which is most suitable for anciently diverged genes. To construct gene trees from our amino acid sequences, we first selected the best-fitting substitution model for each gene according to its Bayesian Information Criterion score using the model selection tool in IQ-Tree (Nguyen *et al.*, 2015). For all the genes we analyzed, the best-fitting model was a mixture model combining the Le and Gascuel (LG) exchangeability matrix (Le and Gascuel 2008) with site-specific composition profiles (the C40, C50, and C60 models (Lartillot and Philippe, 2004; Le and Gascuel, 2008)) to accommodate across-site variation in the substitution process. LG + C60 was used for G1PDH, DGGGP, GpsA, GlpA/GlpD, GlpK, and PlsC. LG + 50 was used for PlsY. LG + C40 was used for GGGPS. A discretised Gamma distribution (Yang, 1994) with four rate categories was used to model across-site rate variation. The trees were inferred with their respective models in PhyloBayes (Lartillot and Philippe, 2004; Lartillot, Brinkmann and Philippe, 2007b); convergence was assessed using the bpcomp and tracecomp programs (maxdiff < 0.1; effective sample sizes > 100), as recommended by the authors. We additionally inferred maximum likelihood (ML) trees in IQ-Tree using the

LG + C60 model for each enzyme for comparison. We used heads-or-tails (Landan and Graur, 2007) to assess the impact of alignment uncertainty: starting with the reversed alignments, we used the same phylogenetics pipeline as described above. Further testing was carried out by removing the metagenomic data from G1PDH, GGGPS, DGGGPS, GpsA, GlpA/GlpD, and GlpK, creating new alignments as described above, and inferring trees from these alignments in IQ-Tree using the LG + C60 model. We did not remove metagenomic data for PlsC or PlsY, as all of the archaeal sequences for these trees are derived from metagenome bins. In some cases, our trees included highly divergent sequences (sometimes forming distinct clades); we checked the E-values for these hits, and if they were close to or at the 10e-7 cut-off, they were removed and the analyses were rerun.

The trees were rooted with an uncorrelated lognormal relaxed molecular clock (RMC), using the LG model with a discretised Gamma distribution (Yang, 1994) with four rate categories, and a Yule tree prior (Drummond and Rambaut, 2007; Drummond *et al.*, 2012). We also rooted the trees using minimal ancestor deviation (MAD) rooting (Tria, Landan and Dagan, 2017). We used two complementary methods: root posterior probabilities averaged over the trees sampled during the Bayesian molecular clock analysis using RootAnnotator (Calvignac-Spencer *et al.*, 2014), and the ambiguity index (AI) implemented in MAD. The AI is defined as the ratio of the MAD value to the second smallest value. “Ties”, that is, where two or more competing root positions with equal deviations, would obtain a score of 1, with smaller values obtained in proportion to the relative quality of the best root position.

For G1PDH, GpsA, and GlpA/GlpD, we also rooted using a subsample of the outgroup sequences used by Yokobori et al. (2016). The outgroups used were two sequences annotated as 3-dehydroquinate synthase, five as glycerol dehydrogenase, and five as alcohol dehydrogenase for G1PDH; six sequences annotated as hydroxyacyl-CoA dehydrogenase and six as uridine diphosphoglucose 6-dehydrogenase sequences for GpsA; and 12 sequences annotated as flavin adenine dinucleotide dependent oxidoreductase for GlpA/GlpD. All three of these trees were inferred under the LG + C60 model to directly compare to the unrooted trees. Trees were also inferred from best-fit models selected in IQTree (LG + C60 for G1PDH and GlpA/GlpD and LG + C50 for GpsA).

Eukaryotic orthologues of prokaryotic phospholipid biosynthesis genes (GlpA/GlpD, GpsA, and PlsC) were identified by performing BlastP searches on 35 eukaryotic genomes from across eukaryotic diversity using *Homo sapiens* query as the sequence in each case, selecting sequences with an *E*-value of 10e-7 or less, and at least 50% coverage. We then performed model testing in IQTree and inferred trees in PhyloBayes using the selected substitution model (LG + C60 for PlsC and LG + C50 for GlpA/GlpD and GpsA).

5.3 Results and Discussion

Distribution of Core Phospholipid Biosynthesis Genes

We performed BlastP searches for the enzymes of the canonical archaeal and bacterial lipid biosynthesis pathways (Fig. 5.1) against all archaeal and bacterial genomes in the NCBI nr database. Our BLAST searches revealed homologs for all of the core phospholipid biosynthesis genes of both pathways in both prokaryotic domains, with the exception of bacterial enzymes PlsB and PlsX, which we did not find in Archaea. Orthologues of the canonical archaeal genes are particularly widespread in many bacterial lineages (Table 5.1). Of the 52 bacterial phyla surveyed, 8 had no orthologues of the archaeal genes (Table 5.1, indicated in red). Six phyla have orthologues of all three archaeal genes distributed across various genomes (Table 5.1, indicated in yellow and green). Of these phyla, Firmicutes (genera *Bacillus* and *Halanaerobium*), Actinobacteria (genus *Streptomyces*), and Fibrobacteres (genera *Chitinispirillum* and *Chitinivibrio*) contain species which have all three genes in their genomes (Table 5.1, indicated in green). Based on the presence of all three core biosynthetic genes, and given their recognised role in the synthesis of archaeal-like lipid components in *B. subtilis* (Guldan, Sterner and Babinger, 2008; Guldan *et al.*, 2011), members of Firmicutes, Actinobacteria, and Fibrobacteres lineages of Bacteria may be capable of making archaea like lipids, although we cannot determine if these are used in the production of membrane phospholipids. Of the 12 FCB group (Fibrobacteres, Chlorobi, Bacteroidetes and related lineages) phyla we surveyed, all 12 have GGGPS and DGGGPS orthologues, but only Fibrobacteres and Cloacimonetes have G1PDH orthologues (see Fig. 5.1 for overview of pathway). In these species lacking G1PDH, it is unclear whether GGGPS and DGGGPS are active

and if so, what they are used for; one possibility is that they catalyze the reverse reaction, catabolising archaeal lipids as an energy source. However, a very recent report (Villanueva, von Meijenfeldt and Westbye, 2018) has shown that the GGGPS and DGGGPS genes from one FCB lineage, Cloacimonetes, support the production of archaeal-type membrane phospholipids and a mixed membrane when heterologously expressed in *E. coli*. This suggests that both *E. coli* and perhaps Cloacimonetes have an alternative, as yet unknown mechanism for making G1P, and that some FCB members may have mixed archaeal and bacterial membranes.

Superphylum	Phylum	Class	G1PDH	GGGPS	DGGGPS	GpsA	GlpA/GlpD	GlpK	PlsC	PlsY
Archaea	Euryarchaeota	Archaeoglobi	✓	✓	✓	✓	✓	✓		
		Halobacteria	✓	✓	✓		✓	✓		
		Methanobacteria	✓	✓	✓	✓				
		Methanococci	✓	✓	✓					
		Methanomicrobia	✓	✓	✓	✓	✓			
		Thermococci	✓	✓	✓		✓	✓		
		Thermoplasmatales	✓	✓	✓	✓	✓	✓	✓	
TACK	Aigarchaeota		✓	✓	✓					
	Crenarchaeota		✓	✓	✓	✓	✓	✓		
	Korarchaeota		✓	✓	✓		✓	✓		
	Thaumarchaeta		✓	✓	✓					
Asgard	Heimdallarchaeota		✓	✓	✓				✓	✓
	Lokiarchaeota		✓	✓	✓		✓	✓	✓	✓
	Odinarchaeota		✓	✓	✓					

	Thorarchaeota	✓	✓	✓					✓
DPANN	Aenigmarchaeota	✓	✓	✓					
	Diapherotrites		✓	✓	✓				✓
	Micrarchaeota	✓	✓	✓					
	Nanoarchaeota								
	Nanohaloarchaeota								
	Pacearchaeota								✓
	Parvarchaeota					✓		✓	
	Woesearchaeota				✓		✓	✓	✓
Bacteria	Acidobacteria	✓			✓		✓	✓	✓
	Actinobacteria	✓	✓	✓	✓	✓	✓	✓	✓
	Aminicenantes	✓			✓	✓	✓		✓
	Aquificae				✓	✓	✓	✓	✓
	Armatimonadetes	✓			✓	✓	✓	✓	✓
	Candidate division KSB1		✓	✓	✓	✓	✓	✓	✓
	Candidate division NC10				✓	✓	✓	✓	✓
	Candidate division TA06	✓	✓	✓	✓			✓	✓
	Candidate division WOR-3	✓	✓	✓	✓			✓	✓
	Candidatus Edwardsbacteria		✓	✓	✓			✓	✓
	Candidatus Handelsmanbacteria		✓	✓		✓	✓	✓	✓

Candidatus Kerfeldbacteria	✓	✓		✓		✓	✓	✓
Candidatus Magnetoovum	✓	✓		✓			✓	✓
Candidatus Raymondbacteria	✓	✓	✓	✓	✓	✓	✓	✓
Chloroflexi	✓	✓	✓	✓	✓	✓	✓	✓
Chrysiogenetes	✓	✓		✓			✓	✓
Cloacimonetes	✓	✓	✓	✓	✓	✓	✓	✓
Cyanobacteria	✓	✓	✓	✓	✓	✓	✓	✓
Deferribacterales				✓	✓	✓	✓	✓
Deinococcus-Thermus			✓	✓	✓	✓	✓	✓
Dictyoglomi	✓			✓		✓	✓	✓
Elusimicrobia		✓	✓	✓	✓	✓	✓	✓
Firmicutes	✓	✓	✓	✓	✓	✓	✓	✓
Fusobacteria		✓		✓	✓	✓		✓
Melainabacteria	✓	✓		✓	✓	✓	✓	✓
Nitrospinae				✓	✓		✓	✓
Nitrospirae	✓			✓	✓	✓	✓	✓
Parcubacteria		✓	✓	✓			✓	✓
Proteobacteria	✓	✓	✓	✓	✓	✓	✓	✓
Rhodothermaeota		✓	✓	✓		✓	✓	✓
Spirochaetes	✓	✓		✓	✓	✓	✓	✓
Synergistetes	✓			✓		✓	✓	✓
Tenericutes				✓	✓	✓	✓	✓
Thermobaculum				✓	✓	✓	✓	✓

	Thermodesulfobacteria				✓			✓	✓
	Thermotogae	✓	✓		✓	✓	✓	✓	✓
	TMED		✓					✓	✓
FCB	Chlorobi		✓	✓	✓	✓	✓	✓	✓
	Caldithrix	✓	✓	✓	✓	✓		✓	✓
	Bacteroidetes		✓	✓	✓		✓	✓	✓
	Candidatus Marinimicrobia		✓	✓	✓	✓	✓	✓	✓
	Fibrobacteres	✓	✓	✓	✓			✓	✓
	Ignavibacteria		✓	✓	✓			✓	✓
	Gemmatimonadetes		✓	✓	✓	✓	✓	✓	✓
	Latescibacteria		✓	✓	✓		✓		✓
	Candidatus Kryptonium		✓	✓	✓			✓	✓
	Candidatus Kryptobacter		✓	✓	✓			✓	✓
	Candidate division Zixibacteria		✓	✓	✓			✓	✓
PVC	Chlamydia				✓		✓	✓	✓
	Planctomycetes	✓			✓	✓	✓	✓	✓
	Lentisphaerae	✓		✓	✓	✓	✓	✓	✓
	Verrucomicrobia				✓		✓	✓	✓

Table 5.1 *Distribution of Phospholipid Biosynthesis Genes in Bacterial and Archaeal Phyla. Ticks represent phyla (class level for Euryarchaeota) with at least one genome which has a sequence for the corresponding gene. Bacterial phyla where all three archaeal genes are found are indicated in yellow and green. Those bacterial phyla where all three archaeal genes are found within the same genome in at least one case are indicted in green. Those bacterial phyla with no archaeal genes are found are indicated in red. It should be noted that in the case of environmental lineages, the lack*

of a tick may not represent absence of genes, given that these represent metagenomics bins, and the lack of said genes may be due to missing data. FCB are Fibrobacteres, Chlorobi, and Bacteroidetes and related lineages. PVC are Planctomycetes, Verrucomicrobia, and Chlamydiae and related lineages. TACK are Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota. DPANN include Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota, as well as several other lineages.

Orthologues of the canonical bacterial genes are less widespread in Archaea (table 1). Of all the genomes surveyed, none contained all homologs. Of the 17 phyla shown in table 1, 8 had no bacterial homologs in any of their genomes. Orthologues of GpsA, GplA/GlpD, and Gpk are found in at least one genome of each of the major archaeal clades (Euryarchaeota, Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota (TACK), Asgardarchaeota, and Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota, as well as several other lineages (DPANN) (Williams *et al.*, 2017)). However, they appear sporadically. Within Euryarchaeota, of the seven classes surveyed, GpsA and GlpK appear in the genomes of four and GlpA/GlpD in five. Within the TACK superphylum, GlpA/GlpD and GlpK appear in Crenarchaeota and Korarchaeota, whereas GpsA appears only in a single crenarchaeote genome (Thermofilum). GpsA and GlpK are also found in at least one genome in two of the eight DPANN phyla surveyed (Woesearchaeota and GW2011, and Woesearchaeota and Parvarchaeota, respectively), whereas GlpA/GlpD is found in a single parvarchaeote genome (Candidatus Parvarchaeum acidiphilum ARMAN-4). Within the Asgardarchaeota superphylum, no orthologues for GpsA are found, and only one of the genomes (Lokiarchaeum sp. GC14_75) has GlpA/GlpD or GlpK. PlsC and PlsY are more restricted, being found mainly in environmental lineages within Euryarchaeota (Marine Groups II/III, all in class Thermoplasmatales), DPANN, and Asgardarchaeota (Table 5.1).

Early Origins of Archaeal-Type Membrane Phospholipid Biosynthesis Genes in Bacteria

To investigate the evolutionary histories of membrane phospholipid biosynthesis, we inferred Bayesian single-gene phylogenies from the amino acid alignments using

PhyloBayes 4.1 (Lartillot and Philippe, 2004; Lartillot, Brinkmann and Philippe, 2007b). We selected the best-fitting substitution model for each gene according to its Bayesian Information Criterion score using the model selection tool in IQ-Tree (Nguyen *et al.*, 2015). We used two complementary approaches to root these single-gene trees: a RMC in BEAST (Drummond and Rambaut, 2007; Drummond *et al.*, 2012) (Table 5.2), and the recently described MAD rooting method of Tria et al. (2017). The MAD algorithm finds the root position that minimises pairwise evolutionary rate variation, averaged over all pairs of taxa in the tree. Many of our single-gene trees were poorly resolved, and we wanted to account for topological uncertainty in our root estimates. To do so, we used two complementary methods: root posterior probabilities (Table 5.2) averaged over the trees sampled during the Bayesian molecular clock analysis, and the AI implemented in MAD, which is defined by Tria et al. (2017) as the ratio of the MAD value to the second smallest value (Table 5.3). For the genes for which an outgroup was available (G1PDH, GpsA, and GlpA/GlpD, following Yokobori et al. 2016), we compared our results to traditional outgroup rooting. For more details, see Materials and Methods.

Gene	RMC	MAD	RMC (metagenomic sequences removed)
G1PDH	0.68	0.62	0.4
GGGPS	0.99	1	1
DGGGPS	0.43	0.79	0.99
GpsA	0.31	0.59	0.41
GlpA/GlpD	0.5	0.44	N/A
GlpK	0.47	0.34	N/A
PlsC	0.28	0.03	N/A
PlsY	0.57	0.85	N/A

Table 5.2 *Maximum marginal posterior probabilities for molecular clock and MAD rooting methods. For several bacterial genes, removing metagenomic sequences would remove all archaeal sequences and are thus marked with "NA"; see section*

entitled "Sensitivity to Model Fitting Approach, Alignment Uncertainty, and the Inclusion of Metagenomic Sequences".

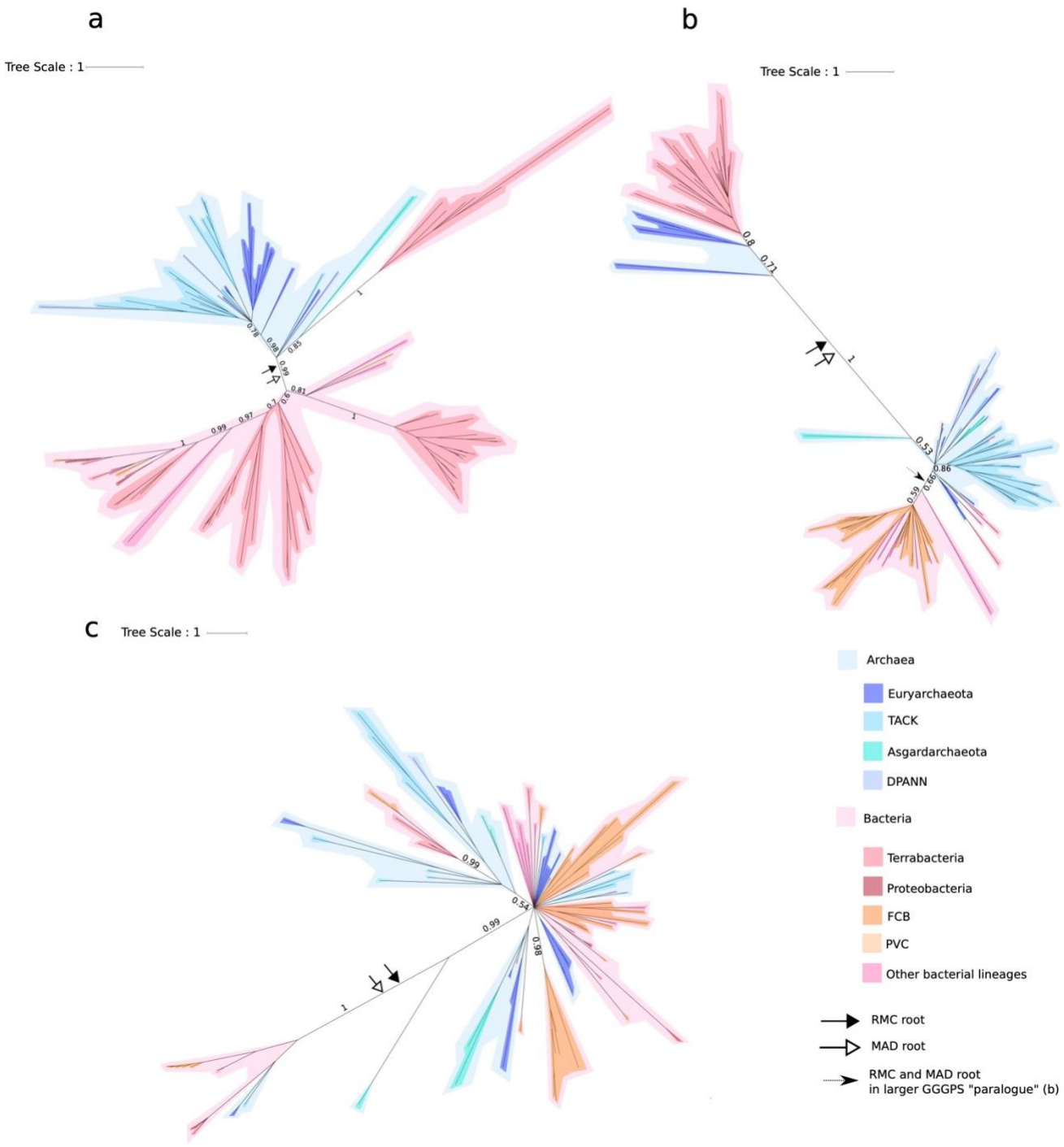
Gene	Bayesian trees	ML trees	HoT ML trees	ML trees, metagenomic data removed
G1PDH	0.997	0.976	0.956	0.973
GGGPS	0.648	0.561	0.681	0.577
DGGGPS	0.947	0.995	0.897	0.659
GpsA	0.979	0.972	0.958	0.983
GlpA/GlpD	0.964	0.906	0.902	N/A
GlpK	0.999	1	0.994	N/A
PlsC	0.996	1	0.999	N/A
PlsY	0.959	0.982	0.926	N/A

Table 5.3 *Ambiguity Index (AI) Scores for MAD roots. For several bacterial genes, removing metagenomic sequences would remove all archaeal sequences and are thus marked with "NA"; see section entitled "Sensitivity to Model Fitting Approach, Alignment Uncertainty, and the Inclusion of Metagenomic Sequences".*

G1PDH is the enzyme that establishes phospholipid stereochemistry in Archaea. Interestingly, the majority of the bacterial G1PDH orthologues do not appear to be recent horizontal acquisitions from Archaea, but instead form a deep-branching clan (Wilkinson *et al.*, 2007) (PP = 1), resolved as sister to an archaeal lineage clan (Fig. 5.2(a)). The relationships within the clans are poorly resolved. The root position that receives the highest posterior support in the RMC analysis is that between the archaeal and bacterial clans, with a marginal posterior probability of 0.68 (Table 5.2). This is substantially higher than the next most probable position, which places the root within the Bacteria with a posterior probability of 0.1. When rooted using MAD, the same root between the bacterial and archaeal clans is recovered with a marginal posterior

probability of 0.62, also substantially higher than the next most probable root of 0.1. Rooting single-gene trees can prove difficult, and this uncertainty is captured in the low root probabilities inferred using both the RMC and MAD methods. However, these analyses can be used to exclude the root from some regions of the trees with a degree of certainty. In the case of G1PDH, a post-LUCA origin of the gene would predict a root on the archaeal stem or within the Archaea. In our analyses, no such root position has a significant probability (i.e., PP > 0.05), and therefore the root is highly unlikely to be within the Archaea. This is similar to topologies recovered by Peretó et al. (2004) and Carbone et al. (2015). The bacterial clan mainly comprises sequences from Firmicutes and Actinobacteria, with most of the other Bacteria grouping together in a single, maximally supported (PP = 1) lineage suggestive of recent horizontal acquisition from the Firmicutes/Actinobacteria clade, followed by further HGT.

Fig. 5.2 (below) *Bayesian consensus trees of archaeal enzymes. Support values are Bayesian posterior probabilities. The black arrow and the white arrow indicate the modal root positions obtained using the RMC and MAD approaches, respectively. The dashed arrow indicates the RMC and MAD roots for the larger GGGPS subclade. Archaea in blue-tones and Bacteria in red/pink-tones. (a) G1PDH tree (111 sequences and 190 positions) inferred under the best-fitting LG + C60 model. (b) GGGPS tree (133 sequences and 129 positions) inferred under the best-fitting LG + C40 model. (c) DGGGPS tree (97 sequences and 119 positions) inferred under the best-fitting LG + C60 model. Terrabacteria are Firmicutes, Actinobacteria, Cyanobacteria, Chloroflexi, and related lineages. FCB are the Fibrobacteria, Chlorobi, Bacteroidetes, and related lineages. PVC are the Planctomycetes, Verrucomicrobia, Chlamydiae, and related lineages. TACK are Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota. DPANN includes Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota, as well as several other lineages. For full trees, see Appendix C, Supplementary figures 1–4. For full unrooted trees, see Appendix C, Supplementary figures 16–18.*


a
Tree Scale : 1
b
Tree Scale : 1
c
Tree Scale : 1
Archaea
Euryarchaeota
TACK
Asgardarchaeota
DPANN
Bacteria
Terrabacteria
Proteobacteria
FCB
PVC
Other bacterial lineages
RMC root
MAD root
RMC and MAD root
in larger GGGPS "paralogue" (b)

This root position is consistent with two scenarios between which we cannot distinguish based on the available data. One possibility is an early transfer of G1PDH from stem Archaea into Bacteria, either into the bacterial stem lineage with subsequent loss in later lineages, or into the ancestor of Actinobacteria and Firmicutes, with subsequent transfers to other Bacteria. Alternatively, G1PDH could have already been present in LUCA, and was subsequently inherited vertically in both Archaea and Bacteria, followed by loss in later bacterial lineages. The Firmicute sequences within the archaeal clade appear to be a later transfer into those Firmicutes, apparently from Thorarchaeota.

GGGPS attaches the first isoprenoid chain to G1P. Phylogenetic analysis of GGGPS (Fig. 5.2(b)) evidenced two deeply divergent paralogues, with the tree confidently rooted between them using both the RMC (PP = 0.99) and MAD methods (PP = 1) (Table 5.2); resolution within each of the paralogs was poor. The recovery of two distinct paralogues has been noted in several previous studies (Nemoto, Oshima and Yamagishi, 2003; Boucher, Kamekura and Doolittle, 2004; Lombard, López-García and Moreira, 2012a; Peterhoff *et al.*, 2014). One of these paralogues comprises sequences from some Euryarchaeota (including members of the Haloarchaea, Methanomicrobia, and Archaeoglobi), along with Firmicutes and Actinobacteria. The other paralogue comprises sequences from the rest of the Archaea, including other Euryarchaeota, and a monophyletic bacterial clade largely consisting of members of the FCB lineage. Taken with the root position between the two paralogues, the tree topology implies an ancestral duplication followed by sorting out of the paralogues and multiple transfers into Bacteria. Because genes from both GGGPS paralogous clades have been experimentally characterised as geranylgeranylglyceryl phosphate synthases (Nemoto, Oshima and Yamagishi, 2003; Boucher, Kamekura and Doolittle, 2004), it appears that this activity was already present in LUCA before the radiation of the bacterial and archaeal domains. It has been suggested, however, that the firmicute sequences (which comprise the majority of the sequences in the smaller paralogue) are used in teichoic acid synthesis (Payandeh *et al.*, 2006). In this case, two apparently diverging paralogues may be an artefact due to changes in the sequences during neofunctionalisation. Lombard et al. (2012b), who also find two divergent homologues, and homologues in a large diversity of FCB bacteria (mostly Bacteroidetes), suggest that one of these homologues was likely present in the last archaeal common ancestor,

whereas the bacterial sequences were likely horizontal transfers. To improve resolution among the deeper branches of the tree, we inferred an additional phylogeny focusing just on the larger of the two clades (Appendix C, Supplementary Fig. 3). The root of this subtree fell between a clade of monophyletic Bacteria and a clade of Archaea in which six bacterial sequences were interleaved, perhaps as the result of later gene transfer (PP = 0.8 for the root split, much higher than the next most likely root, within the Bacteria, with PP = 0.07). This tree might be interpreted as gene presence in LUCA, followed by some more recent transfers from Archaea to Bacteria. Given that this gene is a hallmark of archaeal membrane phospholipid biosynthesis, our data do not exclude the possibility of a very early gene transfer from the archaeal stem to Bacteria, prior to the radiation of the archaeal domain.

DGGGPS attaches the second isoprenoid chain to G1P. DGGGPS is present in all sampled Archaea, with the exception of three of the DPANN metagenome bins. Although the DGGGPS tree is poorly resolved (Fig. 5.2(c)), both the RMC and MAD root the tree between the same two clades (PP = 0.43 and 0.79, respectively) (Table 5.2). The smaller clade comprises mostly bacterial sequences from the Actinobacteria and FCB lineages, as well as two archaeal sequences (from the TACK and Euryarchaeota lineages). The larger clade contains sequences from a diversity of Bacteria, particularly FCB (also reported by Villanueva et al. 2018), as well as Archaea. DGGGPS is part of the UbiA protein superfamily, which are involved in a number of different biosynthetic pathways, including the production of photosynthetic pigments, and are therefore widely distributed in Bacteria, and are known to have undergone extensive HGT (Hemmi *et al.*, 2004). Indeed, several of the sequences used in our analyses (and those in previous studies, such as Villanueva et al. 2017) are annotated on NCBI as other proteins within this superfamily (see Supplementary Table 5). To distinguish orthologues of DGGGPS from other, distantly related members of the UbiA superfamily that might have different functions, we inferred an expanded phylogeny including our initial sequence set and sequences sampled from the other known UbiA subfamilies (Appendix C, Supplementary Fig. 25). Surprisingly, this analysis indicated that the Thaumarchaeota lack an orthologue of the DGGGPS gene that other Archaea use to attach the second isoprenoid chain; the most closely related Thaumarchaeota sequences branch within another UbiA subfamily with high posterior support (PP = 0.99). Thaumarchaeota may be using this paralog to perform the same function, or

may use another unrelated enzyme to catalyse this reaction. The wide distribution of this enzyme across both Archaea and Bacteria, and the occurrence of both domains on either side of the root, for both rooting methods, suggest either multiple transfers into Bacteria from Archaea, or that DGGGPS was present in LUCA and inherited in various archaeal and bacterial lineages, followed by many later losses in and transfers between various lineages.

In sum, our results of archaeal phospholipid biosynthesis genes suggest that there have been repeated, independent inter domain transfers of these genes from Archaea to Bacteria throughout the evolutionary history of life. Furthermore, our phylogenetic analyses do not exclude the possibility that the genes of the archaeal pathway were present in LUCA. If correct, this would imply that LUCA had the capability to make archaeal-type membrane phospholipids.

Transfers of Bacterial Membrane Phospholipid Genes into Archaea

In contrast to our analyses of proteins of the classical archaeal pathway, phylogenies of proteins of bacterial-type membrane phospholipid biosynthesis pathways are more ambiguous and the root positions are not confidently resolved. Homologs of both forms of glycerol-3-phosphate dehydrogenase (G3PDH) and GlpK are broadly distributed in Archaea, however, these three enzymes are not exclusive to phospholipid synthesis and have been shown to be used in glycerol metabolism in some autotrophic Archaea (Nishihara *et al.*, 1999). Of the enzymes thought to function exclusively in bacterial membrane phospholipid biosynthesis, we did not find any archaeal homologs for PlsB or PlsX, and archaeal PlsC and PlsY homologs are patchily distributed and are found only in metagenomic bins. It therefore seems unlikely that any of these genes function in membrane phospholipid synthesis in Archaea.

The root positions for each of the trees using both RMC and MAD have low posterior probabilities (Table 5.2), so that the exact root positions are unclear. *Gps* and *glp* are two genes that code for two forms of glycerol-3-phosphate (G3PDH), GpsA, and GlpA/GlpD, respectively, which establishes phospholipid stereochemistry in Bacteria. The deep relationships between the archaeal and bacterial sequences in the GpsA tree are poorly resolved (Fig. 5.3(a)), while being better resolved for GlpA/GlpD (Fig. 5.3(b)). The root position in both trees is poorly resolved for both rooting methods (Table 5.2). The highest marginal posterior probability for the root positions recovered

in the GpsA tree are 0.31 and 0.59 and for the RMC and MAD, respectively, and 0.5 and 0.44, respectively, for GlpA/GlpD. The tree inferred for GlpK (glycerol synthase, which can synthesise G3P from glycerol (Fig. 5.4(a)), shows a similar pattern to the phylogenies of GpsA and GlpA/GlpD. Again, the root positions have low posterior support (0.47 and 0.34 for the RMC and MAD, respectively). However, in each case, there is evidence of recent transfers from Bacteria to Archaea, as we recover several distinct bacterial and archaeal clades with moderate to high support (0.8–1), as also reported by Villanueva et al. (2017). For all three of these enzymes, the differing root positions are resolved either within the Bacteria, or with bacterial and archaeal sequences on both sides of the root. This suggests that these enzymes may have been present in LUCA, or that the archaeal sequences are later transfers from Bacteria. Due to incongruence between the rooting methods and the low supports, our analyses do not robustly reject either of these scenarios.

Fig. 5.3 (below) *Bayesian consensus trees of both G3PDH enzymes. Support values are Bayesian posterior probabilities. The black arrow and the white arrow indicate the modal root positions obtained using the RMC and MAD approaches, respectively. Archaea in blue-tones and Bacteria in red/pink-tones. (a) GpsA tree (84 sequences and 169 positions) inferred under the best-fitting LG + C60 model. (b) GlpA/GlpD tree (51 sequences and 199 positions) inferred under the best-fitting LG + C60 model. Terrabacteria are Firmicutes, Actinobacteria, Cyanobacteria, Chloroflexi, and related lineages. FCB are the Fibrobacteria, Chlorobi, Bacteroidetes, and related lineages. PVC are the Planctomycetes, Verrucomicrobia, Chlamydiae, and related lineages. TACK are Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota. DPANN includes Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota, as well as several other lineages. For full trees, see Appendix C, Supplementary figures 5 and 6. For full unrooted trees, see Appendix C, Supplementary figures 19 and 20.*

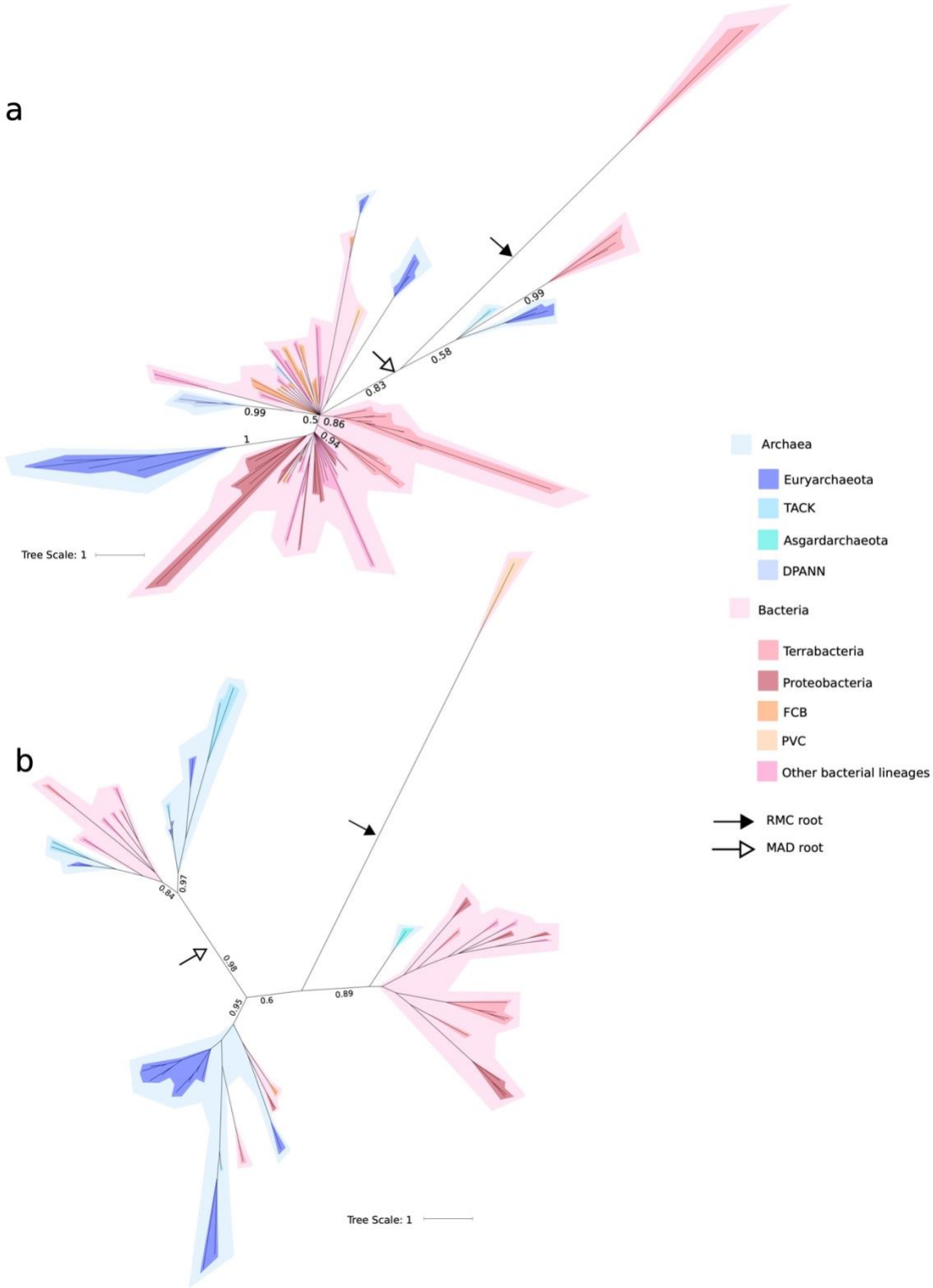
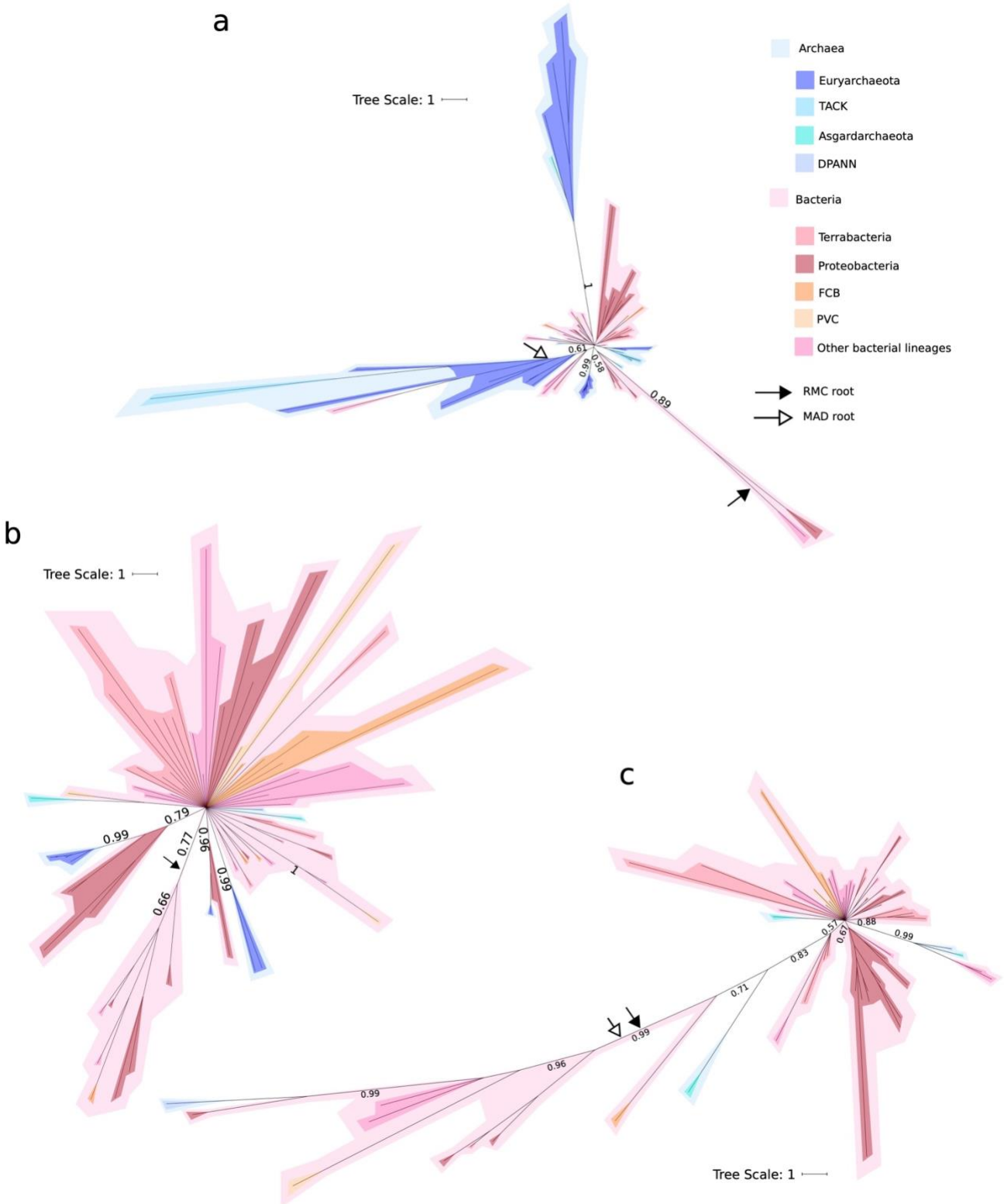

a
0.99
0.58
0.83
0.99
0.5
0.86
1
0.94
Tree Scale: 1
b
0.84
0.97
0.98
0.95
0.6
0.89
Tree Scale: 1
Archaea
Euryarchaeota
TACK
Asgardarchaeota
DPANN
Bacteria
Terrabacteria
Proteobacteria
FCB
PVC
Other bacterial lineages
RMC root
MAD root

Fig. 5.4 (below) *Bayesian consensus trees of GlpK, PlsC, and PlsY enzymes. Support values are Bayesian posterior probabilities. The black arrow and the white arrow indicate the modal root positions obtained using the RMC and MAD approaches, respectively. Archaea in blue-tones and Bacteria in red/pink-tones. (a) GlpK tree (77 sequences and 363 positions) inferred under the best-fitting LG + C60 model. (b) PlsC tree (74 sequences and 57 positions) inferred under the best-fitting LG + C60 model. (c) PlsY tree (60 sequences and 104 positions) inferred under the best-fitting LG + C50 model. Terrabacteria are Firmicutes, Actinobacteria, Cyanobacteria, Chloroflexi, and related lineages. FCB are the Fibrobacteria, Chlorobi, Bacteroidetes, and related lineages. PVC are the Planctomycetes, Verrucomicrobia, Chlamydiae, and related lineages. TACK are Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota. DPANN includes Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota, as well as several other lineages. For full trees, see Appendix C, Supplementary figures 7–9. For full unrooted trees, see Appendix C, Supplementary figures 20–23.*


a
Tree Scale: 1
Archaea
Euryarchaeota
TACK
Asgardarchaeota
DPANN
Bacteria
Terrabacteria
Proteobacteria
FCB
PVC
Other bacterial lineages
RMC root
MAD root
1
0.61
0.58
0.99
0.89
b
Tree Scale: 1
0.79
0.99
0.77
0.96
0.99
1
0.66
c
0.57
0.67
0.88
0.99
0.83
0.71
0.99
0.96
0.99
Tree Scale: 1

PlsC and PlsY (which attach fatty acids to G3P) both have many fewer orthologues among archaeal genomes, all of which are derived from environmental samples (Embley and Martin, 2006; Martin, Garg and Zimorski, 2015; Eme *et al.*, 2018). Both trees are poorly resolved (Fig. 5.4). Both are rooted within the Bacteria, with PlsC (Fig. 5.4(b)) having the low posterior of 0.28 (with the next most likely, also within the Bacteria, being 0.1). The PlsY (Fig. 5.4(c)) has a more certain root position, with a posterior of 0.57, and the next most probable being 0.1. For PlsY, MAD recovers the same root as the molecular clock, with a high posterior probability (0.85). When the PlsC tree is rooted using MAD, the root is resolved between two clades, which are not recovered in the inferred tree topology (see Appendix C, Supplementary Fig. 8) and has a low posterior probability of 0.03. All of the archaeal homologs seem to be horizontal acquisitions from Bacteria.

Sensitivity to Model Fitting Approach, Alignment Uncertainty, and the Inclusion of Metagenomic Sequences

The deep branches of our trees are in general poorly resolved, a problem that is sometimes encountered when inferring phylogenies for ancient single genes (Williams, Martin Embley and Heinz, 2011). We therefore performed sensitivity analyses to evaluate the robustness of our biological conclusions to some of the key decision points in our phylogenetic approach. Our focal analyses are Bayesian, so we also inferred trees using the same models in the maximum likelihood framework using IQ-Tree (see Appendix C, Supplementary figures 30–37 for ML topologies, and Table 5.3 for MAD AI scores). The topologies were closely similar to the Bayesian trees, with the exception of some poorly supported clades that are resolved in the ML tree but are not present in the Bayesian majority rule consensus tree. The root positions on the G1PDH, GGGPS, DGGGPS, GpsA, GlpA/GlpD, and PlsC ML trees were identical to those on the Bayesian trees. The MAD root positions for GlpK and PlsY ML trees differ from the Bayesian trees, but in both cases the root positions are on adjacent branches and the changes do not substantially alter our interpretations (Appendix C, Supplementary figures 35–37).

We evaluated the impact of alignment uncertainty on our results using heads-or-tails (Landan and Graur, 2007). The reverse alignments were used to infer ML trees in IQ-Tree using the LG + C60 model. These were broadly congruent with the ML and

Bayesian trees on the original alignments, with only minor topological differences in poorly resolved areas of the trees (Appendix C, Supplementary figures 38–45). The root positions on the G1PDH, GGGPS, DGGGPS, GlpA/GlpD, GpsA, and PlsC ML trees were identical to those on the Bayesian trees (Appendix C, Supplementary figures 38–40, 42, 44). The MAD root positions for GlpK (Appendix C, Supplementary Fig. 43) and PlsY (Appendix C, Supplementary Fig. 45) ML trees differ for the Bayesian trees, but in the case of PlsY, the root positions is on an adjacent branch. The MAD root position for GlpK falls between a *Korarchaeum* sequence and the rest of the tree.

Due to errors in assembly, metagenome bins sometimes incorporate sequences from more than one underlying organismal genome (Parks *et al.*, 2015). To evaluate whether some apparent gene transfers might be artefacts of metagenome assembly, we repeated our analyses without the inclusion of metagenome-derived sequences, where possible. We performed these analyses for G1PDH, GGGPS, DGGGPS, GpsA, GlpA/GlpD, and GlpK, but not for PlsC or PlsY, because all of the archaeal sequences for these trees are derived from metagenome bins (see Supplementary Table 5). In the six cases where a reasonable comparison can be made, the topologies and roots of the trees were closely similar to those in the full analysis (Appendix C, Supplementary figures 46–52).

These results suggest that, although our analyses do include substantial topological uncertainty, our overall conclusions are not driven by issues with alignment, metagenome-derived sequences, or the choice of model fitting approach (maximum likelihood or Bayesian).

Comparing Outgroup and Outgroup-Free Rooting for Single-Gene Trees

Evolutionary interpretations typically depend on rooted trees, but rooting single-gene trees can prove difficult. The most widely used approach is to place the root on the branch leading to a predefined outgroup (Penny, 1976). However, this can be challenging for ancient genes when closely related outgroups are lacking; either the outgroup method cannot be used at all, or else the long branch leading to the outgroup can induce errors in the ingroup topology (a phenomenon known as long branch attraction (LBA) (Gouy, Baurain and Philippe, 2015)).

In the case of phospholipid biosynthesis, some of the key genes belong to larger protein families whose other members, although distantly related, have conserved structures and related functions (Peretó, López-García and Moreira, 2004). Several previous studies looking at the history of phospholipid biosynthetic genes have used these outgroups for rooting. Due in part to the difficulties of outgroup rooting for ancient genes, these studies have disagreed on the roots for some of these gene trees, leading to very different evolutionary conclusions. Our outgroup-free results are consistent with those of Peretó et al. (2004) and Carbone et al. (2015), but not with those of the recent study of Yokobori et al. (2016). Yokobori et al. used outgroups to root trees for G1PDH, G3PDH (both GpsA and GlpA/GlpD) and GlpK. Their root inferences differed from ours in that they found that bacterial G1PDH sequences formed a monophyletic group that branched from within Archaea, suggesting more recent horizontal transfer from Archaea to Bacteria, as opposed to transfer from stem Archaea or vertical inheritance from LUCA (Fig. 4.2(a)). On the other hand, their analysis of GlpA/GlpD recovered Bacteria on one side of the root, and a clade of Bacteria and Archaea on the other. They interpreted this as evidence for the presence of GlpA/GlpD in LUCA, and therefore that LUCA would have had bacterial-type G3P membrane phospholipids.

Single-matrix models, such as those used by Yokobori et al. (2016), have been shown to be more susceptible to phylogenetic artefacts such as LBA than the profile mixture models used here (Lartillot, Brinkmann and Philippe, 2007b). To investigate whether the differences in root inference between our analyses and those of Yokobori et al. (2016) might be the result of LBA, we performed outgroup rooting analysis on G1PDH, GpsA, and GlpA/GlpD, augmenting our data sets with a subsample of the outgroups used by Yokobori et al. and using the same models used to infer the unrooted trees (LG + C60 in each case). The resulting trees (Appendix C, Supplementary figures 10–12) show different topologies when compared with the unrooted trees (Appendix C, Supplementary figures 16, 19, and 20). This suggests that the long branch outgroup may be distorting the ingroup topology.

We also performed model testing in IQ-Tree and compared the fit of the chosen models to the models used by Yokobori et al. (see Material and Methods). LG + C60 was selected for both G1PDH and GlpA/GlpD, whereas LG + C50 was selected for GpsA (Appendix C, Supplementary Fig. 24). The results of these analyses indicate that the

empirical profile mixture models which we have used here fit each of these alignments significantly better than the single-matrix models of Yokobori et al. (Table 5.4). However, even analyses under the best-fitting available models show distortion of the ingroup topology upon addition of the outgroup (Appendix C, Supplementary figures 10–12 and 24), when compared with the unrooted topologies (Appendix C, Supplementary figures 16, 19, and 20). In each case, we found the root in a different place to those recovered by Yokobori et al. In the G1PDH tree, we find Bacteria (Firmicutes) to be most basal, rather the Crenarchaeota found by Yokobori. In the case of GpsA, Yokobori et al. did not find compelling support for an origin in LUCA, but they did recover one archaeal lineage (the Euryarchaeota) at the base of the ingroup tree with low (bootstrap 48) support. Although our GpsA tree is also poorly resolved, we do not find evidence to support the basal position of the archaeal lineages, and therefore for the presence of GpsA in LUCA. For GlpA/GlpD, which Yokobori et al. trace back to LUCA due to the basal position of the archaeal sequences, the outgroup sequences did not form a monophyletic group, and were instead distributed throughout the tree (Appendix C, Supplementary Fig. 11). Thus, analyses under the best-fitting available models did not support the presence of bacterial lipid biosynthesis genes in LUCA. Further, the distortion of the ingroup topologies suggests that these outgroups may not be suitable for root inference, at least given current data and methods. The RMC and the MAD methods have their own assumptions and limitations, but these results suggest that they may be useful for rooting trees in other contexts, either as part of a sensitivity test or when suitable outgroups are not available.

Gene (with selected model)	BIC score	BIC score LG+gamma (used by Yokobori et al.)
G1PDH (LG+C60)	43392.094	44227.872
GlpA/GlpD (LG+C60)	28433.114	28526.816
GpsA (LG+C50)	34483.395	34604.88

Table 5.4 *Bayesian Information Criterion (BIC) scores for outgroup rooted trees under IQ-Tree model selection*

Origin of Eukaryotic Membrane Phospholipid Biosynthesis Genes

Phylogenetics and comparative genomics suggest that eukaryotes arose from a symbiosis between an archaeal host cell and a bacterial endosymbiont that evolved into the mitochondrion (Embley and Martin, 2006; Martin, Garg and Zimorski, 2015; Eme *et al.*, 2018). Genomic and phylogenetic evidence indicates that the host lineage belonged to the Asgardarchaeota superphylum (Spang *et al.*, 2015; Zaremba-Niedzwiedzka *et al.*, 2017). The origin of bacterial-type membrane phospholipids in eukaryotes is therefore an important evolutionary question that has received considerable attention (Woese, Kandler and Wheelis, 1990; Kandler, 1995; López-García and Moreira, 2006; Baum and Baum, 2014; Gould, Garg and Martin, 2016). Given the evidence for transfer of bacterial-type phospholipid biosynthesis genes into Archaea, one possibility, also raised by the results of Villanueva et al. (2017), is that eukaryotes may have inherited their bacterial lipids vertically from the archaeal host cell. Both our study and that of Villanueva et al. (2017) point to the presence of orthologues for bacterial lipid genes in Asgardarchaeota. These include GlpA/GlpD, PlsC, and PlsY orthologues in Lokiarchaeum sp. GC14_75, PlsC, and PlsY in Heimdallarchaeota archaeon LC_2, and PlsY in Thorarchaeota archaeon SMTZ1-83 (Table 1). However, phylogenies of these genes (Appendix C, Supplementary figures 13–15) do not support a specific relationship between eukaryotes and any of the archaeal sequences, and so do not provide any compelling support for an origin of eukaryotic lipids via the archaeal host cell.

5.4 Conclusions

Our phylogenetic analyses of lipid biosynthesis genes support two main conclusions about prokaryotic cell physiology and early cell evolution. First, our results corroborate previous evidence for extensive horizontal transfer of lipid genes, particularly from Archaea to Bacteria, from potentially very early to more recent evolutionary times. The functions of these genes remain unclear, but in *B. subtilis* (Guldan, Sterner and Babinger, 2008; Guldan *et al.*, 2011) they are involved in making archaeal-type G1P ether-linked lipids, whereas in the FCB lineage Cloacimonetes (Villanueva et al. 2018) they may be involved in synthesizing archaeal-type phospholipids that are incorporated

into the bacterial cell membrane. Evidence that these genes have undergone horizontal transfer, both early in evolution and more recently, provides a potential mechanism for the remarkable diversity of membrane lipids, and especially ether lipids, in environmental settings (Schouten, Wakeham and Sinninghe Damsté, 2001). We also note that it is intriguing that bacterial lipids with archaeal features are particularly abundant in settings characterised by high archaeal abundances, including cold seeps, wetlands and geothermal settings (Schouten, Hopmans and Sinninghe Damsté, 2013), potentially providing ecological opportunity for gene transfer. Experimental work to characterise the enzymes that make these environmental lipids will be needed to test this prediction.

A second, and more tentative, result of our study relates to the antiquity of the canonical archaeal and bacterial pathways. Our analyses suggest that the enzymes for making G1P lipids may have been present in the common ancestor of Archaea and Bacteria. Under the consensus view that the root of the tree of life lies between Bacteria and Archaea, this would imply that LUCA could have made archaeal-type membranes. This finding is intriguing in light of previous work suggesting the presence of isoprenoids produced by the mevalonate pathway in LUCA (Lombard and Moreira, 2011; Castelle and Banfield, 2018). By contrast, we found no positive evidence to suggest that the bacterial pathway was present in LUCA, although our gene trees are poorly resolved and so we cannot exclude this possibility. The consensus universal root between Bacteria and Archaea is supported by analyses of ancient gene duplications (Iwabe *et al.*, 1989; J. P. Gogarten *et al.*, 1989; Zhaxybayeva, Lapierre and Gogarten, 2005) and genome networks (Dagan *et al.*, 2010), but some analyses have supported an alternative placement of the root within Bacteria (Cavalier-Smith, 2006; Lake *et al.*, 2009; Williams *et al.*, 2015). Our trees do not exclude a within-Bacteria root, in which case LUCA would have possessed the bacterial pathway, and the archaeal pathway would have evolved along the archaeal stem, or in a common ancestor of Archaea and Firmicutes (Cavalier-Smith, 2006; Lake *et al.*, 2009).

If one membrane lipid pathway evolved before the other, this would imply that one of the two prokaryotic lineages changed its membrane lipid composition during early evolution. The evolutionary processes that drive such changes remain unclear, in part because we still do not fully understand the functional differences between modern

archaeal and bacterial membranes. Compared with bacterial-type membranes, archaeal-type membranes maintain their physiochemical properties over a broader range of temperatures and may be more robust to other environmental extremes (Vossenberg *et al.*, 1998; Koga, 2012). If the archaeal pathway is older than the bacterial pathway, then that could reflect a LUCA adapted to such extreme settings. It is then intriguing to speculate on the evolutionary drivers for subsequent adoption of bacterial-type membranes, especially because the Bacteria appear to be more successful than the Archaea in terms of abundance and genetic diversity (Danovaro *et al.*, 2016; Hug *et al.*, 2016; Castelle and Banfield, 2018). Moreover, an analogous change has happened at least once in evolutionary history, during the origin of eukaryotic cells (Martin, Garg and Zimorski, 2015). Chemical considerations suggest such bonds ought to be energetically cheaper to make and break, although we know of no published experimental data on these relative biosynthetic costs. Alternatively, bacterial-type membrane lipids comprise a variety of fatty acyl moieties, varying in chain length, unsaturation, degree of branching and cyclisation, and these could impart a degree of flexibility and adaptability that provides a marginal benefit in dynamic mesophilic environments. If so, that advantage could translate to bacterial ether lipids that are also widespread in non-extreme settings and also characterised by a variety of alkyl forms (Pancost *et al.*, 2001). Conversely, if bacterial-type membranes were ancestral, the transition to archaeal-type membranes could have been driven by adaptation to high environmental temperatures: ether bonds are more thermostable than esters (Vossenberg *et al.*, 1998; Koga, 2012) and are also found in the membranes of thermophilic Bacteria (Kaur *et al.*, 2015). In any case, the widespread occurrence of bacterial-type, archaeal-type, and mixed-type membrane lipids in a range of environments, as well as the widespread occurrence of the associated biosynthetic genes across both domains, suggests that except for high temperature and low pH settings, the advantages of either membrane type is marginal.

Chapter 6

Towards an integrated model of early bacterial evolution

This chapter is not part of any publication and has been written by GAC entirely.

Abstract

In this Chapter, we attempt to summarise and integrate the results from the previous chapter to answer the questions laid out at the beginning of the thesis, and better conceptualise our narrative of early bacterial evolution. We outline the contributions made by this thesis to evolutionary biology and microbiology, as well as attempting to address possible caveats in the analyses. We discuss the difficulties in addressing questions of deep phylogeny and possible future directions which will improve research in this field.

6.1 Addressing the challenges of deep-time phylogenetics

Choosing the right substitution models

A key issue in phylogenetic reconstruction is selecting the best substitution model for your analysis (Ripplinger and Sullivan, 2008; Hoff *et al.*, 2016). As topological artefacts, such as long branch attraction (LBA), can be highly problematic when reconstructing deep-time phylogenies (Felsenstein, 1978; Lartillot, Brinkmann and Philippe, 2007), we must choose adequate models to deal with these issues. We show in Chapter 5 that single-matrix models which do not account for across-site variation perform significantly worse than profile mixture models. Furthermore, we demonstrate that when subjecting the data of Yokobori *et al.* to model testing, more complex models with site specific composition profiles are chosen over simpler models they use in their paper, illustrating that these more complex models inevitably fit the data better than the simpler ones. This additionally indicates the importance of using model testing when performing phylogenetic analyses, to ensure the best available models are being used for the data in question. As more complex models also necessitate the use of smaller taxon samples in order to be computationally tractable, large scale analyses of global prokaryotic diversity (Hug *et al.*, 2016; Parks *et al.*, 2017, 2018; Castelle and Banfield, 2018; Zhu *et al.*, 2019) currently preclude the use of these models, leading to possible topological artefacts. In Chapter 2, we therefore use a smaller taxon sampling to make use of the more complex substitution models with site specific composition profiles when inferring our species trees. It should be noted however, that our taxon sampling is still one of the largest bacterial datasets on which such site-heterogeneous models have been used. Sparse taxon sampling can also lead to artefacts, especially regarding under-sampled lineages that may be deep-branching and fast-evolving (Bergsten, 2005). With increased computational resources, we will be able to expand our taxon sampling, which can help to resolve difficult phylogenetic problems (Graybeal, 1998; Hedtke, Townsend and Hillis, 2006), while still applying the best fitting models. However, the choice of taxa may be more important than number, as having many taxa which are long branching or compositionally biased may cause further artefacts (see outgroup rooting section below). Yet even the best models are crude approximations of evolution, and as they improve, so will the results of our analyses.

Outgroup rooting is problematic

Another key theme through this thesis has been exploring the problems associated with outgroup rooting. As previously explained, outgroup rooting is problematic because it requires prior phylogenetic knowledge, can cause LBA artefacts (Gouy, Baurain and Philippe, 2015), and reduces the amount of data that can be used. We demonstrate some of these issues in Chapter 2. We show that results obtained from outgroup rooting are highly sensitive to taxon sampling, and may be affected by composition-driven LBA artefacts as demonstrated by the different results obtained with a recoded alignment. Furthermore, we could not statistically distinguish between several alternative topologies when an archaeal outgroup was used. In particular, the oversampling of CPR in the GTDB-independent dataset (based on Hug *et al.,* 2016) is likely distorting the topology further, demonstrating the use of more taxa many not always be ideal for solving LBA artefacts, especially if these taxa have long branches or are in some way compositionally biased. These results make it clear that using outgroups is not appropriate when rooting deep-time phylogenies such as Bacteria, due to such distance between the ingroup and the outgroup. In Chapter 5, we further demonstrate the difficulties in using outgroups to root gene trees. The use of outgroups from Yokobori *et al.* (2016) with our taxa resulted in substantially different topologies from those that they recovered, again demonstrating the effect different taxon sampling has when using outgroups. Furthermore, we found that the ingroup topologies were distorted, namely that distantly related genes are causing LBA artefacts. This was found to be the case even when the best fitting models, with site specific composition profiles, were used.

Merits and drawbacks of outgroup-free rooting approaches

The relaxed molecular clock (RMC) (Thorne, Kishino and Painter, 1998; Kishino, Thorne and Bruno, 2001) and minimal ancestor (MAD) rooting (Tria, Landan and Dagan, 2017) are two possible alternatives to outgroup rooting, as they do not require outgroups and therefore can avoid issues surrounding prior phylogenetic knowledge and LBA artefacts. As such, in Chapter 5 we present evidence to show that such approaches may be more appropriate for rooting single gene trees. However, when applied to rooting species trees, such as the tree of Bacteria in Chapter 2, we see that both rooting methods are susceptible to changes in taxon sampling and recoding of

the data. The root positions obtained were also often not statistically distinguishable from other roots. This suggests that methods lack statistical power and are not consummate. Indeed, RMC rooting information, in the absence of calibrations, is derived from modelling rates across branches, with may be susceptible to long branches, saturation and compositional biases. Similarly, MAD assumes that the root minimises rate deviation over the tree, which might not be true. These issues are particularly evident with regards to the sampling of CPR, where, as with outgroup rooting, oversampling of this clade leads to LBA artefacts. Additionally, both methods still use only a small amount of data to inform their choice of root. In contrast, we show that amalgamated likelihood estimation (ALE) seems to not be susceptible to these issues, giving the same root region regardless of the taxon sampling used. Even with the oversampling of the CPR in the GTDB-independent dataset, the same result is obtained. The method is also informed by much more data, and allows us to model both vertical and horizontal components of evolution. Furthermore, we show that ALE, along with other species tree aware outgroup methods, outperforms species unaware rooting methods, and may thus be the most appropriate rooting method for deep phylogenies.

6.2 The root of the bacterial tree between two large and diverse clades

In Chapter 2, we show that the tree of Bacteria comprises two large radiations, the Gracilicutes and the Terrabacteria. Branching between these two large clades are several smaller phyla, including the Fusobacteriota, Synergistota and Thermotogota. The larger clades are stable across different taxon samplings and when heterogeneous sites are removed, demonstrating that key features of bacterial topology are not being due to composition or taxon sampling driven LBA artefacts. The smaller phyla, however, are more susceptible to such artefacts. Better taxon sampling and a greater understanding of the biology of these phyla may help place them in the tree with higher confidence. We further demonstrate that the root of the bacterial tree falls between these two clades, although we could not resolve the root in relation to the aforementioned small phyla. We find no support for the root on the

CPR branch, as proposed in recent outgroup rooted analyses (Hug *et al.*, 2016; Parks *et al.*, 2017; Castelle and Banfield, 2018; Zhu *et al.*, 2019), with CPR rather being a derived lineage within Terrabacteria. Instead, our analyses suggest CPR basal topologies may be LBA artefacts. We also do not find evidence to support the placement of the root on the branches to the thermophilic taxa Aquificota and Themotogota (Bocchetta *et al.*, 2000; Bern and Goldberg, 2005; Barion *et al.*, 2007; Battistuzzi and Hedges, 2009), Planctomycetes (Brochier and Philippe, 2002), or Chloroflexota (Cavalier-Smith, 2006).

Furthermore, our analyses in Chapter 2 suggest that HGT has had far-reaching effects on bacterial evolution, and therefore an essential component to model. However, we have demonstrated that the majority of gene families evolved vertically most of the time. Although almost no gene families have escaped HGT completely, even those which are transferred frequently may contain vertical signal. Therefore, a tree my still be an apt representation of bacterial evolution, including the deepest branches of the tree (Creevey *et al.*, 2004; Koonin, Wolf and Puigbò, 2009; Puigbò, Wolf and Koonin, 2010), although additional work is stilll needed to help elucidate this further.

6.3 An acetogenic origin of Bacteria?

Attempting to reconstruct the central metabolic pathways of the earliest Bacteria has proven challenging. Some pathways may be poorly recovered across the early nodes due to their sparse distributions in modern taxa caused by multiple losses. Others may be over-represented due to wide occurrence in modern taxa caused by extensive HGT. While ALE models these variables to circumnavigate these problems, like all models it is a simplified extraction of the true evolutionary process, albeit one that is more comprehensive than non-genomic, non-species aware approaches. Similarly, the reduced taxon sampling necessary to make the analyses tractable also reduced the probability of recovering gene families in the root. We further use different origination priors for each COG functional category derived from theory (Jain, Rivera and Lake, 1999) which, while being a large improvement on using a flat prior, are still too broad, as many gene families within the same category evolving very differently.

These limitations notwithstanding, we believe we have improved on previous attempts to reconstruct ancestral metabolism of prokaryotic cells. Many such attempts are framed in the context of hypotheses regarding the environment of the early Earth, which rely on assumptions for which the data is limited and difficult to interpret. In contrast, while such hypotheses may act as lines of supporting evidence, we infer our constructions based on an explicit model of originations, duplication, transfers and losses (ODTLs) with origination priors on each COG category. Such a model based approach allows us to give predictions with different degrees of support for the presence of different genes, and therefore to test different hypotheses on the evolution of different metabolic pathways. As the parameters of the model are explicit, we can also identify areas where the model can be improved in the future (discussed in the last section of this chapter).

Based on the metabolic reconstruction in Chapters 3 and 4, we tentatively hypothesise that the earliest bacterial cells were anaerobic acetogens. Although the hallmark enzyme of the Wood-Ljungdahl Pathway (WLP), the carbon monoxide dehydrogenase/acetyl-CoA synthase (CODH/ACS) had only moderate root support, the presence of the methyl branch, along with acetate kinase and an Na^+-translocating ferredoxin:NAD^+ oxidoreductase (Rnf) complex suggest that early Bacteria had the capability of facultative acetogenic growth (Schuchmann and Müller, 2014). This is congruent with previous research suggesting the WLP to be the most ancient form of carbon fixation (Fuchs, 2011; Sousa and Martin, 2014; Williams *et al.*, 2017; Adam, Borrel and Gribaldo, 2018). Based on the presence of acetogenesis, these early cells were likely anaerobic with carbon dioxide as a possible electron acceptor. The capability for acetogenesis was subsequently lost in other lineages, for example in Actinobacteriota and the ancestor of CPR and Chloroflexota.

We found some evidence for the presence of a respiratory nitrate reductase in LBCA, suggesting that it may have had the ability for anaerobic respiration of nitrate, although we do not recover this in other nodes. We further recover evidence of terminal oxidases in some nodes, although the distributions are inconsistent and other components of those protein complexes are not found. We find no evidence for the respiration of sulphur in any nodes, contrary to hypotheses regarding the early evolution of sulphur metabolism (Wagner *et al.*, 1998; Shen, Buick and Canfield,

2001). The presence of other carbon fixation pathways has been difficult to determine. Despite some research suggesting some combination of the 3-hydroxypropionic cycle with other pathways (Marakushev and Belonogova, 2011, 2013; Braakman and Smith, 2012), we find no evidence for this pathway. We also find little for the Calvin Cycle. The inference of some components of the TCA cycle and glycine pathway in LBCA, and the knowledge tht they can be both reductive and oxidative (Nunoura *et al.* 2018; Sánchez-Andrea 2020), gives some support to the presence of either or both the reverse TCA cycle (Wächtershäuser, 1990; Cody *et al.*, 2001; Smith and Morowitz, 2004; Nunoura *et al.*, 2018) or the reductive glccyine pathway (Sáchez-Andrea 2020). However, the lack of the hallmark enzymes of the pathways renders it difficult to assess in which direction these pathways were being used.

6.4 The early origin of motile diderm cells

In Chapters 3 and 4, we recover proteins involved in the construction of the outer cell membrane, including for LPS biosynthesis, in LBCA and in subsequent nodes, with the loss of these pathways in Actinobacteriota and the common ancestor of Chloroflexota and CPR. Taken together, these results are consistent with hypotheses in which the earliest bacterial cells were diderms (Cavalier-Smith, 2002; Antunes *et al.*, 2016; Megrian *et al.*, 2020). This is contrary to hypotheses which argue that the outer membrane arose later in bacterial evolution, under antibiotic selection pressure (Gupta, 2011), via endosymbiosis between monoderms (single-membraned bacteria (Lake, 2009)) or via the arrest of sporulation (Tocheva, Ortega and Jensen, 2016). Subsequent diderm-to-monoderm transitions have occurred on multiple occasions within Bacteria (Antunes *et al.*, 2016; Megrian *et al.*, 2020). Furthermore, the presence of flagellar in the deepest nodes further supports the idea that the earliest Bacteria were motile organism (Liu and Ochman, 2007a, 2007b), in contrast to scenarios where the first bacterial cells were non-motile, living on mineral substrates (Martin and Russell, 2007; Lane and Martin, 2012; Sousa *et al.*, 2013; Sousa and Martin, 2014). Flagella and pili have been lost in various later lineages, with cells becoming non-motile or evolving other means of motility (Miyata *et al.*, 2020).

6.6 Diversification of Bacteria through time

In Chapter 4, we use patterns of gene transfer derived from our ALE analysis to infer relative order events within Bacteria diversification. We show that Terrabacteria, and many of the clades within it, diversified earlier than the Gracilicutes. CPR, despite its relatively derived position in the tree, diversified relatively early in bacterial evolutionary history. Given the early branching of DPANN within Archaea (Williams *et al.*, 2017), this implies an early origin for highly reduced and metabolically streamlined organisms, potentially living symbiotically with other contemporaneous cells. The earliest diversifying lineages within Terrabacteria (excluding the CPR) are all inferred to have been acetogenic, further pointing to the ancient origins of this pathway. In contrast, lineages that are less likely to be acetogenic, such as Actinobacteriota and Cyanobacteria, diversify later, implying the later evolution of other metabolic pathways, including the aerobic metabolisms. Additionally, the later origins of the Cyanobacteria and Alphaproteobacteria imply that eukaryotic cells arose late in bacterial diversification. Although the Terrabacteria diversify early, we do not find evidence that this corresponds with an early adaptation to terrestrialisation, as has been previously suggested (Battistuzzi, Feijao and Hedges, 2004; Battistuzzi and Hedges, 2009). Terrestrialisation likely happened later in individual lineages, although the evidence is difficult to assess. In Chapter 4, we also see a gradual increase in genome size through the tree, congruent with results in Chapter 2 that most phyla acquire most of their genes in the crown group. This major exception is in the CPR, where genomes become highly reduced, congruent with previous findings (Castelle *et al.*, 2018). A relationship between terrestrialisation and genomes size (Wu *et al.*, 2014) may be present, but further study would be needed to investigate this further.

6.6 The Lipid divide

In Chapter 5, we provide evidence to show that there has been extensive HGT of genes involved in the synthesis of phospholipids between Archaea and Bacteria, particularly from the former to the latter. Such extensive transfers can provide an explanation for the existence of various phospholipids of mixed characteristics found

in the environment (Schouten *et al.*, 2000; Weijers *et al.*, 2006; Damsté *et al.*, 2007; Schouten, Hopmans and Sinninghe Damsté, 2013), as well as specific examples of Archaea with lipids with some bacterial characteristics (Gattinger, Schloter and Munch, 2002) and vice versa (Guldan, Sterner and Babinger, 2008; Goldfine, 2010; Guldan *et al.*, 2011). More tentatively, we present phylogenetic evidence for the possible earlier appearance of the archaeal pathway and its presence in the last universal common ancestor (LUCA). In this scenario LBCA would have inherited archaeal phospholipids, which would have been lost independently in many bacterial lineages, while being retained in others. The bacterial pathway would have evolved in LBCA and been retained by most modern lineages. However, we cannot rule out the possibility of multiple transfers from Archaea to Bacteria, either deep in the bacterial tree, or more recently. In Chapters 3 and 4, we infer the metabolic capabilities of LBCA and several deep nodes in the bacterial tree, including genes for phospholipid biosynthesis. We recover evidence for the presence of bacterial G3P lipids in early bacterial lineages, excluding the CPR. In contrast, we do not find evidence for any proteins involved in the synthesis of archaeal phospholipids in LBCA, or in any other node, including the ancestors of Firmicutes, and Actinobacteriota and Firmicutes respectively. This implies that the presence of archaeal phospholipid synthesis genes in these bacterial clades may be due to more recent HGTs. Alternatively, it may be that the patchy distribution of the genes in modern species has resulted in low presence posterior probabilities.
Or because we aren't modelling properly – i.e. O_R per COG category rather the per gene family.

We believe an acellular LUCA (Koga *et al.*, 1998; Martin and Russell, 2003) is unlikely. Regardless of the stereochemistry, phospholipids are fundamentally similar in their architecture and biochemistry, and enzymes unique to the Bacteria and Archaeal are part of larger gene families found in both families. We recover membrane bound ATPases in LBCA, congruent with similar results inferred in Archaea (Williams *et al.*, 2017), and LUCA (Sojo, Pomiankowski and Lane, 2014; Weiss *et al.*, 2016). The presence of some genes for lipid biosynthesis have also been inferred in LUCA (Lombard and Moreira, 2011; Lombard, López-García and Moreira, 2012; Koga, 2014; Weiss *et al.*, 2016). These lines of evidence imply that LUCA had a membrane, although its properties may have differed from those of modern, ion-tight prokaryote cell membranes (Lombard, López-García and Moreira, 2012; Koga, 2014; Sojo,

Pomiankowski and Lane, 2014). If our inferences in Chapter 5 of the greater antiquity of the archaeal pathway are true, then we could infer that LUCA had a homochiral membrane with G1P phospholipids which were present in LBCA and lost in various subsequent bacterial lineages. Given the evidence outlined above and in the Chapters 3-5 however, we cannot exclude a possible heterochiral membrane (Wächtershäuser, 2003; Peretó, López-García and Moreira, 2004), or a membrane completely unlike that of modern prokaryotic cells.

6.7 What can we say about the last universal common ancestor?

The consensus view is that the root of the tree of life lies between Archaea and Bacteria (Gogarten *et al.*, 1989; Iwabe *et al.*, 1989; Brown and Doolittle, 1995; Zhaxybayeva, Lapierre and Gogarten, 2005), although alternative scenarios have been proposed (Cavalier-Smith, 2006; Skophammer *et al.*, 2007; Lake *et al.*, 2009). Our analyses in Chapter 2 place the root between Archaea and Bacteria, although they are not comprehensive. Nonetheless, with the assumption that the root does indeed lie between the two domains, in comparing our analysis in this thesis with previous work both on Bacteria and on Archaea, we may be able to make some inferences about LUCA. Our tentative inference of acetogenesis in LBCA is congruent with other research (Adam, Borrel and Gribaldo, 2018), and with research suggesting the presence of the WLP in the last archaeal common ancestor (LACA) (Williams *et al.*, 2017; Adam, Borrel and Gribaldo, 2018), and in LUCA (Fuchs, 2011; Sousa *et al.*, 2013; Sousa and Martin, 2014; Weiss *et al.*, 2016; Adam, Borrel and Gribaldo, 2018). This suggests that LUCA may have been an anaerobic autotrophs using the WLP. The double membrane is likely an innovation in the bacterial stem, as Archaea are largely monderm, with some lineages having independently evolved diderm-like morphologies, although these differ strongly from diderm bacterial cell envelope (Klingl, 2014). Motility machineries seem to have evolved independently in both domains (Thomas, Bardy and Jarrell, 2001). As outlined above, given the presence of essentially similar phospholipids in both domains, and the inference of membrane bound proteins (Sojo, Pomiankowski and Lane, 2014; Weiss *et al.*, 2016), and some genes for lipid biosynthesis (Lombard and Moreira, 2011; Lombard, López-García and

Moreira, 2012; Koga, 2014; Weiss *et al.*, 2016) in LUCA, we believe that LUCA likely possessed a membrane. However, the evidence is not completely clear.

In sum, these analyses suggest that LUCA was a cellular, non-motile anaerobe using the WPL. However, it must be noted that these are pure inferences based on our results in Bacteria, and other previous studies, and thus should be viewed with caution. To determine the answers to questions concerning LUCA, we would need to perform extensive and thorough analyses, including repeating the analyses carried out in this thesis on the entire tree life. Nonetheless, it is intriguing, based on the results we do have, to speculate on these questions.

6.8 Future directions

In summary, in this thesis we have presented a rooted tree of Bacteria, and shown that the retention of vertical single in the data illustrates that the use of the “tree” analogy to describe the evolution of Bacteria may still be apt. However, we also demonstrate that the horizontal signal is still of great importance, and that both signals must be modelled to allow a fuller understanding of prokaryotic evolutionary history. We have additionally shown that using the right methods and evolutionary models is of utmost importance, and in particular that outgroup rooting may not be suitable for rooting deep phylogenies. Finally, we infer that the earliest bacterial cells were free living, motile, diderms, and were tentatively anaerobic acetogens. The loss of the outer membrane, motility machinery, and evolution of other metabolic pathways appear to emerge later in bacterial diversification.

A number of questions still remain standing. Within the bacterial tree, the positions of some small phyla, such as the Fusobacteriota, are still poorly resolved. Better sampling of these lineages, and particularly the discovery of more basal members, may bring needed clarity to this issue. This additional resolution to the tree may also help to resolve more precisely the position of the root. The root of the tree of life is another outstanding question. Although we carried out an analysis to root the tree of life using ALE in Chapter 2, this was only a preliminary analysis. Far more substantial

work would be needed to better answer this question, including increased taxon sampling, increased number of orthologues, and extensive topology testing in order to produce a reliable species tree. Furthermore, although we carried out a relative dating analyses, our results, particularly regarding how they relate to important events in the biogeochemical history of the Earth, could be strengthened by carrying our molecular clock analyses.

With regards to the ALE analyses specifically, a number of future developments could improve upon the analyses presented here. Currently, we employ a two step process, whereby unrooted gene trees are inferred using species-unaware models, before using the gene tree topologies in the reconciliation analyses. However, ideally we would jointly model DTL events, the rooted gene trees and a rooted species tree, but such analyses are not currently tractable. The taxon samplings used in this thesis are reduced due to computational constraints, and thus the verticality analyses and the ancestral gene inferences are almost certainly affected by the sparse taxonomic sampling. The increase in computing power and the writing of more efficient programs for phylogenetic inference may allow the use of a much larger taxon sampling size, which would hopefully increase the accuracy of the results. Additionally, when reconciling the COG families with our species tree, we used an origination prior for each COG category. Although this greatly improved on previous analyses using a flat prior, the COG categories are broad and contain many genes which may be evolving in very different ways. We would therefore ideally use a different origination prior for every family, although this would be computationally expensive. Both the expanded taxon sampling and the individual origination priors for each gene family would likely increase the number of genes recovered at each node and lead to better reconstructions. Furthermore, alternative pipelines for generating gene families could be explored. In this thesis, we used a Markov Clustering (MCL) (van Dongen, 2000) algorithm to generate the gene families, but alternatives, such as HiFix (Miele *et al.*, 2012) may also be explored and results compared. In a similar vein, repeating these analyses using other species aware methods, as such GeneRax (Morel *et al.*, 2020) may also prove useful. Further simulations would also be useful in determining the extent to which horizontal transfer may affect the ability to infer th tree. Aditionally, the further development of the ALE method to account for finer grained evolutionary process, including accounting for orthologus replacement, pseudogenes,

recombination, and transfer beteween closely related strains and lineages, will improve the accuracy of results.

Despite the caveats discussed, these methods are a clear improvement on methods used previously. We use model based approaches were caveats are explicit, and where there are clear and realistic areas for development. We believe they represent important contributions to the fields of evolutionary microbiology and phylogenetics. Particularly, we have further demonstrated the importance of using holistic, multi-pronged approaches to answering such illusive questions surrounding the early evolution of life. Hopefully, with improved models and better taxonomic sampling, and increased computational power allowing for larger datasets, we will be able to improve our results and continue to refine our picture of the evolutionary history of life.

References

Abby, S. S. *et al.* (2012) ‘Lateral gene transfer as a support for the tree of life’, *Proceedings of the National Academy of Sciences of the United States of America*, 109(13), pp. 4962–4967.

Adam, P. S., Borrel, G. and Gribaldo, S. (2018) ‘Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes’, *Proceedings of the National Academy of Sciences of the United States of America*, 115(6), pp. E1166–E1173.

Akanuma, S. *et al.* (2013) ‘Experimental evidence for the thermophilicity of ancestral life’, *Proceedings of the National Academy of Sciences of the United States of America*, 110(27), pp. 11067–11072.

Akanuma, S., Yokobori, S.-I. and Yamagishi, A. (2013) ‘Comparative Genomics of Thermophilic Bacteria and Archaea’, in Satyanarayana, T., Littlechild, J., and Kawarabayasi, Y. (eds) *Thermophilic Microbes in Environmental and Industrial Biotechnology: Biotechnology of Thermophiles*. Dordrecht: Springer Netherlands, pp. 331–349.

Alvarez-Ponce, D. *et al.* (2013) ‘Gene similarity networks provide tools for understanding eukaryote origins and evolution’, *Proceedings of the National Academy of Sciences of the United States of America*, 110(17), pp. E1594–603.

Anantharaman, K. *et al.* (2016) ‘Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system’, *Nature communications*, 7, p. 13219.

Antunes, L. C. *et al.* (2016) ‘Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the Firmicutes’, *eLife*, 5. doi: 10.7554/eLife.14589.

Aramaki, T. *et al.* (2020) ‘KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold’, *Bioinformatics* , 36(7), pp. 2251–2252.

Arndt, N. T. and Nisbet, E. G. (2012) ‘Processes on the Young Earth and the Habitats

of Early Life', *Annual review of earth and planetary sciences*. Annual Reviews, 40(1), pp. 521–549.

Avery Oswald, T., Colin, M. and MacLeod, M. M. C. (1944) 'Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus type III', *J. of Experimental Medicine*, 79(2), pp. 137–158.

Bansal, M. S. *et al.* (2018) 'RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss', *Bioinformatics* , 34(18), pp. 3214–3216.

Bansal, M. S., Alm, E. J. and Kellis, M. (2012) 'Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss', *Bioinformatics* . academic.oup.com, 28(12), pp. i283–91.

Barion, S. *et al.* (2007) 'The first lines of divergence in the Bacteria domain were the hyperthermophilic organisms, the Thermotogales and the Aquificales, and not the mesophilic Planctomycetales', *Bio Systems*, 87(1), pp. 13–19.

Barka, E. A. *et al.* (2016) 'Taxonomy, Physiology, and Natural Products of Actinobacteria', *Microbiology and molecular biology reviews: MMBR*, 80(1), pp. 1–43.

Barrangou, R. *et al.* (2007) 'CRISPR provides acquired resistance against viruses in prokaryotes', *Science*, 315(5819), pp. 1709–1712.

Battistuzzi, F. U., Feijao, A. and Hedges, S. B. (2004) 'A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land', *BMC evolutionary biology*, 4, p. 44.

Battistuzzi, F. U. and Hedges, S. B. (2009) 'A major clade of prokaryotes with ancient adaptations to life on land', *Molecular biology and evolution*, 26(2), pp. 335–343.

Baum, D. A. and Baum, B. (2014) 'An inside-out origin for the eukaryotic cell', *BMC Biology*. doi: 10.1186/s12915-014-0076-2.

Beam, J. P. *et al.* (2020) 'Ancestral absence of electron transport chains in Patescibacteria and DPANN', *bioRxiv*. doi: 10.1101/2020.04.07.029462.

Bell, S. D. and Jackson, S. P. (1998) ‘Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features’, *Trends in microbiology*, 6(6), pp. 222–228.

Bergsten, J. (2005) ‘A review of long-branch attraction’, *Cladistics*, pp. 163–193. doi: 10.1111/j.1096-0031.2005.00059.x.

Bern, M. and Goldberg, D. (2005) ‘Automatic selection of representative proteins for bacterial phylogeny’, *BMC evolutionary biology*, 5, p. 34.

Betts, H. C. *et al.* (2018) ‘Integrated genomic and fossil evidence illuminates life’s early evolution and eukaryote origin’, *Nature ecology & evolution*, 2(10), pp. 1556–1562.

Biegel, E. *et al.* (2011) ‘Biochemistry, evolution and physiological function of the Rnf complex, a novel ion-motive electron transport complex in prokaryotes’, *Cellular and molecular life sciences: CMLS*, 68(4), pp. 613–634.

Biegel, E. and Muller, V. (2010) ‘Bacterial Na -translocating ferredoxin:NAD oxidoreductase’, *Proceedings of the National Academy of Sciences*, pp. 18138–18142. doi: 10.1073/pnas.1010318107.

Bladen, H. A. and Mergenhagen, S. E. (1964) ‘ULTRASTRUCTURE OF VEILLONELLA AND MORPHOLOGICAL CORRELATION OF AN OUTER MEMBRANE WITH PARTICLES ASSOCIATED WITH ENDOTOXIC ACTIVITY’, *Journal of bacteriology*, 88, pp. 1482–1492.

Bocchetta, M. *et al.* (2000) ‘Phylogenetic depth of the bacterial genera Aquifex and Thermotoga inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences’, *Journal of molecular evolution*, 50(4), pp. 366–380.

Boucher, Y., Kamekura, M. and Doolittle, W. F. (2004) ‘Origins and evolution of isoprenoid lipid biosynthesis in archaea’, *Molecular microbiology*, 52(2), pp. 515–527.

Boussau, B., Guéguen, L. and Gouy, M. (2008) ‘Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria’, *BMC evolutionary biology*, 8, p. 272.

Braakman, R. and Smith, E. (2012) ‘The emergence and early evolution of biological carbon-fixation’, *PLoS computational biology*, 8(4), p. e1002455.

Brennan, C. A. and Garrett, W. S. (2019) ‘Fusobacterium nucleatum - symbiont, opportunist and oncobacterium’, *Nature reviews. Microbiology*, 17(3), pp. 156–166.

Brochier, C. and Philippe, H. (2002) ‘Phylogeny: a non-hyperthermophilic ancestor for bacteria’, *Nature*, 417(6886), p. 244.

Brosnan, M. E. and Brosnan, J. T. (2016) ‘Formate: The Neglected Member of One-Carbon Metabolism’, *Annual review of nutrition*, 36, pp. 369–388.

Brown, C. T. *et al.* (2015) ‘Unusual biology across a group comprising more than 15% of domain Bacteria’, *Nature*, 523(7559), pp. 208–211.

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

Buchfink, B., Xie, C. and Huson, D. H. (2015) ‘Fast and sensitive protein alignment using DIAMOND’, *Nature methods*, 12(1), pp. 59–60.

Buckel, W. and Thauer, R. K. (2013) ‘Energy conservation via electron bifurcating ferredoxin reduction and proton/Na+ translocating ferredoxin oxidation’, *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1827(2), pp. 94–113.

Burstein, D. *et al.* (2016) ‘Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems’, *Nature communications*, 7, p. 10613.

Burstein, D. *et al.* (2017) ‘New CRISPR–Cas systems from uncultivated microbes’, *Nature*, pp. 237–241. doi: 10.1038/nature21059.

Butterfield, N. J. (2015) ‘Early evolution of the Eukaryota’, *Palaeontology*. Wiley Online Library. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/pala.12139.

Butterfield, N. J. (2015) ‘Early evolution of the Eukaryota’, *Palaeontology*, 58(1), pp. 5–17.

Caforio, A. *et al.* (2018) ‘Converting Escherichia coli into an archaebacterium with a hybrid heterochiral membrane’, *Proceedings of the National Academy of Sciences of the United States of America*, 115(14), pp. 3704–3709.

Calvignac-Spencer, S. *et al.* (2014) ‘Clock Rooting Further Demonstrates that Guinea 2014 EBOV is a Member of the Zaïre Lineage’, *PLoS currents*, 6. doi: 10.1371/currents.outbreaks.c0e035c86d721668a6ad7353f7f6fe86.

Carbone, V. *et al.* (2015) ‘Structure and Evolution of the Archaeal Lipid Synthesis Enzyme sn-Glycerol-1-phosphate Dehydrogenase’, *The Journal of biological chemistry*, 290(35), pp. 21690–21704.

Castelle, C. J. *et al.* (2015) ‘Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling’, *Current biology: CB*. Elsevier Ltd, 25(6), pp. 690–701.

Castelle, C. J. *et al.* (2018) ‘Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations’, *Nature reviews. Microbiology*. doi: 10.1038/s41579-018-0076-2.

Castelle, C. J. and Banfield, J. F. (2018) ‘Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life’, *Cell*. Elsevier, 172(6), pp. 1181–1197.

Cavalier-Smith, T. (2002) ‘The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification’, *International journal of systematic and evolutionary microbiology*, 52(Pt 1), pp. 7–76.

Cavalier-Smith, T. (2006) ‘Rooting the tree of life by transition analyses’, *Biology direct*, 1, p. 19.

Chauve, C. *et al.* (2017) ‘MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers’. Available at: https://hal.archives-ouvertes.fr/hal-01532738/.

Chauve, C. *et al.* (2017) ‘MaxTiC: Fast Ranking Of A Phylogenetic Tree By Maximum Time Consistency With Lateral Gene Transfers’, *bioRxiv*. doi: 10.1101/127548.

Chernikova, D. *et al.* (2011) ‘A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes’, *Biology direct*, 6, p. 26.

Cody, G. D. *et al.* (2001) ‘Geochemical roots of autotrophic carbon fixation:

hydrothermal experiments in the system citric acid, H2O-(±FeS)−(±NiS)', *Geochimica et cosmochimica acta*, 65(20), pp. 3557–3576.

Coleman, G. A., Pancost, R. D. and Williams, T. A. (2019) 'Investigating the Origins of Membrane Phospholipid Biosynthesis Genes Using Outgroup-Free Rooting', *Genome biology and evolution*, 11(3), pp. 883–898.

Comte, N. *et al.* (2019) 'Treerecs: an integrated phylogenetic tool, from sequences to reconciliations', *bioRxiv*. doi: 10.1101/782946.

Creevey, C. J. *et al.* (2004) 'Does a tree-like phylogeny only exist at the tips in the prokaryotes?', *Proceedings. Biological sciences / The Royal Society*, 271(1557), pp. 2551–2558.

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

Csurös, M. (2010) 'Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood', *Bioinformatics* , 26(15), pp. 1910–1912.

Dagan, T. *et al.* (2010) 'Genome networks root the tree of life between prokaryotic domains', *Genome biology and evolution*, 2, pp. 379–392.

Dagan, T. and Martin, W. (2006) 'The tree of one percent', *Genome biology*, 7(10), p. 118.

Dagan, T. and Martin, W. (2007) 'Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution', *Proceedings of the National Academy of Sciences of the United States of America*, 104(3), pp. 870–875.

Daiyasu, H. *et al.* (2002) 'Analysis of membrane stereochemistry with homology modeling of sn-glycerol-1-phosphate dehydrogenase', *Protein engineering*, 15(12), pp. 987–995.

Damsté, J. S. S., Sinninghe Damsté, J. S., *et al.* (2007) 'Structural characterization of diabolic acid-based tetraester, tetraether and mixed ether/ester, membrane-spanning lipids of bacteria from the order Thermotogales', *Archives of Microbiology*, pp. 629–

641. doi: 10.1007/s00203-007-0284-z.

Damsté, J. S. S., Rijpstra, W. I. C., *et al.* (2007) 'Structural characterization of diabolic acid-based tetraester, tetraether and mixed ether/ester, membrane-spanning lipids of bacteria from the order Thermotogales', *Archives of microbiology*, 188(6), pp. 629–641.

Danovaro, R. *et al.* (2016) 'Macroecological drivers of archaea and bacteria in benthic deep-sea ecosystems', *Science advances*, 2(4), p. e1500961.

David, L. A. and Alm, E. J. (2011) 'Rapid evolutionary innovation during an Archaean genetic expansion', *Nature*, 469(7328), pp. 93–96.

Davín, A. A. *et al.* (2018) 'Gene transfers can date the tree of life', *Nature ecology & evolution*, 2(5), pp. 904–909.

Davis, J. J. *et al.* (2013) 'Genomes of the class Erysipelotrichia clarify the firmicute origin of the class Mollicutes', *International journal of systematic and evolutionary microbiology*, 63(Pt 7), pp. 2727–2741.

Decker, K., Jungermann, K. and Thauer, R. K. (1970) 'Energy Production in Anaerobic Organisms', *Angewandte Chemie International Edition in English*, pp. 138–158. doi: 10.1002/anie.197001381.

Dombrowski, N. *et al.* (2020) 'Undinarchaeota illuminate the evolution of DPANN archaea', *bioRxiv*. doi: 10.1101/2020.03.05.976373.

van Dongen, S. M. (2000) *Graph Clustering by Flow Simulation*.

Doolittle, W. F. (1999) 'Phylogenetic classification and the universal tree', *Science*, 284(5423), pp. 2124–2129.

Doolittle, W. F. and Bapteste, E. (2007) 'Pattern pluralism and the Tree of Life hypothesis', *Proceedings of the National Academy of Sciences of the United States of America*, 104(7), pp. 2043–2049.

Drummond, A. J. *et al.* (2012) 'Bayesian phylogenetics with BEAUti and the BEAST 1.7', *Molecular biology and evolution*, 29(8), pp. 1969–1973.

Drummond, A. J. and Bouckaert, R. R. (2015) *Bayesian Evolutionary Analysis with BEAST*. Cambridge University Press.

Drummond, A. J. and Rambaut, A. (2007) 'BEAST: Bayesian evolutionary analysis by sampling trees', *BMC Evolutionary Biology*, p. 214. doi: 10.1186/1471-2148-7-214.

Embley, T. M. and Martin, W. (2006) 'Eukaryotic evolution, changes and challenges', *Nature*, 440(7084), pp. 623–630.

Eme, L. *et al.* (2014a) 'On the age of eukaryotes: evaluating evidence from fossils and molecular clocks', *Cold Spring Harbor perspectives in biology*, 6(8). doi: 10.1101/cshperspect.a016139.

Eme, L. *et al.* (2014b) 'On the age of eukaryotes: evaluating evidence from fossils and molecular clocks', *Cold Spring Harbor perspectives in biology*, 007.

Eme, L. *et al.* (2018) 'Archaea and the origin of eukaryotes', *Nature reviews. Microbiology*. nature.com, 16(2), p. 120.

Emms, D. M. and Kelly, S. (2017) 'STRIDE: Species Tree Root Inference from Gene Duplication Events', *Molecular biology and evolution*, 34(12), pp. 3267–3278.

Erb, T. J. and Zarzycki, J. (2018) 'A short history of RubisCO: the rise and fall (?) of Nature's predominant CO2 fixing enzyme', *Current opinion in biotechnology*, 49, pp. 100–107.

Errington, J. (2013) 'L-form bacteria, cell walls and the origins of life', *Open biology*, 3(1), p. 120143.

Escalante-Semerena, J. C., Rinehart, K. L., Jr and Wolfe, R. S. (1984) 'Tetrahydromethanopterin, a carbon carrier in methanogenesis', *The Journal of biological chemistry*, 259(15), pp. 9447–9455.

Etiope, G., Schoell, M. and Hosgörmez, H. (2011) 'Abiotic methane flux from the Chimaera seep and Tekirova ophiolites (Turkey): Understanding gas exhalation from low temperature serpentinization and implications for Mars', *Earth and planetary science letters*, 310(1), pp. 96–104.

Fan, Q. *et al.* (1995) ‘Stability against temperature and external agents of vesicles composed of archaeal bolaform lipids and egg PC’, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1240(1), pp. 83–88.

Felsenstein, J. (1978) ‘Cases in which Parsimony or Compatibility Methods will be Positively Misleading’, *Systematic biology*. Oxford Academic, 27(4), pp. 401–410.

Ferry, J. G. and House, C. H. (2006) ‘The stepwise evolution of early life driven by energy conservation’, *Molecular biology and evolution*, 23(6), pp. 1286–1292.

Finn, R. D., Clements, J. and Eddy, S. R. (2011) ‘HMMER web server: interactive sequence similarity searching’, *Nucleic acids research*, 39(Web Server issue), pp. W29–37.

Foster, P. G. (2004) ‘Modeling compositional heterogeneity’, *Systematic biology*, 53(3), pp. 485–495.

Foster, P. G. and Hickey, D. A. (1999) ‘Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions’, *Journal of molecular evolution*, 48(3), pp. 284–290.

Franklin, R. E. and Gosling, R. G. (1953a) ‘Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate’, *Nature*. nature.com, 172(4369), pp. 156–157.

Franklin, R. E. and Gosling, R. G. (1953b) ‘Molecular configuration in sodium thymonucleate’, *Nature*. nature.com, 171(4356), pp. 740–741.

Fuchs, G. (2011) ‘Alternative pathways of carbon dioxide fixation: insights into the early evolution of life?’, *Annual review of microbiology*, 65, pp. 631–658.

Garcia-Vallvé, S., Romeu, A. and Palau, J. (2000) ‘Horizontal gene transfer in bacterial and archaeal complete genomes’, *Genome research*, 10(11), pp. 1719–1725.

Gascuel, O. (2005) *Mathematics of Evolution and Phylogeny*. OUP Oxford.

Gattinger, A., Schloter, M. and Munch, J. C. (2002) ‘Phospholipid etherlipid and phospholipid fatty acid fingerprints in selected euryarchaeotal monocultures for taxonomic profiling’, *FEMS microbiology letters*, 213(1), pp. 133–139.

Naik, Gauri A., *et al.* (1994) 'Transfer of broad-host-range antibiotic resistance plasmids in soil microcosms.' *Current Microbiology* 28.4, pp. 209-215.

Gogarten, J. P. *et al.* (1989) 'Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes', *Proceedings of the National Academy of Sciences of the United States of America*, 86(17), pp. 6661–6665.

Gogarten, J. P. *et al.* (1989) 'Molecular Evolution of H -ATPases. I. Methanococcus and Sulfolobus are Monophyletic with Respect to Eukaryotes and Eubacteria', *Zeitschrift für Naturforschung C*, pp. 641–650. doi: 10.1515/znc-1989-7-816.

Goldfine, H. (2010) 'The appearance, disappearance and reappearance of plasmalogens in evolution', *Progress in lipid research*, 49(4), pp. 493–498.

Gould, S. B., Garg, S. G. and Martin, W. F. (2016) 'Bacterial Vesicle Secretion and the Evolutionary Origin of the Eukaryotic Endomembrane System', *Trends in microbiology*, 24(7), pp. 525–534.

Gouy, R., Baurain, D. and Philippe, H. (2015) 'Rooting the tree of life: the phylogenetic jury is still out', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1678), p. 20140329.

Graybeal, A. (1998) 'Is it better to add taxa or characters to a difficult phylogenetic problem?', *Systematic biology*, 47(1), pp. 9–17.

Greening, C. *et al.* (2016) 'Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival', *The ISME journal*, 10(3), pp. 761–777.

Groussin, M. *et al.* (2015) 'Toward more accurate ancestral protein genotype-phenotype reconstructions with the use of species tree-aware gene trees', *Molecular biology and evolution*, 32(1), pp. 13–22.

Guldan, H. *et al.* (2011) 'Functional assignment of an enzyme that catalyzes the synthesis of an Archaea-type ether lipid in bacteria', *Angewandte Chemie, International Edition*. Wiley Online Library, 50(35), pp. 8188–8191.

Guldan, H., Sterner, R. and Babinger, P. (2008) 'Identification and Characterization of

a Bacterial Glycerol-1-phosphate Dehydrogenase: Ni2+-Dependent AraM from Bacillus subtilis', *Biochemistry*. American Chemical Society, 47(28), pp. 7376–7384.

Gupta, R. S. (2004) 'The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes', *Critical reviews in microbiology*, 30(2), pp. 123–143.

Gupta, R. S. (2011) 'Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes', *Antonie van Leeuwenhoek*, 100(2), pp. 171–182.

Hedtke, S. M., Townsend, T. M. and Hillis, D. M. (2006) 'Resolution of phylogenetic conflict in large data sets by increased taxon sampling', *Systematic biology*, 55(3), pp. 522–529.

Hemmi, H. *et al.* (2004) '(S)-2, 3-Di-O-geranylgeranylglyceryl phosphate synthase from the thermoacidophilic archaeon sulfolobus solfataricus molecular cloning and characterization of a membrane-intrinsic prenyltransferase involved in the biosynthesis of archaeal ether-linked membrane lipids', *The Journal of biological chemistry*. ASBMB, 279(48), pp. 50197–50203.

Herrmann, G. *et al.* (2008) 'Energy conservation via electron-transferring flavoprotein in anaerobic bacteria', *Journal of bacteriology*, 190(3), pp. 784–791.

Hess, V., Schuchmann, K. and Müller, V. (2013) 'The ferredoxin:NAD+ oxidoreductase (Rnf) from the acetogen Acetobacterium woodii requires Na+ and is reversibly coupled to the membrane potential', *The Journal of biological chemistry*, 288(44), pp. 31496–31502.

Heuer, H. and Smalla, K. (2007) 'Horizontal gene transfer between bacteria', *Environmental biosafety research*, 6(1-2), pp. 3–13.

Hoff, M. *et al.* (2016) 'Does the choice of nucleotide substitution models matter topologically?', *BMC bioinformatics*, 17, p. 143.

Huerta-Cepas, J. *et al.* (2016) 'eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences',

Nucleic acids research, 44(D1), pp. D286–93.

Huerta-Cepas, J. *et al.* (2017) 'Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper', *Molecular biology and evolution*, 34(8), pp. 2115–2122.

Huerta-Cepas, J. *et al.* (2019) 'eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses', *Nucleic acids research*, 47(D1), pp. D309–D314.

Hug, L. A. *et al.* (2013) 'Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling', *Microbiome*, 1(1), p. 22.

Hug, L. A. *et al.* (2016) 'A new view of the tree of life', *Nature microbiology*, 1(April), p. 16048.

Iwabe, N. *et al.* (1989) 'Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes', *Proceedings of the National Academy of Sciences of the United States of America*, 86(December), pp. 9355–9359.

Jacox, E. *et al.* (2016) 'ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony', *Bioinformatics* , 32(13), pp. 2056–2058.

Jain, R., Rivera, M. C. and Lake, J. a. (1999) 'Horizontal gene transfer among genomes: the complexity hypothesis', *Proceedings of the National Academy of Sciences of the United States of America*, 96(7), pp. 3801–3806.

Ji, M. *et al.* (2017) 'Atmospheric trace gases support primary production in Antarctic desert surface soil', *Nature*, 552(7685), pp. 400–403.

Jones, P. *et al.* (2014) 'InterProScan 5: genome-scale protein function classification', *Bioinformatics* , 30(9), pp. 1236–1240.

Jones, W. J., Donnelly, M. I. and Wolfe, R. S. (1985) 'Evidence of a common pathway of carbon dioxide reduction to methane in methanogens', *Journal of bacteriology*, 163(1), pp. 126–131.

Jukes, T. H., Cantor, C. R. and Others (1969) ‘Evolution of protein molecules’, *Mammalian protein metabolism*, 3, pp. 21–132.

Kanao, T. *et al.* (2001) ‘ATP-citrate lyase from the green sulfur bacterium Chlorobium limicola is a heteromeric enzyme composed of two distinct gene products’, *European journal of biochemistry / FEBS*, 268(6), pp. 1670–1678.

Kandler, O. (1995) ‘Cell wall biochemistry in Archaea and its phylogenetic implications’, *Journal of biological physics*. Springer, 20(1-4), pp. 165–169.

Kates, M. (1978) ‘The phytanyl ether-linked polar lipids and isoprenoid neutral lipids of extremely halophilic bacteria’, *Progress in the chemistry of fats and other lipids*, 15(4), pp. 301–342.

Katoh, K. *et al.* (2002) ‘MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform’, *Nucleic acids research*, 30(14), pp. 3059–3066.

Katz, L. a. *et al.* (2012) ‘Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life’, *Systematic biology*, 61(4), pp. 653–660.

Kaur, G. *et al.* (2015) ‘Temperature and pH control on lipid composition of silica sinters from diverse hot springs in the Taupo Volcanic Zone, New Zealand’, *Extremophiles: life under extreme conditions*, 19(2), pp. 327–344.

Kazlauskiene, M. *et al.* (2017) ‘A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems’, *Science*, 357(6351), pp. 605–609.

Kellner, S. *et al.* (2018) ‘Genome size evolution in the Archaea’, *Emerging Topics in Life Sciences*. Portland Press Ltd., 2(4), pp. 595–605.

Kelman, L. M. and Kelman, Z. (2014) ‘Archaeal DNA replication’, *Annual review of genetics*, 48, pp. 71–97.

Kishino, H., Thorne, J. L. and Bruno, W. J. (2001) ‘Performance of a divergence time estimation method under a probabilistic model of rate evolution’, *Molecular biology*. academic.oup.com. Available at: https://academic.oup.com/mbe/article-abstract/18/3/352/1073229.

Klein, M. *et al.* (2001) ‘Multiple lateral transfers of dissimilatory sulfite reductase genes between major lineages of sulfate-reducing prokaryotes’, *Journal of bacteriology*, 183(20), pp. 6028–6035.

Klingl, A. (2014) ‘S-layer and cytoplasmic membrane - exceptions from the typical archaeal cell wall with a focus on double membranes’, *Frontiers in microbiology*, 5, p. 624.

Knoll, A. H. (2014) ‘Paleobiological perspectives on early eukaryotic evolution’, *Cold Spring Harbor perspectives in biology*, 6(1). doi: 10.1101/cshperspect.a016121.

Knoll, A. H. and Nowak, M. A. (2017) ‘The timetable of evolution’, *Science advances*, 3(5), p. e1603076.

Koga, Y. *et al.* (1998) ‘Did archaeal and bacterial cells arise independently from noncellular Precursors? A hypothesis stating that the advent of membrane phospholipid with enantiomeric glycerophosphate backbones caused the separation of the two lines of descent’, *Journal of molecular evolution*, 47(5), p. 631.

Koga, Y. (2011) ‘Early evolution of membrane lipids: how did the lipid divide occur?’, *Journal of molecular evolution*, 72(3), pp. 274–282.

Koga, Y. (2012) ‘Thermal adaptation of the archaeal and bacterial lipid membranes’, *Archaea* , 2012, p. 789652.

Koga, Y. (2014) ‘From promiscuity to the lipid divide: on the evolution of distinct membranes in Archaea and Bacteria’, *Journal of molecular evolution*, 78(3-4), pp. 234–242.

Koonin, E. V. (2014) ‘The origins of cellular life’, *Antonie van Leeuwenhoek*, 106(1), pp. 27–41.

Koonin, E. V. and Makarova, K. S. (2019) ‘Origins and evolution of CRISPR-Cas systems’, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 374(1772), p. 20180087.

Koonin, E. V., Makarova, K. S. and Aravind, L. (2002) *Horizontal Gene Transfer in Prokaryotes: Quantification and Classification*. National Center for Biotechnology

Information (US).

Koonin, E. V., Wolf, Y. I. and Puigbò, P. (2009) 'The phylogenetic forest and the quest for the elusive tree of life', *Cold Spring Harbor symposia on quantitative biology*, 74, pp. 205–213.

Krupovic, M., Dolja, V. V. and Koonin, E. V. (2019) 'Origin of viruses: primordial replicators recruiting capsids from hosts', *Nature reviews. Microbiology*, 17(7), pp. 449–458.

Kurland, C. G., Collins, L. J. and Penny, D. (2006) 'Genomics and the irreducible nature of eukaryote cells', *Science*, 312(5776), pp. 1011–1014.

Lafond, M., Swenson, K. M. and El-Mabrouk, N. (2012) 'An Optimal Reconciliation Algorithm for Gene Trees with Polytomies', in *Algorithms in Bioinformatics*. Springer Berlin Heidelberg, pp. 106–122.

Lake, J. A. (2009) 'Evidence for an early prokaryotic endosymbiosis', *Nature*, 460(7258), pp. 967–971.

Lake, J. A. *et al.* (2009) 'Genome beginnings: rooting the tree of life', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1527), pp. 2177–2185.

Landan, G. and Graur, D. (2007) 'Heads or tails: a simple reliability check for multiple sequence alignments', *Molecular biology and evolution*, 24(6), pp. 1380–1383.

Lane, N., Allen, J. F. and Martin, W. (2010) 'How did LUCA make a living? Chemiosmosis in the origin of life', *BioEssays: news and reviews in molecular, cellular and developmental biology*, 32(4), pp. 271–280.

Lane, N. and Martin, W. F. (2012) 'The origin of membrane bioenergetics', *Cell*, 151(7), pp. 1406–1416.

Lang, S. Q. *et al.* (2010) 'Elevated concentrations of formate, acetate and dissolved organic carbon found at the Lost City hydrothermal field', *Geochimica et cosmochimica acta*, 74(3), pp. 941–952.

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

Lartillot, N., Brinkmann, H. and Philippe, H. (2007b) ‘Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model’, *BMC Evolutionary Biology*, p. S4. doi: 10.1186/1471-2148-7-s1-s4.

Lartillot, N. and Philippe, H. (2004) ‘A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process’, *Molecular Biology and Evolution*, pp. 1095–1109. doi: 10.1093/molbev/msh112.

Lechevalier, H. and Lechevalier, M. P. (1965) ‘Classification des actinomycètes aérobies basée sur leur morphologie et leur composition chimique’, in *ANNALES DE L INSTITUT PASTEUR*. MASSON EDITEUR 21 STREET CAMILLE DESMOULINS, ISSY, 92789 MOULINEAUX CEDEX 9 …, p. 662–+.

Le, S. Q. and Gascuel, O. (2008) ‘An improved general amino acid replacement matrix’, *Molecular biology and evolution*, 25(7), pp. 1307–1320.

Le, S. Q., Lartillot, N. and Gascuel, O. (2008) ‘Phylogenetic mixture models for proteins’, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1512), pp. 3965–3976.

Li, F. *et al.* (2008) ‘Coupled ferredoxin and crotonyl coenzyme A (CoA) reduction with NADH catalyzed by the butyryl-CoA dehydrogenase/Etf complex from Clostridium kluyveri’, *Journal of bacteriology*. Am Soc Microbiol, 190(3), pp. 843–850.

Liu, R. and Ochman, H. (2007a) ‘Origins of flagellar gene operons and secondary flagellar systems’, *Journal of bacteriology*, 189(19), pp. 7098–7104.

Liu, R. and Ochman, H. (2007b) ‘Stepwise formation of the bacterial flagellar system’, *Proceedings of the National Academy of Sciences of the United States of America*, 104(17), pp. 7116–7121.

Liu, Y., Beer, L. L. and Whitman, W. B. (2012) ‘Methanogens: a window into ancient sulfur metabolism’, *Trends in microbiology*, 20(5), pp. 251–258.

Lombard, J., López-García, P. and Moreira, D. (2012a) ‘Phylogenomic investigation of phospholipid synthesis in archaea’, *Archaea* , 2012, p. 630910.

Lombard, J., López-García, P. and Moreira, D. (2012b) ‘The early evolution of lipid membranes and the three domains of life’, *Nature reviews. Microbiology*, 10(7), pp. 507–515.

Lombard, J. and Moreira, D. (2011) ‘Origins and early evolution of the mevalonate pathway of isoprenoid biosynthesis in the three domains of life’, *Molecular biology and evolution*, 28(1), pp. 87–99.

López-García, P., Eme, L. and Moreira, D. (2017) ‘Symbiosis in eukaryotic evolution’, *Journal of theoretical biology*. Elsevier, 434, pp. 20–33.

López-García, P. and Moreira, D. (2006) ‘Selective forces for the origin of the eukaryotic nucleus’, *BioEssays*, pp. 525–533. doi: 10.1002/bies.20413.

Maden, B. E. (2000) ‘Tetrahydrofolate and tetrahydromethanopterin compared: functionally distinct carriers in C1 metabolism’, *Biochemical Journal*, 350 Pt 3, pp. 609–629.

Makarova, K. S. *et al.* (2011) ‘Evolution and classification of the CRISPR-Cas systems’, *Nature reviews. Microbiology*, 9(6), pp. 467–477.

Makarova, K. S. *et al.* (2015) ‘An updated evolutionary classification of CRISPR-Cas systems’, *Nature reviews. Microbiology*, 13(11), pp. 722–736.

Marakushev, S. A. and Belonogova, O. V. (2011) ‘Emergence of the chemoautotrophic metabolism in hydrothermal environments and the origin of ancestral bacterial taxa’, *Doklady. Biochemistry and biophysics*, 439, pp. 161–166.

Marakushev, S. A. and Belonogova, O. V. (2013) ‘The origin of ancestral bacterial metabolism’, *Paleontological Journal*. Springer, 47(9), pp. 1001–1010.

Martin, W. F., Garg, S. and Zimorski, V. (2015) ‘Endosymbiotic theories for eukaryote origin’, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. royalsocietypublishing.org, 370(1678), p. 20140330.

Martin, W. and Russell, M. J. (2003) ‘On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells’, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 358(1429), pp. 59–83; discussion 83–5.

Martin, W. and Russell, M. J. (2007) ‘On the origin of biochemistry at an alkaline hydrothermal vent’, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 362(1486), pp. 1887–1925.

McCollom, T. M. and Amend, J. P. (2005) ‘A thermodynamic assessment of energy requirements for biomass synthesis by chemolithoautotrophic micro-organisms in oxic and anoxic environments’, *Geobiology*. Wiley Online Library, 3(2), pp. 135–144.

Megrian, D. *et al.* (2020) ‘One or two membranes? Diderm Firmicutes challenge the Gram-positive/Gram-negative divide’, *Molecular microbiology*, 113(3), pp. 659–671.

Mehrshad, M. *et al.* (2018) ‘Hidden in plain sight-highly abundant and diverse planktonic freshwater Chloroflexi’, *Microbiome*, 6(1), p. 176.

Melville, S. and Craig, L. (2013) ‘Type IV pili in Gram-positive bacteria’, *Microbiology and molecular biology reviews: MMBR*, 77(3), pp. 323–341.

Mendel, G. (1866) ‘Versuche über Pflanzenhybriden Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr, 1865’, *Abhand-lungen*, 3, p. 47.

Miele, V. *et al.* (2012) ‘High-quality sequence clustering guided by network topology and multiple alignment likelihood’, *Bioinformatics* . academic.oup.com, 28(8), pp. 1078–1085.

Miescher, F. (1869) ‘Letter I; to Wilhelm His; Tübingen, February 26th, 1869’, *Die Histochemischen und Physiologischen Arbeiten von Friedrich Miescher--Aus dem sissenschaft--lichen Briefwechsel von F. Miescher*, 1, pp. 33–38.

Miescher-Rüsch, F. (1871) *Ueber die chemische Zusammensetzung der Eiterzellen*.

Miyata, M. *et al.* (2020) ‘Tree of motility - A proposed history of motility systems in the

tree of life', *Genes to cells: devoted to molecular & cellular mechanisms*, 25(1), pp. 6–21.

Morel, B. *et al.* (2019) 'GeneRax: A tool for species tree-aware maximum likelihood based gene tree inference under gene duplication, transfer, and loss', *BioRxiv*. biorxiv.org. Available at: https://www.biorxiv.org/content/10.1101/779066v1.abstract.

Morel, B. *et al.* (2020) 'GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss', *Molecular biology and evolution*, 37(9), pp. 2763–2774.

Mukherjee, S. *et al.* (2017) '1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life', *Nature biotechnology*, 35(7), pp. 676–683.

Müller, V., Chowdhury, N. P. and Basen, M. (2018) 'Electron Bifurcation: A Long-Hidden Energy-Coupling Mechanism', *Annual review of microbiology*, 72, pp. 331–353.

Naser-Khdour, S. *et al.* (2019) 'The Prevalence and Impact of Model Violations in Phylogenetic Analysis', *Genome biology and evolution*, 11(12), pp. 3341–3352.

Nelson-Sathi, S. *et al.* (2015) 'Origins of major archaeal clades correspond to gene acquisitions from bacteria', *Nature*, 517(7532), pp. 77–80.

Nemoto, N., Oshima, T. and Yamagishi, A. (2003) 'Purification and characterization of geranylgeranylglyceryl phosphate synthase from a thermoacidophilic archaeon, Thermoplasma acidophilum', *Journal of biochemistry*, 133(5), pp. 651–657.

Nguyen, L.-T. *et al.* (2015) 'IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Molecular biology and evolution*, 32(1), pp. 268–274.

Niewoehner, O. *et al.* (2017) 'Type III CRISPR–Cas systems produce cyclic oligoadenylate second messengers', *Nature*, pp. 543–548. doi: 10.1038/nature23467.

Nishihara, M. *et al.* (1999) 'sn-glycerol-1-phosphate-forming activities in Archaea: separation of archaeal phospholipid biosynthesis and glycerol catabolism by

glycerophosphate enantiomers', *Journal of bacteriology*, 181(4), pp. 1330–1333.

Nitschke, W. and Russell, M. J. (2013) 'Beating the acetyl coenzyme A-pathway to the origin of life', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1622), p. 20120258.

Norris, P. R. *et al.* (2011) 'Autotrophic, sulfur-oxidizing actinobacteria in acidic environments', *Extremophiles: life under extreme conditions*. Springer, 15(2), pp. 155–163.

Noutahi, E. *et al.* (2016) 'Efficient Gene Tree Correction Guided by Genome Evolution', *PloS one*. journals.plos.org, 11(8), p. e0159559.

Nuñez, J. K. *et al.* (2014) 'Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity', *Nature Structural & Molecular Biology*, pp. 528–534. doi: 10.1038/nsmb.2820.

Nunoura, T. *et al.* (2018) 'A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile', *Science*, 359(6375), pp. 559–563.

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

de Oliveira Martins, L. and Posada, D. (2017) 'Species Tree Estimation from Genome-Wide Data with guenomu', *Methods in molecular biology* , 1525, pp. 461–478.

Pancost, R. D. *et al.* (2001) 'Three series of non-isoprenoidal dialkyl glycerol diethers in cold-seep carbonate crusts', *Organic Geochemistry*, pp. 695–707. doi: 10.1016/s0146-6380(01)00015-8.

Parfrey, L. W. *et al.* (2011) 'Estimating the timing of early eukaryotic diversification with multigene molecular clocks', *Proceedings of the National Academy of Sciences of the United States of America*, 108(33), pp. 13624–13629.

Parks, D. H. *et al.* (2015) 'CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome research*, 25(7), pp. 1043–1055.

Parks, D. H. *et al.* (2017) ‘Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life’, *Nature microbiology*, 2(11), pp. 1533–1542.

Parks, D. H. *et al.* (2018) ‘A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life’, *Nature biotechnology*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. doi: 10.1038/nbt.4229.

Parsons, J. B. and Rock, C. O. (2013) ‘Bacterial lipids: metabolism and membrane homeostasis’, *Progress in lipid research*, 52(3), pp. 249–276.

Payandeh, J. *et al.* (2006) ‘The crystal structure of (S)-3-O-geranylgeranylglyceryl phosphate synthase reveals an ancient fold for an ancient enzyme’, *The Journal of biological chemistry*, 281(9), pp. 6070–6078.

Penny, D. (1976) ‘Criteria for optimising phylogenetic trees and the problem of determining the root of a tree’, *Journal of molecular evolution*. Springer. Available at: https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1007/BF01739097&casa_token=l-YG_6Vt4MUAAAAA:oRUdQBia7wQ4ufBDjQbpwqeCkzcMBjYiG7TXv60fkD0jd8quaickcPOkcTK-bOkqfFC4it58lJgSTtuHqQ.

Penny, D. *et al.* (2001) ‘Mathematical elegance with biochemical realism: the covarion model of molecular evolution’, *Journal of molecular evolution*, 53(6), pp. 711–723.

Peretó, J., López-García, P. and Moreira, D. (2004) ‘Ancestral lipid biosynthesis and early membrane evolution’, *Trends in biochemical sciences*, 29(9), pp. 469–477.

Peterhoff, D. *et al.* (2014) ‘A comprehensive analysis of the geranylgeranylglyceryl phosphate synthase enzyme family identifies novel members and reveals mechanisms of substrate specificity and quaternary structure organization’, *Molecular microbiology*. Wiley Online Library, 92(4), pp. 885–899.

Proskurowski, G. *et al.* (2008) ‘Abiogenic hydrocarbon production at lost city hydrothermal field’, *Science*, 319(5863), pp. 604–607.

Puigbò, P., Wolf, Y. I. and Koonin, E. V. (2010) ‘The tree and net components of

prokaryote evolution', *Genome biology and evolution*, 2, pp. 745–756.

Rabus, R., Hansen, T. A. and Widdel, F. (2006) 'Dissimilatory Sulfate-and Sulfur-Reducing Prokaryotes In M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Sheleifer, and E. Stackebrandt (ed.), The …'. Springer, NewYork.

Raymann, K., Brochier-Armanet, C. and Gribaldo, S. (2015) 'The two-domain tree of life is linked to a new root for the Archaea', *Proceedings of the National Academy of Sciences of the United States of America*, 112(21), pp. 6670–6675.

Reeve, J. N., Sandman, K. and Daniels, C. J. (1997) 'Archaeal histones, nucleosomes, and transcription initiation', *Cell*, 89(7), pp. 999–1002.

Ripplinger, J. and Sullivan, J. (2008) 'Does choice in model selection affect maximum likelihood analysis?', *Systematic biology*. academic.oup.com, 57(1), pp. 76–85.

Roger, A. J., Muñoz-Gómez, S. A. and Kamikawa, R. (2017) 'The Origin and Diversification of Mitochondria', *Current biology: CB*. Elsevier, 27(21), pp. R1177–R1192.

Romano, A. H. and Conway, T. (1996) 'Evolution of carbohydrate metabolic pathways', *Research in microbiology*, 147(6-7), pp. 448–455.

Roth, A. C. J., Gonnet, G. H. and Dessimoz, C. (2008) 'Algorithm of OMA for large-scale orthology inference', *BMC bioinformatics*, 9, p. 518.

Russell, M. J. *et al.* (1994) 'A hydrothermally precipitated catalytic iron sulphide membrane as a first step toward life', *Journal of molecular evolution*. Springer, 39(3), pp. 231–243.

Russell, M. J. and Hall, A. J. (1997) 'The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front', *Journal of the Geological Society*, 154(3), pp. 377–402.

Sánchez-Andrea, I *et al.* (2020) 'The reductive glycine pathway allows autotrophic growth of Desulfovibrio desulfuricans.' *Nat. Commun.* **11**, 5090 (2020)

Sánchez-Baracaldo, P. *et al.* (2017) 'Early photosynthetic eukaryotes inhabited low-

salinity habitats', *Proceedings of the National Academy of Sciences of the United States of America*, 114(37), pp. E7737–E7745.

Schirrmeister, B. E. *et al.* (2013) 'Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event', *Proceedings of the National Academy of Sciences of the United States of America*, 110(5), pp. 1791–1796.

Schouten, S. *et al.* (2000) 'Widespread occurrence of structurally diverse tetraether membrane lipids: evidence for the ubiquitous presence of low-temperature relatives of hyperthermophiles', *Proceedings of the National Academy of Sciences of the United States of America*, 97(26), pp. 14421–14426.

Schouten, S., Hopmans, E. C. and Sinninghe Damsté, J. S. (2013) 'The organic geochemistry of glycerol dialkyl glycerol tetraether lipids: A review', *Organic geochemistry*, 54, pp. 19–61.

Schouten, S., Wakeham, S. G. and Sinninghe Damsté, J. S. (2001) 'Evidence for anaerobic methane oxidation by archaea in euxinic waters of the Black Sea', *Organic Geochemistry*, pp. 1277–1281. doi: 10.1016/s0146-6380(01)00110-3.

Schrenk, M. O., Brazelton, W. J. and Lang, S. Q. (2013) 'Serpentinization, Carbon, and Deep Life', *Reviews in Mineralogy and Geochemistry*. GeoScienceWorld, 75(1), pp. 575–606.

Schuchmann, K. and Müller, V. (2014) 'Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria', *Nature reviews. Microbiology*, 12(12), pp. 809–821.

Sela, I., Wolf, Y. I. and Koonin, E. V. (2016) 'Theory of prokaryotic genome evolution', *Proceedings of the National Academy of Sciences of the United States of America*, 113(41), pp. 11399–11407.

Shen, Y., Buick, R. and Canfield, D. E. (2001) 'Isotopic evidence for microbial sulphate reduction in the early Archaean era', *Nature*, 410(6824), pp. 77–81.

Shimada, H. and Yamagishi, A. (2011) 'Stability of heterochiral hybrid membrane

made of bacterial sn-G3P lipids and archaeal sn-G1P lipids', *Biochemistry*, 50(19), pp. 4114–4120.

Shimodaira, H. (2002) 'An approximately unbiased test of phylogenetic tree selection', *Systematic biology*, 51(3), pp. 492–508.

Silhavy, T. J., Kahne, D. and Walker, S. (2010) 'The bacterial cell envelope', *Cold Spring Harbor perspectives in biology*, 2(5), p. a000414.

Sinninghe Damsté, J. S. *et al.* (2002) 'Linearly concatenated cyclobutane lipids form a dense bacterial membrane', *Nature*, 419(6908), pp. 708–712.

Skophammer, R. G. *et al.* (2007) 'Evidence for a gram-positive, eubacterial root of the tree of life', *Molecular biology and evolution*, 24(8), pp. 1761–1768.

Smith, E. and Morowitz, H. J. (2004) 'Universality in intermediary metabolism', *Proceedings of the National Academy of Sciences of the United States of America*, 101(36), pp. 13168–13173.

Sojo, V., Pomiankowski, A. and Lane, N. (2014) 'A bioenergetic basis for membrane divergence in archaea and bacteria', *PLoS biology*, 12(8), p. e1001926.

Sousa, F. L. *et al.* (2013) 'Early bioenergetic evolution', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1622), p. 20130088.

Sousa, F. L. and Martin, W. F. (2014) 'Biochemical fossils of the ancient transition from geoenergetics to bioenergetics in prokaryotic one carbon compound metabolism', *Biochimica et biophysica acta*, 1837(7), pp. 964–981.

Sousa, F. L., Nelson-Sathi, S. and Martin, W. F. (2016) 'One step beyond a ribosome: The ancient anaerobic core', *Biochimica et biophysica acta*, 1857(8), pp. 1027–1038.

Spang, A. *et al.* (2015) 'Complex archaea that bridge the gap between prokaryotes and eukaryotes', *Nature*, 521(7551), pp. 173–179.

Sparacino-Watkins, C., Stolz, J. F. and Basu, P. (2014) 'Nitrate and periplasmic nitrate reductases', *Chemical Society reviews*, 43(2), pp. 676–706.

Stover, P. J. (2009) 'One-carbon metabolism-genome interactions in folate-associated

pathologies', *The Journal of nutrition*, 139(12), pp. 2402–2405.

Susko, E. and Roger, A. J. (2007) 'On reduced amino acid alphabets for phylogenetic inference', *Molecular biology and evolution*, 24(9), pp. 2139–2150.

Sutcliffe, I. C. (2010) 'A phylum level perspective on bacterial cell envelope architecture', *Trends in microbiology*, 18(10), pp. 464–470.

Szöllosi, G. J. *et al.* (2012) 'Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations', *Proceedings of the National Academy of Sciences of the United States of America*, 109(43), pp. 17513–17518.

Szöllősi, G. J. *et al.* (2013) 'Efficient Exploration of the Space of Reconciled Gene Trees', *Systematic biology*. Oxford Academic, 62(6), pp. 901–912.

Szöllősi, G. J., Davín, A. A., *et al.* (2015) 'Genome-scale phylogenetic analysis finds extensive gene transfer among fungi', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. royalsocietypublishing.org, 370(1678), p. 20140335.

Szöllősi, G. J., Tannier, E., *et al.* (2015) 'The inference of gene trees with species trees', *Systematic biology*, 64(1), pp. e42–62.

Szöllősi, G. J., Boussau, B. and Abby, S. S. (2012) 'Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations', *Proceedings of the*. National Acad Sciences. Available at: https://www.pnas.org/content/109/43/17513.short.

Tatusov, R. L. *et al.* (2003) 'The COG database: an updated version includes eukaryotes', *BMC bioinformatics*, 4, p. 41.

Tavaré, S. (1986) 'Some probabilistic and statistical problems in the analysis of DNA sequences', *Lectures on mathematics in the life sciences*, 17(2), pp. 57–86.

Thauer, R. K. *et al.* (2008) 'Methanogenic archaea: ecologically relevant differences in energy conservation', *Nature reviews. Microbiology*, 6(8), pp. 579–591.

Thomas, N. A., Bardy, S. L. and Jarrell, K. F. (2001) 'The archaeal flagellum: a different

kind of prokaryotic motility structure', *FEMS microbiology reviews*, 25(2), pp. 147–174.

Thorne, J. L., Kishino, H. and Painter, I. S. (1998) 'Estimating the rate of evolution of the rate of molecular evolution', *Molecular biology*. academic.oup.com. Available at: https://academic.oup.com/mbe/article-abstract/15/12/1647/963101.

Tocheva, E. I. *et al.* (2011) 'Peptidoglycan remodeling and conversion of an inner membrane into an outer membrane during sporulation', *Cell*, 146(5), pp. 799–812.

Tocheva, E. I., Ortega, D. R. and Jensen, G. J. (2016) 'Sporulation, bacterial cell envelopes and the origin of life', *Nature reviews. Microbiology*, 14(8), pp. 535–542.

Tria, F. D. K., Landan, G. and Dagan, T. (2017) 'Phylogenetic rooting using minimal ancestor deviation', *Nature ecology & evolution*, 1, p. 193.

'Ueber die isolirte Färbung der Schizomyceten in Schnitt- und Trockenpräparaten von Dr. Gram, Kopenhagen. — Fortschritte der Medicin 1884 No. 6. Ref. Dr. Becker' (1884) *DMW - Deutsche Medizinische Wochenschrift*, pp. 234–235. doi: 10.1055/s-0029-1209285.

Ueno, Y. *et al.* (2006) 'Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era', *Nature*, 440(7083), pp. 516–519.

Varga, M., *et al.* (2016) 'Molecular characterization of a new efficiently transducing bacteriophage identified in meticillin-resistant Staphylococcus aureus.' *Journal of General Virology* 97.1, pp. 258-268.

Vignais, P. M., Billoud, B. and Meyer, J. (2001) 'Classification and phylogeny of hydrogenases', *FEMS microbiology reviews*, 25(4), pp. 455–501.

Villanueva, L., von Meijenfeldt, F. A. B. and Westbye, A. B. (2018) 'Bridging the divide: bacteria synthesizing archaeal membrane lipids', *bioRxiv*. biorxiv.org. Available at: https://www.biorxiv.org/content/10.1101/448035v1.abstract.

Villanueva, L., Schouten, S. and Damsté, J. S. S. (2017) 'Phylogenomic analysis of lipid biosynthetic genes of Archaea shed light on the "lipid divide"', *Environmental microbiology*, 19(1), pp. 54–69.

Vossenberg, J. L. C. M. van de *et al.* (1998) 'The essence of being extremophilic: the

role of the unique archaeal membrane lipids', *Extremophiles*, pp. 163–170. doi: 10.1007/s007920050056.

Wächtershäuser, G. (1990) 'Evolution of the first metabolic cycles', *Proceedings of the National Academy of Sciences of the United States of America*, 87(1), pp. 200–204.

Wächtershäuser, G. (2003) 'From pre-cells to Eukarya--a tale of two lipids', *Molecular microbiology*. Wiley Online Library, 47(1), pp. 13–22.

Wade, T. *et al.* (2020) 'Assessing the accuracy of phylogenetic rooting methods on prokaryotic gene families', *PloS one*, 15(5), p. e0232950.

Wagner, M. *et al.* (1998) 'Phylogeny of dissimilatory sulfite reductases supports an early origin of sulfate respiration', *Journal of bacteriology*, 180(11), pp. 2975–2982.

Wagner, M. and Horn, M. (2006) 'The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance', *Current opinion in biotechnology*, 17(3), pp. 241–249.

Wang, H.-C. *et al.* (2018) 'Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation', *Systematic biology*, 67(2), pp. 216–235.

Watson, J. D. and Crick, F. H. (1953) 'Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid', *Nature*, 171(4356), pp. 737–738.

Weijers, J. W. H. *et al.* (2006) 'Membrane lipids of mesophilic anaerobic bacteria thriving in peats have typical archaeal traits', *Environmental microbiology*, 8(4), pp. 648–657.

Weiss, M. C. *et al.* (2016) 'The physiology and habitat of the last universal common ancestor', *Nature microbiology*, 1(9), p. 16116.

Westphal, L. *et al.* (2018) 'The Rnf Complex Is an Energy-Coupled Transhydrogenase Essential To Reversibly Link Cellular NADH and Ferredoxin Pools in the Acetogen Acetobacterium woodii', *Journal of bacteriology*, 200(21). doi: 10.1128/JB.00357-18.

White, W. T. *et al.* (2007) 'Treeness triangles: visualizing the loss of phylogenetic

signal', *Molecular biology and evolution*, 24(9), pp. 2029–2039.

Wilkinson, M. *et al.* (2007) 'Of clades and clans: terms for phylogenetic relationships in unrooted trees', *Trends in ecology & evolution*, 22(3), pp. 114–115.

Williams, T. A. *et al.* (2013) 'An archaeal origin of eukaryotes supports only two primary domains of life', *Nature*, 504(7479), pp. 231–236.

Williams, T. A. *et al.* (2015) 'New substitution models for rooting phylogenetic trees', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1678), p. 20140336.

Williams, T. A. *et al.* (2017) 'Integrative modeling of gene and genome evolution roots the archaeal tree of life', *Proceedings of the National Academy of Sciences of the United States of America*, 114(23), pp. E4602–E4611.

Williams, T. A., Martin Embley, T. and Heinz, E. (2011) 'Informational Gene Phylogenies Do Not Support a Fourth Domain of Life for Nucleocytoplasmic Large DNA Viruses', *PLoS ONE*, p. e21080. doi: 10.1371/journal.pone.0021080.

Woese, C. R., Kandler, O. and Wheelis, M. L. (1990) 'Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya', *Proceedings of the National Academy of Sciences of the United States of America*, 87(12), pp. 4576–4579.

Wolfe, J. M. and Fournier, G. P. (2018) 'Horizontal gene transfer constrains the timing of methanogen evolution', *Nature ecology & evolution*, 2(5), pp. 897–903.

Wu, H. *et al.* (2014) 'The quest for a unified view of bacterial land colonization', *The ISME journal*, 8(7), pp. 1358–1369.

Yang, Z. (1994) 'Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods', *Journal of molecular evolution*, 39(3), pp. 306–314.

Yao, J. and Rock, C. O. (2013) 'Phosphatidic acid synthesis in bacteria', *Biochimica et biophysica acta*, 1831(3), pp. 495–502.

Yokobori, S.-I. *et al.* (2016) ‘Birth of archaeal cells: molecular phylogenetic analyses of G1P dehydrogenase, G3P dehydrogenases, and glycerol kinase suggest derived features of archaeal membranes having G1P polar lipids’, *Archaea* . Hindawi, 2016. Available at: https://www.hindawi.com/journals/archaea/2016/1802675/abs/.

Zaremba-Niedzwiedzka, K. *et al.* (2017) ‘Asgard archaea illuminate the origin of eukaryotic cellular complexity’, *Nature*, 541(7637), pp. 353–358.

Zhang, C., Sayyari, E. and Mirarab, S. (2017) ‘ASTRAL-III: Increased Scalability and Impacts of Contracting Low Support Branches’, in *Comparative Genomics*. *RECOMB International Workshop on Comparative Genomics*, Springer, Cham (Lecture Notes in Computer Science), pp. 53–75.

Zhaxybayeva, O., Lapierre, P. and Gogarten, J. P. (2005) ‘Ancient gene duplications and the root(s) of the tree of life’, *Protoplasma*, 227(1), pp. 53–64.

Zhu, Q. *et al.* (2019) ‘Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea’, *Nature communications*, 10(1), p. 5477.

Zverlov, V. *et al.* (2005) ‘Lateral gene transfer of dissimilatory (bi)sulfite reductase revisited’, *Journal of bacteriology*, 187(6), pp. 2203–2208.

Zwaenepoel, A. and Van de Peer, Y. (2019) ‘Inference of Ancient Whole-Genome Duplications and the Evolution of Gene Duplication and Loss Rates’, *Molecular biology and evolution*. academic.oup.com, 36(7), pp. 1384–1404.

Appendix A

Taxon tables used in analyses for Chapter 2

Table 1 Taxon sampling for the GTDB dataset, Chapter 2

Species code	GTDB Taxonomy
n0	d__Bacteria;p__Verrucomicrobiota;c__Kiritimatiellae;o__RFP12;f__UBA1067;g__UBA1200;s__UBA1200
n1	d__Bacteria;p__Verrucomicrobiota;c__Kiritimatiellae;o__Kiritimatiellales;f__UBA1859;g__SZUA-127;s__SZUA-127
n2	d__Bacteria;p__Verrucomicrobiota;c__Lentisphaeria;o__Victivallales;f__Victivallaceae;g__UBA7640;s__UBA7640
n3	d__Bacteria;p__Verrucomicrobiota;c__Lentisphaeria;o__UBA2565;f__UBA2565;g__UBA2565;s__UBA2565
n4	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__UBA10450;g__Palsa-1382;s__Palsa-1382
n5	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Pedosphaerales;f__GCA-2715965;g__GCA-2715965;s__GCA-2715965
n6	d__Bacteria;p__Verrucomicrobiota_A;c__Chlamydiia;o__Parachlamydiales;f__Parachlamydiaceae;g__Parachlamydia;s__Parachlamydia
n7	d__Bacteria;p__Verrucomicrobiota_A;c__Chlamydiia;o__Chlamydiales;f__Chlamydiaceae;g__Chlamydophila;s__Chlamydophila
n8	d__Bacteria;p__Planctomycetota;c__Planctomycetes;o__Pirellulales;f__UBA7805;g__UBA7805;s__UBA7805
n9	d__Bacteria;p__Planctomycetota;c__Planctomycetes;o__Gemmatales;f__HRBIN36;g__HRBIN36;s__HRBIN36
n10	d__Bacteria;p__Planctomycetota;c__Phycisphaerae;o__Phycisphaerales;f__SM1A02;g__UBA5793;s__UBA5793
n11	d__Bacteria;p__Planctomycetota;c__Phycisphaerae;o__SG8-4;f__SG8-4;g__PLanc-01;s__PLanc-01
n12	d__Bacteria;p__Planctomycetota;c__FEN-1346;o__FEN-1346;f__FEN-1346;g__FEN-1346;s__FEN-1346
n13	d__Bacteria;p__Planctomycetota;c__Brocadiae;o__Brocadiales;f__2-02-FULL-50-16-A;g__2-02-FULL-50-16-A;s__2-02-FULL-50-16-A
n14	d__Bacteria;p__Planctomycetota;c__UBA1135;o__UBA1135;f__GCA-002686595;g__GCA-2686265;s__GCA-2686265
n15	d__Bacteria;p__Planctomycetota;c__UBA1135;o__UBA2386;f__GCA-002748355;g__GCA-2748355;s__GCA-2748355
n16	d__Bacteria;p__Planctomycetota;c__GCA-002687715;o__GCA-002687715;f__GCA-002687715;g__GCA-2683135;s__GCA-2683135
n17	d__Bacteria;p__Planctomycetota;c__UBA8108;o__UBA1146;f__UBA1146;g__UBA1146;s__UBA1146
n18	d__Bacteria;p__Planctomycetota;c__UBA8108;o__UBA8108;f__UBA8108;g__UBA8108;s__UBA8108
n19	d__Bacteria;p__Planctomycetota;c__UBA11346;o__GCA-2746535;f__GCA-2746535;g__GCA-2746535;s__GCA-2746535
n20	d__Bacteria;p__Planctomycetota;c__UBA11346;o__UBA11346;f__UBA11346;g__BOG-1363;s__BOG-1363
n21	d__Bacteria;p__Planctomycetota;c__SZUA-567;o__SZUA-567;f__SZUA-567;g__SZUA-421;s__SZUA-421
n22	d__Bacteria;p__Spirochaetota;c__Spirochaetia;o__Sphaerochaetales;f__Sphaerochaetaceae;g__Sphaerochaeta;s__Sphaerochaeta
n23	d__Bacteria;p__Spirochaetota;c__Spirochaetia;o__Borreliales;f__Borreliaceae;g__Borrelia;s__Borrelia
n24	d__Bacteria;p__Spirochaetota;c__GWE2-31-10;o__GWE2-31-10;f__GWE2-31-10;g__GWC1-27-15;s__GWC1-27-15
n25	d__Bacteria;p__Spirochaetota;c__UBA6919;o__GWB1-36-13;f__GWB1-36-13;g__GWB1-36-13;s__GWB1-36-13
n26	d__Bacteria;p__Spirochaetota;c__UBA6919;o__UBA6919;f__UBA6919;g__UBA6919;s__UBA6919
n27	d__Bacteria;p__Spirochaetota;c__Brevinematia;o__Brevinematales;f__GWF1-51-8;g__GWF1-51-8;s__GWF1-51-8
n28	d__Bacteria;p__Spirochaetota;c__Brachyspirae;o__Brachyspirales;f__Brachyspiraceae;g__Brachyspira;s__Brachyspira
n29	d__Bacteria;p__Spirochaetota;c__Leptospirae;o__Turneriellales;f__Turneriellaceae;g__Turneriella;s__Turneriella
n30	d__Bacteria;p__Spirochaetota;c__Leptospirae;o__Leptospirales;f__Leptospiraceae;g__Leptospira_A;s__Leptospira_A
n31	d__Bacteria;p__Spirochaetota;c__UBA4802;o__UBA4802;f__UBA5368;g__UBA5368;s__UBA5368

n32 d__Bacteria;p__Firmicutes_A;c__Clostridia;o__SK-Y3;f__SK-Y3;g__SK-Y3;s__SK-Y3

n33 d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Tissierellales;f__PP17-6a;g__PP17-6a;s__PP17-6a

n34 d__Bacteria;p__Firmicutes_A;c__Mahellia;o__Mahellales;f__Mahellaceae;g__Mahella;s__Mahella

n35 d__Bacteria;p__Firmicutes_A;c__Mahellia;o__Caldicoprobacterales;f__UBA3920;g__UBA3920;s__UBA3920

n36 d__Bacteria;p__Firmicutes_A;c__Thermoanaerobacteria;o__Thermoanaerobacterales;f__Thermoanaerobacteraceae;g__Thermoanaerobacterium;s__Thermoanaerobacterium

n37 d__Bacteria;p__Firmicutes_A;c__Thermoanaerobacteria;o__Caldicellulosiruptorales;f__Caldicellulosiruptoraceae;g__Caldicellulosiruptor;s__Caldicellulosiruptor

n38 d__Bacteria;p__Firmicutes_A;c__Thermovenabulia;o__UBA3567;f__UBA3567;g__UBA3567;s__UBA3567

n39 d__Bacteria;p__Firmicutes_A;c__Thermovenabulia;o__Thermovenabulales;f__Tepidanaerobacteraceae;g__Tepidanaerobacter;s__Tepidanaerobacter

n40 d__Bacteria;p__Firmicutes_I;c__Bacilli_A;o__Paenibacillales;f__Paenibacillaceae;g__Paenibacillus_Q;s__Paenibacillus_Q

n41 d__Bacteria;p__Firmicutes_I;c__Bacilli_A;o__Thermoactinomycetales;f__Thermoactinomycetaceae;g__Marininema;s__Marininema

n42 d__Bacteria;p__Firmicutes_K;c__Alicyclobacillia;o__Alicyclobacillales;f__Acidibacillaceae;g__Acidibacillus;s__Acidibacillus

n43 d__Bacteria;p__Firmicutes_K;c__Alicyclobacillia;o__Kyrpidiales;f__Kyrpidiaceae;g__Kyrpidia;s__Kyrpidia

n44 d__Bacteria;p__Firmicutes_D;c__Dethiobacteria;o__Dethiobacterales;f__Dethiobacteraceae;g__Dethiobacter;s__Dethiobacter

n45 d__Bacteria;p__Firmicutes_D;c__Dethiobacteria;o__DTU022;f__DTU022;g__DTU022;s__DTU022

n46 d__Bacteria;p__Firmicutes_D;c__Natranaerobiia;o__Natranaerobiales;f__Natranaerobiaceae;g__Natranaerobius_A;s__Natranaerobius_A

n47 d__Bacteria;p__Firmicutes_D;c__Proteinivoracia;o__UBA4975;f__UBA4975;g__UBA4975;s__UBA4975

n48 d__Bacteria;p__Firmicutes_D;c__Proteinivoracia;o__Proteinivoracales;f__Proteinivoraceae;g__Anaerobranca;s__Anaerobranca

n49 d__Bacteria;p__Firmicutes_D;c__UBA994;o__UBA994;f__UBA994;g__UBA994;s__UBA994

n50 d__Bacteria;p__Firmicutes_G;c__Limnochordia;o__DTU010;f__DTU012;g__DTU012;s__DTU012

n51 d__Bacteria;p__Firmicutes_G;c__Limnochordia;o__Limnochordales;f__Limnochordaceae;g__Limnochorda;s__Limnochorda

n52 d__Bacteria;p__Firmicutes_G;c__DTU065;o__DTU065;f__DTU065;g__DTU065;s__DTU065

n53 d__Bacteria;p__Firmicutes_G;c__SHA-98;o__UBA4971;f__UBA4971;g__UBA4971;s__UBA4971

n54 d__Bacteria;p__Firmicutes_G;c__SHA-98;o__DTUO25;f__DTU025;g__UBA5389;s__UBA5389

n55 d__Bacteria;p__Firmicutes_G;c__DTU014;o__DTU014;f__DTU014;g__DTU014;s__DTU014

n56 d__Bacteria;p__Firmicutes_G;c__UBA5435;o__UBA5435;f__UBA5435;g__UBA5435;s__UBA5435

n57 d__Bacteria;p__Firmicutes_G;c__UBA4882;o__UBA4882;f__UBA4882;g__UBA4882;s__UBA4882

n58 d__Bacteria;p__Firmicutes_G;c__UBA4882;o__UBA10575;f__UBA3943;g__UBA3943;s__UBA3943

n59 d__Bacteria;p__Firmicutes_E;c__Sulfobacillia;o__Sulfobacillales;f__Sulfobacillaceae;g__Sulfobacillus_A;s__Sulfobacillus_A

n60 d__Bacteria;p__Firmicutes_E;c__Symbiobacteriia;o__Symbiobacteriales;f__ZC4RG38;g__ZC4RG38;s__ZC4RG38

n61 d__Bacteria;p__Firmicutes_E;c__Thermaerobacteria;o__Thermaerobacterales;f__Thermaerobacteraceae;g__Thermaerobacter;s__Thermaerobacter

n62 d__Bacteria;p__Firmicutes_E;c__DTU015;o__D8A-2;f__D2;g__DTU015;s__DTU015

n63 d__Bacteria;p__Firmicutes_E;c__DTU015;o__UBA995;f__UBA995;g__UBA995;s__UBA995

n64 d__Bacteria;p__Firmicutes_E;c__UBA3569;o__UBA3569;f__UBA3569;g__UBA3569;s__UBA3569

n65 d__Bacteria;p__Firmicutes_E;c__UBA3569;o__UBA3575;f__UBA3575;g__UBA3575;s__UBA3575

n66 d__Bacteria;p__Firmicutes_B;c__Desulfitobacteriia;o__Heliobacteriales;f__Heliobacteriaceae;g__Heliobacterium;s__Heliobacterium

n67 d__Bacteria;p__Firmicutes_B;c__Desulfitobacteriia;o__Desulfitobacteriales;f__Desulfitobacteriaceae;g__Desulfosporosinus;s__Desulfosporosinus

n68 d__Bacteria;p__Firmicutes_B;c__Peptococcia;o__DRI-13;f__DRI-13;g__DRI-13;s__DRI-13

n69 d__Bacteria;p__Firmicutes_B;c__Peptococcia;o__Peptococcales;f__Peptococcaceae;g__UBA7185;s__UBA7185

n70 d__Bacteria;p__Firmicutes_B;c__Moorellia;o__Moorellales;f__Moorellaceae;g__Moorella;s__Moorella

n71 d__Bacteria;p__Firmicutes_B;c__Moorellia;o__Desulfitibacterales;f__Desulfitibacteraceae;g__Desulfitibacter;s__Desulfitibacter

n72	d__Bacteria;p__Firmicutes_B;c__Dehalobacteriia;o__UBA7702;f__UBA7702;g__UBA7702;s__UBA7702
n73	d__Bacteria;p__Firmicutes_B;c__Dehalobacteriia;o__UBA4068;f__UBA5755;g__UBA5755;s__UBA5755
n74	d__Bacteria;p__Firmicutes_B;c__Syntrophomonadia;o__Syntrophomonadales;f__Syntrophothermaceae;g__Syntrophothermus;s__Syntrophothermus
n75	d__Bacteria;p__Firmicutes_B;c__Syntrophomonadia;o__Thermacetogeniales;f__Thermacetogeniaceae;g__Thermacetogenium;s__Thermacetogenium
n76	d__Bacteria;p__Firmicutes_B;c__Desulfotomaculia;o__Ammonifexales;f__Ammonificaceae;g__Ammonifex;s__Ammonifex
n77	d__Bacteria;p__Firmicutes_B;c__Desulfotomaculia;o__Carboxydothermales;f__Carboxydothermaceae;g__Carboxydothermus;s__Carboxydothermus
n78	d__Bacteria;p__Firmicutes_B;c__Thermincolia_A;o__Carboxydocellales;f__Carboxydocellaceae;g__Carboxydocella;s__Carboxydocella
n79	d__Bacteria;p__Firmicutes_B;c__Thermincolia;o__Thermincolales;f__Thermincolaceae;g__Thermincola;s__Thermincola
n80	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__UBA11029;f__UBA11029;g__UBA11029;s__UBA11029
n81	d__Bacteria;p__Firmicutes_C;c__Negativicutes;o__Sporomusales;f__Thermosinaceae;g__Sporolituus;s__Sporolituus
n82	d__Bacteria;p__Firmicutes_F;c__Halanaerobiia;o__Halanaerobiales;f__DTU029;g__DTU029;s__DTU029
n83	d__Bacteria;p__Firmicutes_F;c__Halanaerobiia;o__Halobacteroidales;f__Acetohalobiaceae;g__Acetohalobium;s__Acetohalobium
n84	d__Bacteria;p__Fusobacteriota;c__Fusobacteriia;o__bac17;f__UBA816;g__UBA816;s__UBA816
n85	d__Bacteria;p__Fusobacteriota;c__Fusobacteriia;o__Fusobacteriales;f__Leptotrichiaceae;g__Leptotrichia;s__Leptotrichia
n86	d__Bacteria;p__Firmicutes;c__Bacilli;o__RFN20;f__CAG-288;g__CAG-288;s__CAG-288
n87	d__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Erysipelothrix;s__Erysipelothrix
n88	d__Bacteria;p__Cyanobacteria;c__Vampirovibrionia;o__Gastranaerophilales;f__UBA9579;g__UBA9579;s__UBA9579
n89	d__Bacteria;p__Cyanobacteria;c__Vampirovibrionia;o__Caenarcaniphilales;f__2-02-FULL-35-15;g__2-02-FULL-34-12;s__2-02-FULL-34-12
n90	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Gloeomargaritales;f__Gloeomargaritaceae;g__Gloeomargarita;s__Gloeomargarita
n91	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Neosynechococcales;f__Neosynechococcaceae;g__Neosynechococcus;s__Neosynechococcus
n92	d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Elainellales;f__Elainellaceae;g__Elainella;s__Elainella
n93	d__Bacteria;p__Cyanobacteria;c__Sericytochromatia;o__S15B-MN24;f__UBA4093;g__UBA4093;s__UBA4093
n94	d__Bacteria;p__Margulisbacteria;c__WOR-1;o__O2-12-FULL-45-9;f__XYB2-FULL-48-7;g__XYB2-FULL-45-9;s__XYB2-FULL-45-9
n95	d__Bacteria;p__Margulisbacteria;c__WOR-1;o__XYC2-FULL-46-14;f__XYC2-FULL-46-14;g__XYC2-FULL-46-14;s__XYC2-FULL-46-14
n96	d__Bacteria;p__Margulisbacteria;c__ZB3;o__UBA6595;f__UBA6595;g__UBA6595;s__UBA6595
n97	d__Bacteria;p__Margulisbacteria;c__ZB3;o__UBA817;f__GCA-002719695;g__AG-343-D04;s__AG-343-D04
n98	d__Bacteria;p__Margulisbacteria;c__GWF2-35-9;o__GWF2-39-127;f__GWF2-39-127;g__GWF2-39-127;s__GWF2-39-127
n99	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Streptosporangiales;f__Streptosporangiaceae;g__Nocardiopsis;s__Nocardiopsis
n100	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Nanopelagicales;f__Nanopelagicaceae;g__Planktophila;s__Planktophila
n101	d__Bacteria;p__Actinobacteriota;c__UBA4738;o__UBA4738;f__UBA4738;g__UBA5182;s__UBA5182
n102	d__Bacteria;p__Actinobacteriota;c__Acidimicrobiia;o__Acidimicrobiales;f__Bog-793;g__Bog-793;s__Bog-793
n103	d__Bacteria;p__Actinobacteriota;c__Acidimicrobiia;o__UBA5794;f__UBA4744;g__BMS3Bbin01;s__BMS3Bbin01
n104	d__Bacteria;p__Actinobacteriota;c__Thermoleophilia;o__20CM-4-69-9;f__20CM-4-69-9;g__Palsa-739;s__Palsa-739
n105	d__Bacteria;p__Actinobacteriota;c__Thermoleophilia;o__UBA2241;f__UBA2241;g__UBA2241;s__UBA2241
n106	d__Bacteria;p__Actinobacteriota;c__Rubrobacteria;o__Rubrobacterales;f__Rubrobacteraceae;g__Rubrobacter;s__Rubrobacter
n107	d__Bacteria;p__Actinobacteriota;c__Coriobacteriia;o__Coriobacteriales;f__Atopobiaceae;g__Olegusella;s__Olegusella
n108	d__Bacteria;p__Actinobacteriota;c__Coriobacteriia;o__OPB41;f__PALSA-660;g__PALSA-660;s__PALSA-660
n109	d__Bacteria;p__Actinobacteriota;c__UBA9087;o__UBA10029;f__UBA10029;g__CG2-30-50-142;s__CG2-30-50-142
n110	d__Bacteria;p__Actinobacteriota;c__UBA9087;o__UBA9087;f__UBA2594;g__UBA2594;s__UBA2594

n111 d__Bacteria;p__Actinobacteriota;c__UBA1414;o__UBA1414;f__UBA1414;g__UBA1414;s__UBA1414

n112 d__Bacteria;p__Actinobacteriota;c__RBG-13-55-18;o__RBG-13-55-18;f__RBG-13-55-18;g__RBG-13-55-18;s__RBG-13-55-18

n113 d__Bacteria;p__Actinobacteriota;c__RBG-13-55-18;o__Fen-727;f__Fen-727;g__FEN-680;s__FEN-680

n114 d__Bacteria;p__Eremiobacterota;c__Eremiobacteria;o__Eremiobacterales;f__Eremiobacteraceae;g__Eremiobacter;s__Eremiobacter

n115 d__Bacteria;p__Eremiobacterota;c__Eremiobacteria;o__UBP12;f__UBA5184;g__Palsa-1483;s__Palsa-1483

n116 d__Bacteria;p__Eremiobacterota;c__UBP9;o__UBA4705;f__UBA4705;g__UBA4705;s__UBA4705

n117 d__Bacteria;p__Armatimonadota;c__Fimbriimonadia;o__Fimbriimonadales;f__Fimbriimonadaceae;g__UBA2017;s__UBA2017

n118 d__Bacteria;p__Armatimonadota;c__Fimbriimonadia;o__OS-L;f__GBS-DC;g__HRBIN15;s__HRBIN15

n119 d__Bacteria;p__Armatimonadota;c__HRBIN16;o__HRBIN16;f__HRBIN16;g__HRBIN16;s__HRBIN16

n120 d__Bacteria;p__Armatimonadota;c__Chthonomonadetes;o__Chthonomonadales;f__Chthonomonadaceae;g__Chthonomonas;s__Chthonomonas

n121 d__Bacteria;p__Armatimonadota;c__BOG-944;o__BOG-944;f__BOG-944;g__BOG-944;s__BOG-944

n122 d__Bacteria;p__Armatimonadota;c__UBA5377;o__UBA1398;f__UBA1398;g__UBA1398;s__UBA1398

n123 d__Bacteria;p__Armatimonadota;c__UBA5377;o__UBA5377;f__UBA5352;g__UBA5352;s__UBA5352

n124 d__Bacteria;p__Armatimonadota;c__Abditibacteria;o__Abditibacteriales;f__Abditibacteriaceae;g__Abditibacterium;s__Abditibacterium

n125 d__Bacteria;p__Armatimonadota;c__Abditibacteria;o__CG2-30-59-28;f__CG2-30-59-28;g__CG2-30-59-28;s__CG2-30-59-28

n126 d__Bacteria;p__Armatimonadota;c__HRBIN17;o__HRBIN17;f__HRBIN17;g__HRBIN17;s__HRBIN17

n127 d__Bacteria;p__Patescibacteria;c__Paceibacteria;o__WO2-41-13;f__UBA11713;g__2-01-FULL-46-13;s__2-01-FULL-46-13

n128 d__Bacteria;p__Patescibacteria;c__Paceibacteria;o__GWC2-36-17;f__M10-OD1-4;g__WO2-46-25;s__WO2-46-25

n129 d__Bacteria;p__Patescibacteria;c__Paceibacteria_A;o__Moranbacterales;f__UBA2206;g__GCA-002790615;s__GCA-002790615

n130 d__Bacteria;p__Patescibacteria;c__ABY1;o__SG8-24;f__GWF2-40-263;g__UBA12396;s__UBA12396

n131 d__Bacteria;p__Patescibacteria;c__ABY1;o__UBA2591;f__UBA2591;g__UBA2591;s__UBA2591

n132 d__Bacteria;p__Patescibacteria;c__Andersenbacteria;o__UBA10190;f__UBA10190;g__HO2-45-11b;s__HO2-45-11b

n133 d__Bacteria;p__Patescibacteria;c__Gracilibacteria;o__Peribacterales;f__Peribacteraceae;g__GWB1-54-5;s__GWB1-54-5

n134 d__Bacteria;p__Patescibacteria;c__Gracilibacteria;o__UBA1369;f__UBA12473;g__UBA12473;s__UBA12473

n135 d__Bacteria;p__Patescibacteria;c__Saccharimonadia;o__Saccharimonadales;f__Saccharimonadaceae;g__UBA2866;s__UBA2866

n136 d__Bacteria;p__Patescibacteria;c__CG2-30-33-46;o__CG2-30-33-46;f__CG2-30-33-46;g__CG2-30-33-46;s__CG2-30-33-46

n137 d__Bacteria;p__Patescibacteria;c__UBA1384;o__UBA12157;f__GCA-002773005;g__GCA-002773005;s__GCA-002773005

n138 d__Bacteria;p__Patescibacteria;c__UBA1384;o__UBA1384;f__1-14-0-10-41-12;g__1-14-0-10-41-12;s__1-14-0-10-41-12

n139 d__Bacteria;p__Patescibacteria;c__WWE3;o__UBA101185;f__UBA10185;g__UBA8498;s__UBA8498

n140 d__Bacteria;p__Patescibacteria;c__4484-211;o__4484-211;f__4484-211;g__4484-211;s__4484-211

n141 d__Bacteria;p__Patescibacteria;c__Microgenomatia;o__UBA1406;f__HO2-37-13b;g__2-12-FULL-37-7b;s__2-12-FULL-37-7b

n142 d__Bacteria;p__Patescibacteria;c__Microgenomatia;o__Levybacterales;f__UBA12049;g__PJMG01;s__PJMG01

n143 d__Bacteria;p__Patescibacteria;c__Dojkabacteria;o__SC72;f__SC72;g__UBA5209;s__UBA5209

n144 d__Bacteria;p__Chloroflexota;c__UBA6077;o__UBA6077;f__UBA6077;g__UBA6077;s__UBA6077

n145 d__Bacteria;p__Chloroflexota;c__Dehalococcoidia;o__UBA2963;f__UBA2963;g__Bin16;s__Bin16

n146 d__Bacteria;p__Chloroflexota;c__Dehalococcoidia;o__SAR202-VII-2;f__SAR202-VII-2;g__SAR202-VII-2;s__SAR202-VII-2

n147 d__Bacteria;p__Chloroflexota;c__Anaerolineae;o__UBA3071;f__CG2-30-64-16;g__CG2-30-64-16;s__CG2-30-64-16

n148 d__Bacteria;p__Chloroflexota;c__Anaerolineae;o__UBA4142;f__UBA4142;g__UBA4142;s__UBA4142

n149 d__Bacteria;p__Chloroflexota;c__Chloroflexia;o__Chloroflexales;f__Roseiflexaceae;g__Roseiflexus;s__Roseiflexus

n150 d__Bacteria;p__Chloroflexota;c__Chloroflexia;o__54-19;f__54-19;g__54-19;s__54-19

d__Bacteria;p__Chloroflexota;c__Ktedonobacteria;o__Ktedonobacterales;f__Ktedonobacteraceae;g__Thermogemma
n151 tispora;s__Thermogemmatispora

n152 d__Bacteria;p__Chloroflexota;c__UBA2235;o__UBA12225;f__UBA12225;g__UBA12225;s__UBA12225

n153 d__Bacteria;p__Dormibacterota;c__Dormibacteria;o__UBA8260;f__Bog-877;g__Bog-877;s__Bog-877
d__Bacteria;p__Dormibacterota;c__Dormibacteria;o__Dormibacterales;f__Dormibacteraceae;g__Palsa-872;s__Palsa-
n154 872

n155 d__Bacteria;p__Chloroflexota_A;c__Ellin6529;o__QHBO01;f__QHBO01;g__QHBO01;s__QHBO01

n156 d__Bacteria;p__Chloroflexota_A;c__Ellin6529;o__CSP1-4;f__CSP1-4;g__Fen-1039;s__Fen-1039
d__Bacteria;p__Deinococcota;c__Deinococci;o__Deinococcales;f__Marinithermaceae;g__Marinithermus;s__Marinith
n157 ermus
d__Bacteria;p__Thermotogota;c__Thermotogae;o__Thermotogales;f__Thermotogaceae;g__Pseudothermotoga;s__Ps
n158 eudothermotoga
d__Bacteria;p__Thermotogota;c__Thermotogae;o__Mesoaciditogales;f__Mesoaciditogaceae;g__Mesoaciditoga;s__M
n159 esoaciditoga

n160 d__Bacteria;p__Synergistota;c__Synergistia;o__Synergistales;f__54-24;g__54-24;s__54-24

n161 d__Bacteria;p__Synergistota;c__GBS-1;o__GBS-1;f__GBS-1;g__GBS-1;s__GBS-1
d__Bacteria;p__Fibrobacterota;c__Fibrobacteria;o__Fibrobacterales;f__Fibrobacteraceae;g__Fibrobacter;s__Fibrobac
n162 ter

n163 d__Bacteria;p__Fibrobacterota;c__Fibrobacteria;o__UBA5070;f__UBA5070;g__UBA5070;s__UBA5070
d__Bacteria;p__Fibrobacterota;c__OXYB2-FULL-49-7;o__OXYB2-FULL-49-7;f__XYB2-FULL-49-35;g__XYB2-FULL-49-
n164 35;s__XYB2-FULL-49-35
d__Bacteria;p__Fibrobacterota;c__Chitinivibrionia;o__Chitinivibrionales;f__Chitinivibrionaceae;g__Chitinivibrio;s__Ch
n165 itinivibrio
d__Bacteria;p__Cloacimonadota;c__Cloacimonadia;o__Cloacimonadales;f__Cloacimonadaceae;g__UBA1060;s__UBA1
n166 060

n167 d__Bacteria;p__Cloacimonadota;c__Cloacimonadia;o__JGIOTU-2;f__JGIOTU-2;g__JGIOTU-2;s__JGIOTU-2

n168 d__Bacteria;p__Marinisomatota;c__Marinisomatia;o__Marinisomatales;f__S15-B10;g__S15-B10;s__S15-B10
d__Bacteria;p__Marinisomatota;c__Marinisomatia;o__SCGC-AAA003-L08;f__SCGC-AAA003-L08;g__SCGC-AAA003-
n169 L08;s__SCGC-AAA003-L08

n170 d__Bacteria;p__Marinisomatota;c__UBA8477;o__UBA8477;f__UBA8477;g__UBA8477;s__UBA8477

n171 d__Bacteria;p__Marinisomatota;c__AB16;o__AB16;f__TCS52;g__TCS52;s__TCS52

n172 d__Bacteria;p__Marinisomatota;c__UBA2242;o__UBA2242;f__UBA2242;g__UBA2242;s__UBA2242
d__Bacteria;p__Bacteroidota;c__Rhodothermia;o__Rhodothermales;f__UBA10348;g__1-14-0-65-60-17;s__1-14-0-65-
n173 60-17

n174 d__Bacteria;p__Bacteroidota;c__Rhodothermia;o__Balneolales;f__HLUCCA01;g__HLUCCA01;s__HLUCCA01

n175 d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__NS11-12g;f__UBA9320;g__UBA9320;s__UBA9320

n176 d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Thermonemataceae;g__Raineya;s__Raineya

n177 d__Bacteria;p__Bacteroidota;c__Chlorobia;o__Chlorobiales;f__Chlorobiaceae;g__Chlorobium;s__Chlorobium
d__Bacteria;p__Bacteroidota;c__Ignavibacteria;o__Ignavibacteriales;f__Ignavibacteriaceae;g__XYD12-FULL-36-
n178 8;s__XYD12-FULL-36-8

n179 d__Bacteria;p__Bacteroidota;c__Ignavibacteria;o__SJA-28;f__OLB5;g__OLB5;s__OLB5
d__Bacteria;p__Bacteroidota;c__UBA10030;o__0-14-0-80-59-12;f__0-14-0-80-59-12;g__0-14-0-80-59-12;s__0-14-0-
n180 80-59-12

n181 d__Bacteria;p__Bacteroidota;c__UBA10030;o__UBA10030;f__UBA10030;g__Fen-1268;s__Fen-1268

n182 d__Bacteria;p__Bacteroidota;c__Kryptonia;o__Kryptoniales;f__Kryptoniaceae;g__Kryptonium;s__Kryptonium

n183 d__Bacteria;p__Bacteroidota;c__Kapabacteria;o__Kapabacteriales;f__NICIL-2;g__NICIL-2;s__NICIL-2

n184 d__Bacteria;p__Bacteroidota;c__Kapabacteria;o__Palsa-1295;f__Palsa-1295;g__PALSA-1295;s__PALSA-1295

n185 d__Bacteria;p__Bacteroidota;c__SZUA-365;o__SZUA-365;f__SZUA-365;g__SZUA-365;s__SZUA-365
d__Bacteria;p__Gemmatimonadota;c__Gemmatimonadetes;o__Gemmatimonadales;f__GWC2-71-9;g__20CM-2-65-
n186 7;s__20CM-2-65-7

n187 d__Bacteria;p__Gemmatimonadota;c__Gemmatimonadetes;o__RSA9;f__RSA9;g__RSA9;s__RSA9
d__Bacteria;p__Gemmatimonadota;c__Glassbacteria;o__GWA2-58-10;f__GWA2-58-10;g__GWA2-58-10;s__GWA2-
n188 58-10

n189 d__Bacteria;p__Acidobacteriota;c__Thermoanaerobaculia;o__UBA2201;f__UBA2201;g__UBA2201;s__UBA2201

n190 d__Bacteria;p__Acidobacteriota;c__Thermoanaerobaculia;o__Gp7-AA8;f__Gp7-AA8;g__QHVT01;s__QHVT01

n191 d__Bacteria;p__Acidobacteriota;c__Holophagae;o__Holophagales;f__Holophagaceae;g__Geothrix;s__Geothrix
n192 d__Bacteria;p__Acidobacteriota;c__GCA-2747255;o__GCA-2747255;f__GCA-2747255;g__GCA-2747255;s__GCA-2747255
n193 d__Bacteria;p__Acidobacteriota;c__Acidobacteriae;o__Acidobacteriales;f__Gp1-AA117;g__Gp1-AA17;s__Gp1-AA17
n194 d__Bacteria;p__Acidobacteriota;c__Acidobacteriae;o__2-12-FULL-54-10;f__2-12-FULL-54-10;g__2-02-FULL-61-28;s__2-02-FULL-61-28
n195 d__Bacteria;p__Acidobacteriota;c__Blastocatellia;o__Chloracidobacteriales;f__Chloracidobacteriaceae;g__Chloracidobacterium;s__Chloracidobacterium
n196 d__Bacteria;p__Acidobacteriota;c__Blastocatellia;o__HR10;f__HR10;g__HR10;s__HR10
n197 d__Bacteria;p__Acidobacteriota;c__Vicinamibacteria;o__Vicinamibacterales;f__SCN-69-37;g__SCN-69-37;s__SCN-69-37
n198 d__Bacteria;p__Acidobacteriota;c__Mor1;o__Gp22-AA2;f__Gp22-AA2;g__Gp22-AA3;s__Gp22-AA3
n199 d__Bacteria;p__Acidobacteriota;c__Mor1;o__Mor1;f__Mor1;g__Mor1;s__Mor1
n200 d__Bacteria;p__Acidobacteriota;c__Aminicenantia;o__Aminicenantales;f__RBG-16-66-30;g__RBG-16-66-30;s__RBG-16-66-30
n201 d__Bacteria;p__Acidobacteriota;c__Aminicenantia;o__UBA2199;f__UBA2199;g__UBA2199;s__UBA2199
n202 d__Bacteria;p__Acidobacteriota;c__HRBIN11;o__HRBIN11;f__HRBIN11;g__HRBIN11;s__HRBIN11
n203 d__Bacteria;p__Proteobacteria;c__Zetaproteobacteria;o__CG1-02-64-396;f__CG1-02-64-396;g__CG1-02-64-396;s__CG1-02-64-396
n204 d__Bacteria;p__Proteobacteria;c__Zetaproteobacteria;o__Mariprofundales;f__Mariprofundaceae;g__Mariprofundus;s__Mariprofundus
n205 d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__UBA5158;f__UBA5158;g__2-12-FULL-41-20;s__2-12-FULL-41-20
n206 d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Ga0077554;f__Ga007554;g__Ga0077554;s__Ga0077554
n207 d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__SP197;f__SP197;g__SP197;s__SP197
n208 d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Holosporales;f__Holosporaceae;g__40-19;s__40-19
n209 d__Bacteria;p__Proteobacteria;c__Magnetococcia;o__Magnetococcales;f__DC0425bin3;g__DC0425bin3;s__DC0425bin3
n210 d__Bacteria;p__Dependentiae;c__Babeliae;o__Babeliales;f__GCA-2401785;g__GCA-2401785;s__GCA-2401785
n211 d__Bacteria;p__Campylobacterota;c__Campylobacteria;o__Campylobacterales;f__Nitratiruptoraceae;g__Nitratiruptor;s__Nitratiruptor
n212 d__Bacteria;p__Campylobacterota;c__Campylobacteria;o__Nautiliales;f__Nautiliaceae;g__Lebetimonas;s__Lebetimonas
n213 d__Bacteria;p__Campylobacterota;c__Desulfurellia;o__Desulfurellales;f__Hippeaceae;g__Hippea_A;s__Hippea_A
n214 d__Bacteria;p__Aquificota;c__Aquificae;o__Aquificales;f__Aquificaceae;g__Thermocrinis_A;s__Thermocrinis_A
n215 d__Bacteria;p__Aquificota;c__Aquificae;o__Hydrogenothermales;f__Hydrogenothermaceae;g__Persephonella;s__Persephonella
n216 d__Bacteria;p__Aquificota;c__Desulfurobacteriia;o__Desulfurobacteriales;f__Desulfurobacteriaceae;g__Thermovibrio;s__Thermovibrio
n217 d__Bacteria;p__Deferribacterota;c__Deferribacteres;o__Deferribacterales;f__Deferribacteraceae;g__Deferribacter;s__Deferribacter
n218 d__Bacteria;p__Desulfobacterota;c__Desulfobacteria;o__C00003060;f__S7086C20;g__S7086C20;s__S7086C20
n219 d__Bacteria;p__Desulfobacterota;c__Desulfobacteria;o__Desulfatiglandales;f__NaphS2;g__NaphS2;s__NaphS2
n220 d__Bacteria;p__Desulfobacterota;c__Syntrophobacteria;o__BM002;f__BM002;g__BM002;s__BM002
n221 d__Bacteria;p__Desulfobacterota;c__Syntrophobacteria;o__Syntrophobacterales;f__Syntrophobacteraceae;g__Desulfacinum;s__Desulfacinum
n222 d__Bacteria;p__Desulfobacterota;c__Desulfarculia;o__Adiutricales;f__Adiutricaceae;g__Adiutrix;s__Adiutrix
n223 d__Bacteria;p__Desulfobacterota;c__Desulfarculia;o__Desulfarculales;f__Desulfarculaceae;g__Desulfarculus;s__Desulfarculus
n224 d__Bacteria;p__Desulfobacterota;c__Desulfobaccia;o__Desulfobaccales;f__Desulfobaccaceae;g__Desulfobacca;s__Desulfobacca
n225 d__Bacteria;p__Desulfobacterota;c__Thermodesulfobacteria;o__Thermodesulfobacteriales;f__Thermodesulfatatoraceae;g__Thermodesulfatator;s__Thermodesulfatator
n226 d__Bacteria;p__Desulfobacterota;c__Dissulfuribacteria;o__Dissulfuribacterales;f__UBA5754;g__UBA5754;s__UBA5754
n227 d__Bacteria;p__Desulfobacterota;c__Desulfobulbia;o__Desulfobulbales;f__Desulfobulbaceae;g__Desulfobulbus_A;s__Desulfobulbus_A
n228 d__Bacteria;p__Desulfobacterota;c__Desulfofervidia;o__Desulfofervidales;f__Desulfofervidaceae;g__Desulfofervidus;s__Desulfofervidus

n229 d__Bacteria;p__Desulfobacterota;c__JdFR-97;o__JdFR-97;f__JdFR-97;g__JdFR-97;s__JdFR-97

n230 d__Bacteria;p__Desulfobacterota;c__Syntrophia;o__Syntrophales;f__UBA2192;g__UBA2192;s__UBA2192

n231 d__Bacteria;p__Desulfobacterota;c__BSN033;o__BSN033;f__UBA1163;g__UBA1163;s__UBA1163

n232 d__Bacteria;p__Desulfobacterota;c__BSN033;o__UBA8473;f__UBA8473;g__UBA8473;s__UBA8473

n233 d__Bacteria;p__Desulfobacterota;c__Desulfomonilia;o__Desulfomonilales;f__Desulfomonilaceae;g__Desulfomonile;s__Desulfomonile

n234 d__Bacteria;p__Desulfobacterota;c__Desulfomonilia;o__UBA1062;f__UBA1062;g__UBA1062;s__UBA1062

n235 d__Bacteria;p__Desulfobacterota;c__Syntrophorhabdia;o__Syntrophorhabdales;f__Syntrophorhabdaceae;g__UBA8905;s__UBA8905

n236 d__Bacteria;p__Desulfuromonadota;c__Desulfuromonadia;o__Desulfuromonadales;f__Desulfuromonadaceae_C;g__Desulfuromonas_B;s__Desulfuromonas_B

n237 d__Bacteria;p__Desulfuromonadota;c__Desulfuromonadia;o__Geobacterales;f__Pelobacteraceae;g__Pelobacter_D;s__Pelobacter_D

n238 d__Bacteria;p__Nitrospirota;c__RBG-16-64-22;o__RBG-16-64-22;f__RBG-16-64-22;g__RBG-16-64-22;s__RBG-16-64-22

n239 d__Bacteria;p__Nitrospirota;c__UBA9217;o__UBA9217;f__UBA9217;g__GWC2-57-13;s__GWC2-57-13

n240 d__Bacteria;p__Nitrospirota;c__Nitrospiria;o__Nitrospirales;f__Nitrospiraceae;g__Palsa-1315;s__Palsa-1315

n241 d__Bacteria;p__Nitrospirota;c__Thermodesulfovibrionia;o__Thermodesulfovibrionales;f__JdFR-86;g__JdFR-86;s__JdFR-86

n242 d__Bacteria;p__Nitrospirota;c__Thermodesulfovibrionia;o__UBA6902;f__UBA6902;g__UBA6902;s__UBA6902

n243 d__Bacteria;p__Myxococcota;c__Myxococcia;o__Myxococcales;f__Vulgatibacteraceae;g__ZC4RG40;s__ZC4RG40

n244 d__Bacteria;p__Myxococcota;c__UBA727;o__UBA727;f__GCA-2721815;g__GCA-2721815;s__GCA-2721815

n245 d__Bacteria;p__Myxococcota;c__Polyangia;o__Haliangiales;f__Haliangiaceae;g__Haliangium;s__Haliangium

n246 d__Bacteria;p__Myxococcota;c__Polyangia;o__Nannocystales;f__Nannocystaceae;g__Ga0077550;s__Ga0077550

n247 d__Bacteria;p__Myxococcota;c__Bradimonadia;o__Bradymonadales;f__Bradymonadaceae;g__Bradymonas;s__Bradymonas

n248 d__Bacteria;p__Myxococcota;c__UBA796;o__UBA796;f__UBA796;g__UBA2385;s__UBA2385

n249 d__Bacteria;p__Myxococcota;c__UBA796;o__UBA9615;f__UBA9615;g__UBA6601;s__UBA6601

n250 d__Bacteria;p__Myxococcota;c__UBA9160;o__UBA9160;f__UBA4427;g__UBA4427;s__UBA4427

n251 d__Bacteria;p__Bdellovibrionota;c__Bacteriovoracia;o__Bacteriovoracales;f__Bacteriovoracaceae;g__21-14-all-39-27;s__21-14-all-39-27

n252 d__Bacteria;p__Bdellovibrionota;c__Bacteriovoracia;o__UBA1018;f__UBA923;g__UBA923;s__UBA923

n253 d__Bacteria;p__Bdellovibrionota;c__UBA2394;o__UBA2394;f__UBA2394;g__1-14-0-20-45-16;s__1-14-0-20-45-16

n254 d__Bacteria;p__Bdellovibrionota;c__Bdellovibrionia;o__Bdellovibrionales;f__1-14-0-10-45-34;g__1-14-0-10-45-34;s__1-14-0-10-45-34

n255 d__Bacteria;p__Desulfobacterota_A;c__Desulfovibrionia;o__Desulfovibrionales;f__Desulfoplanaceae;g__Desulfoplanes;s__Desulfoplanes

n256 d__Bacteria;p__Omnitrophota;c__koll11;o__UBA1572;f__UBA1572;g__UBA6210;s__UBA6210

n257 d__Bacteria;p__Omnitrophota;c__koll11;o__UBA10183;f__UBA10183;g__UBA10183;s__UBA10183

n258 d__Bacteria;p__Omnitrophota;c__Omnitrophia;o__Omnitrophales;f__2-12-FULL-44-17;g__2-12-FULL-44-17;s__2-12-FULL-44-17

n259 d__Bacteria;p__Elusimicrobiota;c__Elusimicrobia;o__UBA1565;f__UBA9628;g__UBA9628;s__UBA9628

n260 d__Bacteria;p__Elusimicrobiota;c__Elusimicrobia;o__F11;f__FEN-1173;g__FEN-1173;s__FEN-1173

n261 d__Bacteria;p__Elusimicrobiota;c__Endomicrobia;o__PHAN01;f__PHAN01;g__PHAN01;s__PHAN01

n262 d__Bacteria;p__Elusimicrobiota;c__Endomicrobia;o__CG1-02-37-114;f__CG1-02-37-114;g__CG1-02-37-114;s__CG1-02-37-114

n263 d__Bacteria;p__Elusimicrobiota;c__UBA5214;o__UBA5214;f__UBA5214;g__UBA5214;s__UBA5214

n264 d__Bacteria;p__Elusimicrobiota;c__UBA8919;o__UBA8919;f__UBA8919;g__UBA8919;s__UBA8919

Table 2 Taxon sampling for the GTDB-independent dataset, Chapter 2

Species code	NCBI Taxonomy
00-UBP1	Bacteria_Candidatus_UBP1_bacterium_UBA2172
00-UBP10	Bacteria_Candidatus_UBP10_bacterium_UBA1160
00-UBP11	Bacteria_Candidatus_UBP11_bacterium_UBA4055
00-UBP12	Bacteria_Candidatus_UBP12_bacterium_UBA5184
00-UBP13	Bacteria_Candidatus_UBP13_bacterium_UBA5359
00-UBP14	Bacteria_Candidatus_UBP14_bacterium_UBA6098
00-UBP15	Bacteria_Candidatus_UBP15_bacterium_UBA6099
00-UBP16	Bacteria_Candidatus_UBP16_bacterium_UBA6123
00-UBP17	Bacteria_Candidatus_UBP17_bacterium_UBA6191
00-UBP2	Bacteria_Candidatus_UBP2_bacterium_UBA2255
00-UBP3	Bacteria_Candidatus_UBP3_bacterium_UBA1439
00-UBP4	Bacteria_Candidatus_UBP4_bacterium_UBA6127
00-UBP5	Bacteria_Candidatus_UBP5_bacterium_UBA1559
00-UBP6	Bacteria_Candidatus_UBP6_bacterium_UBA1177
00-UBP7	Bacteria_Candidatus_UBP7_bacterium_UBA6624
00-UBP8	Bacteria_Candidatus_UBP8_bacterium_UBA6595
00-UBP9	Bacteria_Candidatus_UBP9_bacterium_UBA1085
ACID1	Bacteria_Acidobacteria_RBG_16_Acidobacteria_68_9
ACID2	Bacteria_Acidobacteria_RIFCSPLOWO2_02_FULL_Acidobacteria_68_18
ACTI1	Bacteria_Actinobacteria_Actinobacteria_Acidimicrobidae_Acidimicrobiales_Acidimicrobineae_Acidimicrobiaceae_Acidimicrobium_ferrooxidans_DSM_10331
ACTI10	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Actinomycetales_Propionibacterineae_Propionibacteriaceae_Aestuariimicrobium_kwangyangense_DSM_21549
ACTI11	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Actinomycetales_Pseudonocardineae_Pseudonocardiaceae_Prauserella_rugosa_DSM_43194
ACTI12	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Actinomycetales_Streptomycineae_Streptomycetaceae_Streptomycetaceae_bacterium_MP113_05
ACTI13	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Actinomycetales_Streptosporangineae_Thermomonosporaceae_Thermomonospora_curvata_DSM_43183
ACTI14	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Bifidobacteriales_Bifidobacteriaceae_Bifidobacterium_animalis_animalis_ATCC_25527
ACTI15	Bacteria_Actinobacteria_Actinobacteria_Coriobacteridae_Coriobacteriales_Coriobacterineae_Coriobacteriaceae_Eggerthella_sp._YY7918
ACTI16	Bacteria_Actinobacteria_Actinobacteria_Micrococcales_Microbacteriaceae_Candidatus_Rhodoluna_lacicola_MWH_Ta8_Sequence_finished_Aug._2010
ACTI17	Bacteria_Actinobacteria_Coriobacteriia_Coriobacteriales_Coriobacteriaceae_Enorma_massiliensis_phI
ACTI2	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Actinomycetales_Actinomycineae_Actinomycetaceae_Arcanobacterium_haemolyticum_DSM_20595
ACTI3	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Actinomycetales_Corynebacterineae_Corynebacteriaceae_Corynebacterium_argentoratense_DSM_44202
ACTI4	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Actinomycetales_Frankineae_Geodermatophilaceae_Blastococcus_saxobsidens_DD2
ACTI5	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Actinomycetales_Micrococcineae_Intrasporangiaceae_Serinicoccus_marinus_MCCC_1A05965
ACTI6	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Actinomycetales_Micrococcineae_Jonesiaceae_Jonesia_denitrificans_DSM_20603
ACTI7	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Actinomycetales_Micrococcineae_Microbacteriaceae_Cryobacterium_sp._MLB_32
ACTI8	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Actinomycetales_Micrococcineae_Microbacteriaceae_Gulosibacter_molinativorax_DSM_13485
ACTI9	Bacteria_Actinobacteria_Actinobacteria_Actinobacteridae_Actinomycetales_Micrococcineae_Microbacteriaceae_Rathayibacter_toxicus_DSM_7488

AMIN1	Bacteria_Aminicenantes_OP8_RBG_16_Aminicenantes_66_30
AQUI1	Bacteria_Aquificae_Aquificae_Aquificales_Aquificaceae_Hydrogenobacter_thermophilus_TK_6
AQUI2	Bacteria_Aquificae_Aquificae_Aquificales_Desulfurobacteriaceae_Thermovibrio_ammonificans_HB_1
ARMA1	Bacteria_Armatimonadetes_13_1_40CM_Armatimonadetes_64_14
BACT1	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Bacteroidia_Bacteroidales_Bacteroidaceae_Bacteroides_fragilis_NCTC_9343
BACT10	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Flavobacteriia_Flavobacteriales_Flavobacteriaceae_Galbibacter_sp._ck_I2_15
BACT11	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Flavobacteriia_Flavobacteriales_Flavobacteriaceae_Mesoflavibacter_zeaxanthinifaciens_DSM_18436
BACT12	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Flavobacteriia_Flavobacteriales_Flavobacteriaceae_Polaribacter_sp._MED152
BACT13	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Flavobacteriia_Flavobacteriales_Flavobacteriaceae_Weeksella_virosa_DSM_16922
BACT14	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Flavobacteriia_Flavobacteriales_Flavobacteriaceae_Zunongwangia_profunda_SM_A87
BACT15	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Sphingobacteriia_Sphingobacteriales_Chitinophagaceae_Gracilimonas_tropica_DSM_19535
BACT16	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Sphingobacteriia_Sphingobacteriales_Chitinophagaceae_Terrimonas_ferruginea_DSM_30193
BACT17	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Sphingobacteriia_Sphingobacteriales_Saprospiraceae_Aureispira_sp._CCB_QB1
BACT18	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Sphingobacteriia_Sphingobacteriales_Sphingobacteriaceae_Arcticibacter_svalbardensis_MN12_7
BACT19	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_unclassified_Bacteroidetes_Prolixibacter_bellariivorans_ATCC_BAA_1284
BACT2	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Bacteroidia_Bacteroidales_Marinilabiaceae_Marinilabilia_salmonicolor_JCM_21150
BACT3	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Bacteroidia_Bacteroidales_Porphyromonadaceae_Butyricimonas_virosa_DSM_23226
BACT4	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Bacteroidia_Bacteroidales_Porphyromonadaceae_Parabacteroides_distasonis_ATCC_8503
BACT5	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Cytophagia_Cytophagales_Cyclobacteriaceae_Cyclobacterium_marinum_DSM_745
BACT6	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Cytophagia_Cytophagales_Cyclobacteriaceae_Indibacter_alkaliphilus_LW1_Draft1
BACT7	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Cytophagia_Cytophagales_Cytophagaceae_Fibrella_aestuarina
BACT8	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Cytophagia_Cytophagales_Flammeovirgaceae_Cesiribacter_andamanensis_AMV16
BACT9	Bacteria_Bacteroidetes_Chlorobi_group_Bacteroidetes_Flavobacteriia_Flavobacteriales_Flavobacteriaceae_Epilithonimonas_tenax_DSM_16811
BCHL1	Bacteria_Bacteroidetes_Chlorobi_group_Chlorobi_Chlorobia_Chlorobiales_Chlorobiaceae_Chloroherpeton_thalassium_ATCC_35110
CHLA1	Bacteria_Chlamydiae_Chlamydiia_RIFCSPLOWO2_02_FULL_Chlamydiae_45_22
CHLA2	Bacteria_Chlamydiae_Verrucomicrobia_group_Chlamydiae_Chlamydiia_Chlamydiales_Chlamydiaceae_Chlamydia Chlamydophila_group_Chlamydophila_abortus_S263
CHLAV1	Bacteria_Chlamydiae_Verrucomicrobia_group_Verrucomicrobia_Verrucomicrobiae_Verrucomicrobiales_Verrucomicrobiaceae_Rubritalea_marina_DSM_17716_SAORIC_165
CHLO1	Bacteria_Chloroflexi_Anaerolineae_Anaerolineales_Anaerolineaceae_Anaerolinea_thermophila_UNI_1
CHLO2	Bacteria_Chloroflexi_Dehalococcoidia_uncultured_DG_22
CHLO3	Bacteria_Chloroflexi_RBG_13_Chloroflexi_50_10
CHLO4	Bacteria_Chloroflexi_RBG_13_Chloroflexi_52_12
CHLO5	Bacteria_Chloroflexi_RBG_16_Chloroflexi_54_11
CHLO6	Bacteria_Chloroflexi_RBG_16_Chloroflexi_64_43
CHRY1	Bacteria_Chrysiogenetes_Chrysiogenetes_Chrysiogenales_Chrysiogenaceae_Chrysiogenes_arsenatis_DSM_11915
CP1	Bacteria_CP_RIF26_DOLJORAL78_RIF26_32_13
CP2	Bacteria_CP_RIF32_CG2_30_FULL_CP_RIF32_54_10
CP3	Bacteria_CP_TM6_GWF2_TM6_38_10
CPR21	Bacteria_CPR2_CG2_30_FULL_CPR2_33_46

CPR22	Bacteria_CPR2_CG2_30_FULL_CPR2_37_16
CPR23	Bacteria_CPR2_GWC2_CPR2_39_10
CPRBERK1	Bacteria_CPR_Berkelbacteria_ACD58_GWA2_ACD58_46_7
CPRBERK2	Bacteria_CPR_Berkelbacteria_ACD58_GWA2_Berkelbacteria_35_9
CPRBERK3	Bacteria_CPR_Berkelbacteria_ACD58_GWA2_Berkelbacteria_38_9
CPRBERK4	Bacteria_CPR_Berkelbacteria_ACD58_GWE1_Berkelbacteria_39_12_complete
CPRBERK5	Bacteria_CPR_Berkelbacteria_CG1_02_FULL_Berkelbacteria_42_45
CPRBERK6	Bacteria_CPR_Berkelbacteria_CG2_30_FULL_Berkelbacteria_39_44
CPRBERK7	Bacteria_CPR_Berkelbacteria_CG2_30_FULL_Berkelbacteria_43_20
CPRBERK8	Bacteria_CPR_Berkelbacteria_RIFCSPLOWO2_01_FULL_Berkelbacteria_50_28
CPRBERK9	Bacteria_CPR_Berkelbacteria_RIFOXYA2_FULL_Berkelbacteria_43_10
CPRCPR1	Bacteria_CPR_CPR1_GWA2_CPR1_42_17
CPRCPR2	Bacteria_CPR_CPR3_CG1_02_FULL_CPR3_42_9
CPRCPR3	Bacteria_CPR_CPR3_GWF2_CPR3_35_18
CPRDOJK1	Bacteria_CPR_Dojkabacteria_WS6_GWA2_WS6_37_6
CPRDOJK2	Bacteria_CPR_Dojkabacteria_WS6_GWE1_WS6_34_7
CPRDOJK3	Bacteria_CPR_Dojkabacteria_WS6_GWF1_WS6_35_23_partial
CPRKAZA1	Bacteria_CPR_Kazan_GWA1_Kazan_50_15_complete
CPRKAZA2	Bacteria_CPR_Kazan_GWA1_Kazan_rel_44_22_partial
CPRKAZA3	Bacteria_CPR_Kazan_RBG_13_KAZAN_50_9
CPRMICR1	Bacteria_CPR_Microgenomates_OP11_Amesbacteria_RIFOXYB1_FULL_OP11_Amesbacteria_44_23
CPRMICR10	Bacteria_CPR_Microgenomates_OP11_Gottesmanbacteria_RIFCSPHIGHO2_01_FULL_OP11_Gottesmanbacteria_42_12
CPRMICR11	Bacteria_CPR_Microgenomates_OP11_Gottesmanbacteria_RIFCSPHIGHO2_02_FULL_OP11_Gottesmanbacteria_39_14
CPRMICR12	Bacteria_CPR_Microgenomates_OP11_Gottesmanbacteria_RIFCSPLOWO2_01_FULL_OP11_Gottesmanbacteria_39_12
CPRMICR13	Bacteria_CPR_Microgenomates_OP11_Gottesmanbacteria_RIFCSPLOWO2_01_FULL_OP11_Gottesmanbacteria_42_22
CPRMICR14	Bacteria_CPR_Microgenomates_OP11_GWA2_OP11_40_6
CPRMICR15	Bacteria_CPR_Microgenomates_OP11_GWA2_OP11_46_16_partial
CPRMICR16	Bacteria_CPR_Microgenomates_OP11_GWB1_OP11_45_5
CPRMICR17	Bacteria_CPR_Microgenomates_OP11_GWF1_OP11_44_10
CPRMICR18	Bacteria_CPR_Microgenomates_OP11_Levybacteria_CG2_30_FULL_Levybacteria_OP11_37_29
CPRMICR19	Bacteria_CPR_Microgenomates_OP11_Levybacteria_RIFCSPHIGHO2_02_FULL_OP11_Levybacteria_37_10
CPRMICR2	Bacteria_CPR_Microgenomates_OP11_Beckwithbacteria_GWC1_OP11_49_16_complete
CPRMICR20	Bacteria_CPR_Microgenomates_OP11_Pacebacteria_CG1_02_FULL_Pacebacteria_OP11_43_31
CPRMICR21	Bacteria_CPR_Microgenomates_OP11_RBG_13_OP11_40_15
CPRMICR22	Bacteria_CPR_Microgenomates_OP11_RBG_13_OP11_40_7b
CPRMICR23	Bacteria_CPR_Microgenomates_OP11_RIF35_RIFCSPHIGHO2_12_FULL_RIF35_44_25
CPRMICR24	Bacteria_CPR_Microgenomates_OP11_RIF36_RIFCSPLOWO2_01_FULL_RIF36_49_14
CPRMICR25	Bacteria_CPR_Microgenomates_OP11_RIFCSPHIGHO2_01_FULL_OP11_45_11
CPRMICR26	Bacteria_CPR_Microgenomates_OP11_RIFCSPLOWO2_01_FULL_OP11_43_14
CPRMICR27	Bacteria_CPR_Microgenomates_OP11_Roizmannbacteria_RIFCSPHIGHO2_01_FULL_OP11_Roizmanbacteria_39_12b
CPRMICR28	Bacteria_CPR_Microgenomates_OP11_Roizmannbacteria_RIFCSPLOWO2_01_FULL_OP11_Roizmanbacteria_45_11
CPRMICR29	Bacteria_CPR_Microgenomates_OP11_Roizmannbacteria_RIFOXYA1_FULL_OP11_Roizmanbacteria_41_12

CPRMICR3	Bacteria_CPR_Microgenomates_OP11_Collierbacteria_RIFOXYB1_FULL_OP11_Collierbacteria_49_13
CPRMICR30	Bacteria_CPR_Microgenomates_OP11_Shapirobacteria_GWF2_OP11_rel_37_20
CPRMICR31	Bacteria_CPR_Microgenomates_OP11_Shapirobacteria_Microgenomates_bacterium_SCGC_AAA255_J07_SAK_001_132
CPRMICR32	Bacteria_CPR_Microgenomates_OP11_Woesebacteria_GWD2_OP11_40_19
CPRMICR33	Bacteria_CPR_Microgenomates_OP11_Woesebacteria_RIFCSPHIGHO2_01_FULL_OP11_Woesebacteria_41_10
CPRMICR34	Bacteria_CPR_Microgenomates_OP11_Woesebacteria_RIFOXYA1_FULL_OP11_Woesebacteria_43_9
CPRMICR4	Bacteria_CPR_Microgenomates_OP11_Curtissbacteria_RBG_16_OP11_Curtissbacteria_39_7
CPRMICR5	Bacteria_CPR_Microgenomates_OP11_Curtissbacteria_RIFCSPHIGHO2_12_FULL_OP11_Curtissbacteria_41_17
CPRMICR6	Bacteria_CPR_Microgenomates_OP11_Curtissbacteria_RIFCSPLOWO2_01_FULL_OP11_Curtissbacteria_42_26
CPRMICR7	Bacteria_CPR_Microgenomates_OP11_Daviesbacteria_RIFCSPHIGHO2_12_FULL_OP11_Daviesbacteria_37_16
CPRMICR8	Bacteria_CPR_Microgenomates_OP11_Gottesmanbacteria_GWA1_OP11_43_11
CPRMICR9	Bacteria_CPR_Microgenomates_OP11_Gottesmanbacteria_GWA2_OP11_44_17_partial
CPRPARC1	Bacteria_CPR_Parcubacteria_OD1_CG1_02_FULL_Parcubacteria_OD1_36_42
CPRPARC10	Bacteria_CPR_Parcubacteria_OD1_GWA2_OD1_43_17_A823
CPRPARC11	Bacteria_CPR_Parcubacteria_OD1_GWB1_OD1_49_12
CPRPARC12	Bacteria_CPR_Parcubacteria_OD1_GWB1_OD1_rel_46_8
CPRPARC13	Bacteria_CPR_Parcubacteria_OD1_GWE2_OD1_38_18
CPRPARC14	Bacteria_CPR_Parcubacteria_OD1_Jorgensenbacteria_GWA1_OD1_Jorgensenbacteria_54_12
CPRPARC15	Bacteria_CPR_Parcubacteria_OD1_Jorgensenbacteria_RIFCSPHIGHO2_02_FULL_OD1_Jorgensenbacteria_45_20
CPRPARC16	Bacteria_CPR_Parcubacteria_OD1_Kaiserbacteria_RIFCSPHIGHO2_01_FULL_OD1_Kaiserbacteria_56_24
CPRPARC17	Bacteria_CPR_Parcubacteria_OD1_L1_Parcubacteria_bacterium_SCGC_AAA011_A09_DUSEL_001_189
CPRPARC18	Bacteria_CPR_Parcubacteria_OD1_Magasanikbacteria_GWA2_OD1_45_39_plus
CPRPARC19	Bacteria_CPR_Parcubacteria_OD1_Magasanikbacteria_GWA2_OD1_46_17_plus
CPRPARC2	Bacteria_CPR_Parcubacteria_OD1_CG1_02_FULL_Parcubacteria_OD1_41_12
CPRPARC20	Bacteria_CPR_Parcubacteria_OD1_Moranbacteria_RIFCSPHIGHO2_01_FULL_OD1_Moranbacteria_55_24
CPRPARC21	Bacteria_CPR_Parcubacteria_OD1_MoranbacteriaOD1_i_GWC2_OD1_i_37_82
CPRPARC22	Bacteria_CPR_Parcubacteria_OD1_Nomurabacteria_CG1_02_FULL_Nomurabacteria_OD1_31_12
CPRPARC23	Bacteria_CPR_Parcubacteria_OD1_Nomurabacteria_RIFCSPLOWO2_01_FULL_OD1_Nomurabacteria_36_10b
CPRPARC24	Bacteria_CPR_Parcubacteria_OD1_Parcubacteria_bacterium_SCGC_AAA011_N16_Dusel_001_262
CPRPARC25	Bacteria_CPR_Parcubacteria_OD1_Parcubacteria_bacterium_SCGC_AB_164_E21_SAK_001_209
CPRPARC26	Bacteria_CPR_Parcubacteria_OD1_RBG_16_OD1_47_10
CPRPARC27	Bacteria_CPR_Parcubacteria_OD1_RIF-OD1-2_RIFCSPLOWO2_01_FULL_RIF_OD1_02_44_13
CPRPARC28	Bacteria_CPR_Parcubacteria_OD1_RIF-OD1-2_RIFOXYD2_FULL_RIF_OD1_02_43_21
CPRPARC29	Bacteria_CPR_Parcubacteria_OD1_RIF-OD1-9_RIFCSPLOWO2_02_FULL_RIF_OD1_09_51_11
CPRPARC3	Bacteria_CPR_Parcubacteria_OD1_CG2_30_FULL_Parcubacteria_OD1_36_38
CPRPARC30	Bacteria_CPR_Parcubacteria_OD1_RIF10_RIFCSPLOWO2_01_FULL_RIF10_44_230
CPRPARC31	Bacteria_CPR_Parcubacteria_OD1_RIF15_RIFCSPLOWO2_02_RIF15_39_10
CPRPARC32	Bacteria_CPR_Parcubacteria_OD1_RIF16_RIFCSPLOWO2_12_FULL_RIF16_43_20
CPRPARC33	Bacteria_CPR_Parcubacteria_OD1_RIF17_RIFCSPLOWO2_01_FULL_RIF17_60_25
CPRPARC34	Bacteria_CPR_Parcubacteria_OD1_RIF19_RIFCSPLOWO2_01_FULL_RIF19_46_10
CPRPARC35	Bacteria_CPR_Parcubacteria_OD1_RIF22_RBG_13_RIF22_41_18
CPRPARC36	Bacteria_CPR_Parcubacteria_OD1_RIF4_RIFCSPHIGHO2_12_FULL_RIF4_48_17
CPRPARC37	Bacteria_CPR_Parcubacteria_OD1_RIF4_RIFCSPLOWO2_01_FULL_RIF4_48_11

CPRPARC38	Bacteria_CPR_Parcubacteria_OD1_RIF4_RIFCSPLOWO2_02_FULL_RIF4_42_19
CPRPARC39	Bacteria_CPR_Parcubacteria_OD1_RIF6_RIFCSPLOWO2_01_FULL_RIF6_53_11
CPRPARC4	Bacteria_CPR_Parcubacteria_OD1_Falkowbacteria_GWE2_OD1_38_254
CPRPARC40	Bacteria_CPR_Parcubacteria_OD1_RIF6_RIFOXYC2_FULL_RIF6_42_11
CPRPARC41	Bacteria_CPR_Parcubacteria_OD1_RIF9_RIFCSPHIGHO2_01_FULL_RIF9_46_36
CPRPARC42	Bacteria_CPR_Parcubacteria_OD1_RIF9_RIFCSPHIGHO2_02_FULL_RIF9_46_16
CPRPARC43	Bacteria_CPR_Parcubacteria_OD1_RIFCSPHIGHO2_01_FULL_OD1_47_10b
CPRPARC44	Bacteria_CPR_Parcubacteria_OD1_RIFCSPLOWO2_01_FULL_OD1_45_48
CPRPARC45	Bacteria_CPR_Parcubacteria_OD1_RIFCSPLOWO2_01_FULL_OD1_46_25
CPRPARC46	Bacteria_CPR_Parcubacteria_OD1_RIFCSPLOWO2_01_FULL_OD1_48_14
CPRPARC47	Bacteria_CPR_Parcubacteria_OD1_RIFCSPLOWO2_01_FULL_OD1_48_18
CPRPARC48	Bacteria_CPR_Parcubacteria_OD1_Uhrbacteria_RIFCSPLOWO2_02_FULL_OD1_Uhrbacteria_49_11
CPRPARC49	Bacteria_CPR_Parcubacteria_OD1_Uhrbacteria_RIFCSPLOWO2_02_FULL_OD1_Uhrbacteria_54_37
CPRPARC5	Bacteria_CPR_Parcubacteria_OD1_Falkowbacteria_RIFCSPLOWO2_12_FULL_OD1_Falkowbacteria_42_13
CPRPARC50	Bacteria_CPR_Parcubacteria_OD1_Uhrbacteria_RIFOXYC2_FULL_OD1_Uhrbacteria_47_19
CPRPARC51	Bacteria_CPR_Parcubacteria_OD1_Wolfebacteria_RIFOXYB1_FULL_OD1_Wolfebacteria_54_12
CPRPARC52	Bacteria_CPR_Parcubacteria_OD1_Yanofskybacteria_RIFCSPHIGHO2_01_FULL_OD1_Yanofskybacteria_44_17
CPRPARC53	Bacteria_CPR_Parcubacteria_OD1_Yanofskybacteria_RIFCSPHIGHO2_01_FULL_OD1_Yanofskybacteria_44_22
CPRPARC6	Bacteria_CPR_Parcubacteria_OD1_Falkowbacteria_RIFOXYC2_FULL_OD1_Falkowbacteria_36_12
CPRPARC7	Bacteria_CPR_Parcubacteria_OD1_Giovannonibacteria_RIFCSPHIGHO2_02_OD1_Giovannonibacteria_43_16
CPRPARC8	Bacteria_CPR_Parcubacteria_OD1_GWA2_OD1_31_28_partial
CPRPARC9	Bacteria_CPR_Parcubacteria_OD1_GWA2_OD1_37_10
CPRPER1	Bacteria_CPR_PER-ii_GWB1_PER_54_5
CPRPER2	Bacteria_CPR_Peregrinibacteria_CG1_02_FULL_Peregrinibacteria_PER_41_10
CPRPER3	Bacteria_CPR_Peregrinibacteria_CG1_02_FULL_Peregrinibacteria_PER_54_53
CPRPER4	Bacteria_CPR_PeregrinibacteriaPER_GWA2_PER_44_7
CPRPER5	Bacteria_CPR_PeregrinibacteriaPER_PER_GWC2_39_14
CPRPER6	Bacteria_CPR_PeregrinibacteriaPER_PER_GWF2_33_10
CPRSACC1	Bacteria_CPR_Saccharibacteria_TM7_Candidatus_Saccharimonas_aalborgensis_complete
CPRSACC2	Bacteria_CPR_Saccharibacteria_TM7_CG2_30_FULL_Saccharibacteria_TM7_41_52
CPRSACC3	Bacteria_CPR_Saccharibacteria_TM7_RIFCSPHIGHO2_12_FULL_Saccharibacteria_49_19
CPRUNC1	Bacteria_CPR_unclassified_bacteria_RBG_16_CPR_42_10
CPRUNC2	Bacteria_CPR_unclassified_bacteria_RIFCSPHIGHO2_01_FULL_PER_46_8
CPRWWE31	Bacteria_CPR_WWE3_CG2_30_FULL_WWE3_related_40_12
CPRWWE32	Bacteria_CPR_WWE3_CSP1_7
CPRWWE33	Bacteria_CPR_WWE3_RBG_16_WWE3_37_10
CPRWWE34	Bacteria_CPR_WWE3_RIFCSPHIGHO2_01_FULL_WWE3_48_15
CPRWWE35	Bacteria_CPR_WWE3_RIFCSPHIGHO2_12_FULL_WWE3_38_15
CPRWWE36	Bacteria_CPR_WWE3_RIFCSPLOWO2_01_FULL_WWE3_39_13
CPRWWE37	Bacteria_CPR_WWE3_RIFCSPLOWO2_01_FULL_WWE3_42_11
CYAN1	Bacteria_Cyanobacteria_Chroococcales_Crocosphaera_watsonii_WH_8501
CYAN2	Bacteria_Cyanobacteria_Chroococcales_Synechococcus_sp._CC9902
CYAN3	Bacteria_Cyanobacteria_Chroococcales_Thermosynechococcus_elongatus_BP_1

CYAN4	Bacteria_Cyanobacteria_Nostocales_Nostocaceae_Aphanizomenon_flos_aquae_NIES_81
CYAN5	Bacteria_Cyanobacteria_Oscillatoriales_Crinalium_epipsammum_PCC_9333
CYAN6	Bacteria_Cyanobacteria_Oscillatoriales_Oscillatoria_sp._PCC_7112
CYAN7	Bacteria_Cyanobacteria_Pleurocapsales_Stanieria_cyanosphaera_PCC_7437
DEFE1	Bacteria_Deferribacteres_Deferribacteres_Deferribacterales_Deferribacteraceae_Denitrovibrio_acetiphilus_DSM_12809
DEIN1	Bacteria_Deinococcus_Thermus_Deinococci_Deinococcales_Deinococcaceae_Deinococcus_geothermalis_DSM_11300
DEIN2	Bacteria_Deinococcus_Thermus_Deinococci_Thermales_Thermaceae_Marinithermus_hydrothermalis_T1_DSM_14884
ELUS1	Bacteria_Elusimicrobia_GWC2_Elusimicrobia_61_19
ELUS2	Bacteria_Elusimicrobia_GWC2_Elusimicrobia_65_9
ELUS3	Bacteria_Elusimicrobia_RIFCSPLOWO2_01_FULL_Elusimicrobia_60_11
ELUS4	Bacteria_Elusimicrobia_RIFOXYB2_FULL_Elusimicrobia_48_7
ELUS5	Bacteria_Elusimicrobia_RIFOXYB2_FULL_Elusimicrobia_62_6
FIBR1	Bacteria_Fibrobacteres_Acidobacteria_Fibrobacteres_Fibrobacteria_Fibrobacterales_CG2_30_FULL_Fibrobacteres_45_31
FIBR2	Bacteria_FibrobacteresAcidobacteria_group_Acidobacteria_Acidobacteriia_Acidobacteriales_Acidobacteriaceae_Acidobacterium_sp._MP5ACTX8
FIBR3	Bacteria_FibrobacteresAcidobacteria_group_Acidobacteria_Holophagae_Holophagales_Holophagaceae_Holophaga_foetida_TMBS4_DSM_6591
FIRM1	Bacteria_Firmicutes_Bacilli_Bacillales_Bacillaceae_Bacillus_anthracis_52_G
FIRM10	Bacteria_Firmicutes_Clostridia_Clostridiales_Clostridiaceae_Youngiibacter_fragile_232.1
FIRM11	Bacteria_Firmicutes_Clostridia_Clostridiales_Clostridiales_incertae_sedis_Clostridiales_Family_XIII._Incertae_Sedis_Mogibacterium_timidum_ATCC_33093
FIRM12	Bacteria_Firmicutes_Clostridia_Clostridiales_Clostridiales_incertae_sedis_Clostridiales_Family_XVII._Incertae_Sedis_Thermaerobacter_marianensis_DSM_12885
FIRM13	Bacteria_Firmicutes_Clostridia_Clostridiales_Lachnospiraceae_Dorea_longicatena_AGR2136
FIRM14	Bacteria_Firmicutes_Clostridia_Clostridiales_Lachnospiraceae_Robinsoniella_sp._KNHs210
FIRM15	Bacteria_Firmicutes_Clostridia_Clostridiales_Peptococcaceae_Pelotomaculum_thermopropionicum_SI
FIRM16	Bacteria_Firmicutes_Clostridia_Clostridiales_Peptococcaceae_Syntrophobotulus_glycolicus_DSM_8271
FIRM17	Bacteria_Firmicutes_Clostridia_Clostridiales_Peptostreptococcaceae_Filifactor_alocis_ATCC_35896
FIRM18	Bacteria_Firmicutes_Clostridia_Clostridiales_Ruminococcaceae_Acetivibrio_cellulolyticus_CD2_DSM_1870_ORNL_annotation
FIRM19	Bacteria_Firmicutes_Clostridia_Clostridiales_Ruminococcaceae_Subdoligranulum_sp._4_3_54A2FAA
FIRM2	Bacteria_Firmicutes_Bacilli_Bacillales_Bacillaceae_Marinococcus_halotolerans_DSM_16375
FIRM20	Bacteria_Firmicutes_Clostridia_Halanaerobiales_Halobacteroidaceae_Orenia_marismortui_DSM_5156
FIRM21	Bacteria_Firmicutes_Clostridia_Thermoanaerobacterales_Thermoanaerobacterales_Family_III._Incertae_Sedis_Thermoanaerobacterium_thermosaccharolyticum_M0795
FIRM22	Bacteria_Firmicutes_Negativicutes_Selenomonadales_Veillonellaceae_Anaerovibrio_sp._RM50
FIRM23	Bacteria_Firmicutes_Negativicutes_Selenomonadales_Veillonellaceae_Veillonella_parvula_DSM_2008
FIRM3	Bacteria_Firmicutes_Bacilli_Bacillales_Bacillaceae_Terribacillus_aidingensis_MP602
FIRM4	Bacteria_Firmicutes_Bacilli_Bacillales_Staphylococcaceae_Staphylococcus_aureus_502A
FIRM5	Bacteria_Firmicutes_Bacilli_Bacillales_Thermoactinomycetaceae_Thermoactinomyces_sp._Gus2_1
FIRM6	Bacteria_Firmicutes_Bacilli_Lactobacillales_Aerococcaceae_Globicatella_sulfidifaciens_DSM_15739
FIRM7	Bacteria_Firmicutes_Bacilli_Lactobacillales_Lactobacillaceae_Sharpea_azabuensis_DSM_18934
FIRM8	Bacteria_Firmicutes_Bacilli_Lactobacillales_Leuconostocaceae_Fructobacillus_fructosus_KCTC_3544_contig00014
FIRM9	Bacteria_Firmicutes_Bacilli_Lactobacillales_Streptococcaceae_Streptococcus_mutans_GS_5
FUSO1	Bacteria_Fusobacteria_Fusobacteriia_Fusobacteriales_Fusobacteriaceae_Fusobacterium_nucleatum_nucleatum_ATCC_25586
FUSO2	Bacteria_Fusobacteria_Fusobacteriia_Fusobacteriales_Leptotrichiaceae_Leptotrichia_goodfellowii_F0264_contig00021

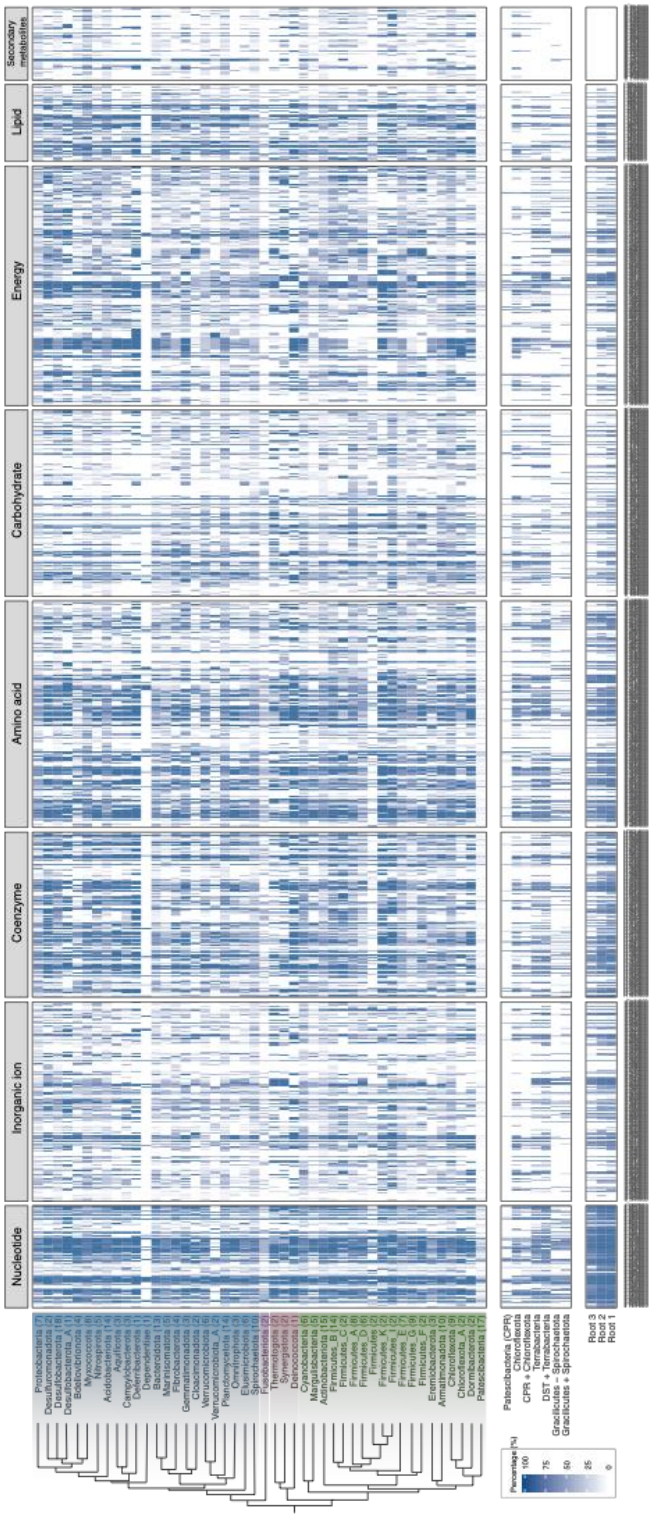
GEMA1	Bacteria_Gemmatimonadetes_RIFCSPLOWO2_12_FULL_Gemmatimonadetes_68_9
GEMM1	Bacteria_Gemmatiomonas_uncultured_SG8_23
GN021	Bacteria_GN02_CG1_02_SUB10_Gracilibacteria_GN02_38_174
GN022	Bacteria_GN02_CG2_30_FULL_Gracilibacteria_GN02_37_12
IGNA1	Bacteria_Ignavibacteria_GWA2_Ignavibacteriae_55_25
IGNA2	Bacteria_Ignavibacteria_RBG_16_Ignavibacteria_34_14
LENT1	Bacteria_Lentisphaerae_GWF2_Lentisphaerae_49_21
LENT2	Bacteria_Lentisphaerae_GWF2_Lentisphaerae_52_8
MARG1	Bacteria_Marine_group_A_SAR406_SAR406_cluster_bacterium_JGI_0000039_D08_Combined_Assembly_SAR406_1__SAR406
MELA1	Bacteria_Melainabacteria_GWA2_Melainabacteria_34_9
MELA2	Bacteria_Melainabacteria_MEL_A1
MELA3	Bacteria_Modulibacteria_KSB3_bacterium_UASB270
MODU1	Bacteria_Modulibacteria_KSB3_bacterium_UASB270
NITRN1	Bacteria_Nitrospinae_RIFCSPLOWO2_12_FULL_Nitrospinae_47_7
NITRN2	Bacteria_Nitrospinae_RifCSPlowO2_12_Nitrospinae_39_15
NITRR1	Bacteria_Nitrospirae_CG1_02_FULL_Nitrospirae_44_142
NITRR2	Bacteria_Nitrospirae_GWC2_Nitrospirae_57_9
NITRR3	Bacteria_Nitrospirae_Nitrospira_Candidatus_Nitrospira_defluvii
OMNI1	Bacteria_Omnitrophica_CG1_02_FULL_Omnitrophica_46_14
OMNI2	Bacteria_Omnitrophica_RIFCSPLOWO2_01_FULL_Omnitrophica_50_24
OMNI3	Bacteria_Omnitrophica_WOR-2_RIFCSPHIGHO2_02_FULL_WOR_2_48_11
PLAN1	Bacteria_Planctomycetes_Brocadiales_GWB2_Planctomycetes_41_19
PLAN2	Bacteria_Plantomycetes_Phycisphaerae_uncultured_SMTZ1_79
PLAN3	Bacteria_Plantomycetes_uncultured_DG_23
PROA1	Bacteria_Proteobacteria_Acidithiobacillia_Acidithiobacillales_Thermithiobacillaceae_Thermithiobacillus_tepidarius_DSM_3134
PROA10	Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Rhodovulum_sp._PH10
PROA11	Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Roseobacter_Thalassiobium_sp._R2A62_genomic_scaffold_scf_1112329232034
PROA12	Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Rubellimicrobium_thermophilum_DSM_16684
PROA13	Bacteria_Proteobacteria_Alphaproteobacteria_Rhodospirillales_Acetobacteraceae_Granulibacter_bethesdensis_CGDNIH1
PROA14	Bacteria_Proteobacteria_Alphaproteobacteria_Rhodospirillales_Acetobacteraceae_Saccharibacter_floricola_DSM_15669
PROA15	Bacteria_Proteobacteria_Alphaproteobacteria_Rhodospirillales_Rhodospirillaceae_Inquilinus_limosus_DSM_16000
PROA16	Bacteria_Proteobacteria_Alphaproteobacteria_Rhodospirillales_Rhodospirillaceae_Nisaea_denitrificans_DSM_18348
PROA17	Bacteria_Proteobacteria_Alphaproteobacteria_Rhodospirillales_Rhodospirillaceae_Rhodospirillum_rubrum_S1_ATCC_11170
PROA18	Bacteria_Proteobacteria_Alphaproteobacteria_Rhodospirillales_Rhodospirillaceae_Tistrella_mobilis_KA081020_065
PROA19	Bacteria_Proteobacteria_Alphaproteobacteria_Rickettsiales_Anaplasmataceae_Anaplasma_marginale_St._Maries
PROA2	Bacteria_Proteobacteria_Alphaproteobacteria_Kordiimonadales_Kordiimonas_gwangyangensis_DSM_19435
PROA20	Bacteria_Proteobacteria_Alphaproteobacteria_Rickettsiales_Rickettsiaceae_Rickettsieae_Rickettsia_conorii_Malish_7
PROA21	Bacteria_Proteobacteria_Alphaproteobacteria_Sphingomonadales_Sphingomonadaceae_Citromicrobium_sp._JLT1363
PROA3	Bacteria_Proteobacteria_Alphaproteobacteria_Rhizobiales_Bradyrhizobiaceae_Bosea_sp._UNC402CLCol
PROA4	Bacteria_Proteobacteria_Alphaproteobacteria_Rhizobiales_Hyphomicrobiaceae_Pelagibacterium_halotolerans_B2

PROA5	Bacteria_Proteobacteria_Alphaproteobacteria_Rhizobiales_Phyllobacteriaceae_Mesorhizobium_australicum_WSM2073
PROA6	Bacteria_Proteobacteria_Alphaproteobacteria_Rhizobiales_Rhizobiaceae_Kaistia_adipata_DSM_17808
PROA7	Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Hyphomonadaceae_Hellea_balneolensis_DSM_19091
PROA8	Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Hyphomonadaceae_Hirschia_baltica_ATCC_49814
PROA9	Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_CG2_30_SUB100_Rhodobacteraceae_10_405
PROB1	Bacteria_Proteobacteria_Betaproteobacteria_Burkholderiales_Alcaligenaceae_Bordetella_bronchiseptica_RB50
PROB10	Bacteria_Proteobacteria_Betaproteobacteria_Rhodocyclales_Rhodocyclaceae_Azovibrio_restrictus_DSM_23866
PROB2	Bacteria_Proteobacteria_Betaproteobacteria_Burkholderiales_Comamonadaceae_Comamonas_testosteroni_TK102
PROB3	Bacteria_Proteobacteria_Betaproteobacteria_Burkholderiales_Comamonadaceae_Hylemonella_gracilis_ATCC_19624
PROB4	Bacteria_Proteobacteria_Betaproteobacteria_Burkholderiales_Oxalobacteraceae_Massilia_alkalitolerans_DSM_17462
PROB5	Bacteria_Proteobacteria_Betaproteobacteria_Burkholderiales_unclassified_Burkholderiales_Burkholderiales_Genera_incertae_sedis_Ideonella_sp._B508_1
PROB6	Bacteria_Proteobacteria_Betaproteobacteria_Gallionellales_Gallionellaceae_Gallionella_capsiferriformans_ES_2
PROB7	Bacteria_Proteobacteria_Betaproteobacteria_Methylophilales_Methylophilaceae_Methylovorus_sp._SIP3_4
PROB8	Bacteria_Proteobacteria_Betaproteobacteria_Neisseriales_Neisseriaceae_Deefgea_rivuli_DSM_18356
PROB9	Bacteria_Proteobacteria_Betaproteobacteria_Neisseriales_Neisseriaceae_Stenoxybacter_acetivorans_DSM_19021
PROD1	Bacteria_Proteobacteria_deltaepsilon_subdivisions_Deltaproteobacteria_Bdellovibrionales_Bacteriovoracaceae_Bacteriovorax_marinus_SJ
PROD2	Bacteria_Proteobacteria_deltaepsilon_subdivisions_Deltaproteobacteria_Desulfobacterales_Desulfobacteraceae_Desulfotignum_phosphitoxidans_FiPS_3
PROD3	Bacteria_Proteobacteria_deltaepsilon_subdivisions_Deltaproteobacteria_Desulfovibrionales_Desulfohalobiaceae_Desulfonatronovibrio_hydrogenovorans_DSM_9292
PROD4	Bacteria_Proteobacteria_deltaepsilon_subdivisions_Deltaproteobacteria_Desulfuromonadales_Geobacteraceae_Geoalkalibacter_ferrihydriticus_DSM_17813
PROD5	Bacteria_Proteobacteria_deltaepsilon_subdivisions_Deltaproteobacteria_Myxococcales_Cystobacterineae_Myxococcaceae_Myxococcus_xanthus_DK_1622
PROD6	Bacteria_Proteobacteria_deltaepsilon_subdivisions_Deltaproteobacteria_Syntrophobacterales_Syntrophaceae_Syntrophus_aciditrophicus_SB
PROE1	Bacteria_Proteobacteria_deltaepsilon_subdivisions_Epsilonproteobacteria_Campylobacterales_Helicobacteraceae_Sulfuricurvum_sp._RIFRC_1
PROE2	Bacteria_Proteobacteria_deltaepsilon_subdivisions_Epsilonproteobacteria_Nautiliales_Nautiliaceae_Lebetimonas_sp._JS032
PROG1	Bacteria_Proteobacteria_Gammaproteobacteria_Aeromonadales_Succinivibrionaceae_Anaerobiospirillum_succiniciproducens_DSM_6400
PROG10	Bacteria_Proteobacteria_Gammaproteobacteria_Methylococcales_Methylococcaceae_Methyloglobulus_morosus_KoM1
PROG11	Bacteria_Proteobacteria_Gammaproteobacteria_Oceanospirillales_Oceanospirillaceae_Oceanospirillum_maris_DSM_6286
PROG12	Bacteria_Proteobacteria_Gammaproteobacteria_Oceanospirillales_Saccharospirillaceae_Saccharospirillum_impatiens_DSM_12546
PROG13	Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Moraxellaceae_Alkanindiges_illinoisensis_DSM_15370
PROG14	Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Moraxellaceae_Psychrobacter_cryohalolentis_K5
PROG15	Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Pseudomonadaceae_Pseudomonas_flectens_ATCC_12775
PROG16	Bacteria_Proteobacteria_Gammaproteobacteria_Vibrionales_Vibrionaceae_Photobacterium_profundum_SS9
PROG17	Bacteria_Proteobacteria_Gammaproteobacteria_Xanthomonadales_Xanthomonadaceae_Frateuria_terrea_CGMCC_1.7053
PROG18	Bacteria_Proteobacteria_Gammaproteobacteria_Xanthomonadales_Xanthomonadaceae_Pseudoxanthomonas_spadix_BD_a59
PROG2	Bacteria_Proteobacteria_Gammaproteobacteria_Alteromonadales_Alteromonadaceae_Haliea_rubra_CM41_15a_DSM_19751
PROG3	Bacteria_Proteobacteria_Gammaproteobacteria_Alteromonadales_Alteromonadales_genera_incertae_sedis_Teredinibacter_turnerae_T7902
PROG4	Bacteria_Proteobacteria_Gammaproteobacteria_Alteromonadales_Moritellaceae_Moritella_marina_ATCC_15381
PROG5	Bacteria_Proteobacteria_Gammaproteobacteria_Chromatiales_Chromatiaceae_Thiocapsa_marina_5811_DSM_5653

PROG6	Bacteria_Proteobacteria_Gammaproteobacteria_Chromatiales_Ectothiorhodospiraceae_Arhodomonas_aquaeolei_DSM_8974
PROG7	Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacteriales_Enterobacteriaceae_Arsenophonus_nasoniae_DSM_15247
PROG8	Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacteriales_Enterobacteriaceae_Buchnera_Buchnera_aphidicola_str._APS_Acyrthosiphon_pisum
PROG9	Bacteria_Proteobacteria_Gammaproteobacteria_Legionellales_Legionellaceae_Legionella_micdadei_ATCC_33218
PROZ1	Bacteria_Proteobacteria_Zetaproteobacteria_Mariprofundales_Mariprofundaceae_Mariprofundus_ferrooxydans_M34
PROZ2	Bacteria_Proteobacteria_Zetaproteobacteria_Zetaproteobacteria_bacterium_TAG_1
SM2F11	Bacteria_SM2F11_RIFCSPHIGHO2_01_FULL_SM2F11_50_11
SPIR1	Bacteria_Spirochaetes_Spirochaetia_Spirochaetales_Brachyspiraceae_Brachyspira_murdochii_DSM_12563
SPIR2	Bacteria_Spirochaetes_Spirochaetia_Spirochaetales_Leptospiraceae_Leptospira_biflexa_serovar_Patoc_strain_Patoc_1_Paris_I
SPIR3	Bacteria_Spirochaetes_Spirochaetia_Spirochaetales_Spirochaetaceae_Spirochaeta_sp._L21_RPul_D2
SPIR4	Bacteria_Spirochatetes_Spirochaetia_Spirochaetales_RIFOXYC1_FULL_Spirochaetes_54_7
SYNE1	Bacteria_Synergistetes_Synergistia_Synergistales_Synergistaceae_Aminobacterium_colombiense_DSM_12261
SYNE2	Bacteria_Synergistetes_Synergistia_Synergistales_Synergistaceae_Thermanaerovibrio_acidaminovorans_DSM_6589
TENE1	Bacteria_Tenericutes_Mollicutes_Entomoplasmatales_Entomoplasmataceae_Mesoplasma_florum_W37
TENE2	Bacteria_Tenericutes_Mollicutes_Mycoplasmatales_Mycoplasmataceae_Candidatus_Hepatoplasma_crinochetorum_Av
THER1	Bacteria_Thermotogae_Thermotogae_Thermotogales_Thermotogaceae_Kosmotoga_olearia_TBF_19.5.1
THER2	Bacteria_Thermotogae_Thermotogae_Thermotogales_Thermotogaceae_Thermotoga_maritima_MSB8
THERSUL1	Bacteria_Thermodesulfobacteria_Thermodesulfobacteria_Thermodesulfobacteriales_Thermodesulfobacteriaceae_Thermodesulfatator_indicus_CIR29812_DSM_15286
UNC1	Bacteria_unclassified_Bacteria_Poribacteria_Candidatus_Poribacteria_WGA_3G_final_clean_version
WIRT	Bacteria_Wirthbacteria_CG2_30_FULL_Wirthbacteria_54_11
WOR1	Bacteria_WOR_1_uncultured_DG_54_3
WOR2	Bacteria_WOR_2_uncultured_SMTZ_29
WOR3	Bacteria_WOR_3_uncultured_SMTZ_42
ZIXI1	Bacteria_Zixibacteria_uncultured_SMTZ_73_2

Appendix B

Full heatmap for all COGs used in Chapter 3



Appendix C

Supplementary figures for Chapter 5.

Supplementary Figure 1. G1PDH full tree, 111 sequences, 190 positions, inferred under LG+C60 model. Root position inferred using relaxed uncorrelated lognormal clock model with a Yule prior and LG substitution model. MAD root indicated with asterisk.

Supplementary Figure 2. GGGPS full tree, 133 sequences, 129 positions, inferred under LG+C40 model. Root position inferred using relaxed uncorrelated lognormal clock model with a Yule prior and LG substitution model. MAD root indicated with asterisk.

Supplementary Figure 3. GGGPS large subclade, 98 sequences, 166 positions, inferred under LG+C40 model. Root position inferred using relaxed uncorrelated lognormal clock model with a Yule prior and LG substitution model. MAD root indicated with asterisk.

Supplementary Figure 4. DGGGPS full tree, 97 sequences, 119 positions, inferred under LG+C60 model. Root position inferred using relaxed uncorrelated lognormal clock model with a Yule prior and LG substitution model. MAD root indicated with asterisk.

Supplementary Figure 5. GpsA full tree, 84 sequences, 169 positions, inferred under LG+C60 model. Root position inferred using relaxed uncorrelated lognormal clock model with a Yule prior and LG substitution model. MAD root indicated with asterisk.

Supplementary Figure 6. Glp full tree, 51 sequences, 199 positions, inferred under LG+C40 model. Root position inferred using relaxed uncorrelated lognormal clock model with a Yule prior and LG substitution model. MAD root indicated with asterisk.

Supplementary Figure 7. GlpK full tree, 77 sequences, 363 positions, inferred under LG+C60 model. Root position inferred using relaxed uncorrelated lognormal clock model with a Yule prior and LG substitution model. MAD root indicated with asterisk.

Supplementary Figure 8. PlsC full tree, 74 sequences, 57 positions, inferred under LG+C60 model. Root position inferred using relaxed uncorrelated lognormal clock model with a Yule prior and LG substitution model. MAD root between clade comprising of sequences in red and other sequences.

Supplementary Figure 9. PlsY full tree, 60 sequences, 104 positions, inferred under LG+C50 model. Root position inferred using relaxed uncorrelated lognormal clock model with a Yule prior and LG substitution model. MAD root indicated with asterisk.

Supplementary Figure 10. G1PDH full tree, 123 sequences, 173 positions, inferred under LG+C60 model. Root position inferred using 3-dehydroquinate synthase (DHQS), five glucerol dehydrogenase (GDH) and five alcohol dehydrogenase (ALDH) sequences as an outgroup

Supplementary Figure 11. Glp full tree, 63 sequences, 183 positions, inferred under LG+C60 model. Root position inferred using 12 FAD-dependent oxidoreductase sequences as an outgroup

Supplementary Figure 12. Gpsa full tree, 96 sequences, 148 positions, inferred under LG+C60 model. Root position inferred using six hydroxyacyl-CoA dehydrogenase (HACDH) and 6 UDP-glucose 6-dehydrogenase (UDPGDH) sequences as an outgroup

Supplementary Figure 13. Gpsa full tree with eukaryotic sequences, 113 sequences, 159 positions, inferred under LG+C50 model.

Supplementary Figure 14. Glp full tree with eukaryotic sequences, 80 sequences, 190 positions, inferred under LG+C50 model.

Supplementary Figure 15. PlsC full tree with eukaryotic sequences, 96 sequences, 54 positions, inferred under LG+C60 model.

Supplementary Figure 16. Unrooted G1PDH full tree, 111 sequences, 190 positions, inferred under LG+C60 model.

Supplementary Figure 17. Unrooted GGGPS full tree, 133 sequences, 129 positions, inferred under LG+C40 model.

Supplementary Figure 18. Unrooted DGGGPS full tree, 97 sequences, 119 positions, inferred under LG+C60 model.

Supplementary Figure 19. Unrooted GpsA full tree, 84 sequences, 169 positions, inferred under LG+C60 model.

Supplementary Figure 20. Unrooted Glp full tree, 51 sequences, 199 positions, inferred under LG+C40 model.

Supplementary Figure 21. Unrooted GlpK full tree, 77 sequences, 363 positions, inferred under LG+C60 model.

Supplementary Figure 22. Unrooted PlsC full tree, 74 sequences, 57 positions, inferred under LG+C60 model.

Supplementary Figure 23. Unrooted PlsY full tree, 60 sequences, 104 positions, inferred under LG+C50 model

Supplementary Figure 24. Gpsa full tree, 96 sequences, 148 positions, inferred under LG+C50 model. Root position inferred using six hydroxyacyl-CoA dehydrogenase (HACDH) and 6 UDP-glucose 6-dehydrogenase (UDPGDH) sequences as an outgroup

Supplementary Figure 25. UbiA full tree, 227 sequences, 69 positions, inferred under LG+C60 model. DGGGP sequences in blue, chlorophyll a synthase in green, protoheme IX farnesyltransferase in red, and 4-hydroxybenzoate octaprenyltransferase in black.

Supplementary Figure 26. Unrooted GGGPS full tree, 133 sequences, 129 positions, inferred under LG+C60 model.

Supplementary Figure 27. Unrooted PlsY full tree, 60 sequences, 104 positions, inferred under LG+C60 model

Supplementary Figure 28. GlpA/GlpD full tree with eukaryotic sequences, 80 sequences, 190 positions, inferred under LG+C60 model

Supplementary Figure 29. Gpsa full tree with eukaryotic sequences, 113 sequences, 159 positions, inferred under LG+C60 model

Supplementary Figure 30. Maximum likelihood G1PDH tree inferred in IQ-Tree under the LG+C60 model. Rooted using MAD

Supplementary Figure 31. Maximum likelihood GGGPS tree inferred in IQ-Tree under the LG+C60 model. Rooted using MAD

Supplementary Figure 32. Maximum likelihood DGGGPS tree inferred in IQ-Tree under the LG+C60 model. Rooted using MAD

Supplementary Figure 33. Maximum likelihood GpsA tree inferred in IQ-Tree under the LG+C60 model. Rooted using MAD

Supplementary Figure 34. Maximum likelihood GlpA/GlpD tree inferred in IQ-Tree under the LG+C60 model. Rooted using MAD

Supplementary Figure 35. Maximum likelihood GlpK tree inferred in IQ-Tree under the LG+C60 model. Rooted using MAD

Supplementary Figure 36. Maximum likelihood PlsC tree inferred in IQ-Tree under the LG+C60 model. Rooted using MAD

Supplementary Figure 37. Maximum likelihood PlsY tree inferred in IQ-Tree under the LG+C60 model. Rooted using MAD

Supplementary Figure 38. Maximum likelihood G1PDH tree inferred in IQ-Tree under the LG+C60 model from HoT alignments. Rooted using MAD

Supplementary Figure 39. Maximum likelihood GGGPS tree inferred in IQ-Tree under the LG+C60 model from HoT alignments. Rooted using MAD

Supplementary Figure 40. Maximum likelihood DGGGPS tree inferred in IQ-Tree under the LG+C60 model from HoT alignments. Rooted using MAD

Supplementary Figure 41. Maximum likelihood GpsA tree inferred in IQ-Tree under the LG+C60 model from HoT alignments. Rooted using MAD

Supplementary Figure 42. Maximum likelihood GlpA/GlpD tree inferred in IQ-Tree under the LG+C60 model from HoT alignments. Rooted using MAD

Supplementary Figure 43. Maximum likelihood GlpK tree inferred in IQ-Tree under the LG+C60 model from HoT alignments. Rooted using MAD

Supplementary Figure 44. Maximum likelihood PlsC tree inferred in IQ-Tree under the LG+C60 model from HoT alignments. Rooted using MAD

Supplementary Figure 45. Maximum likelihood PlsY tree inferred in IQ-Tree under the LG+C60 model from HoT alignments. Rooted using MAD

Supplementary Figure 46. Maximum likelihood G1PDH tree inferred in IQ-Tree under the LG+C60 model from alignments with metagenomic data removed. Rooted using MAD, with lognormal relaxed molecular clock show with an asterisk

Supplementary Figure 47. Maximum likelihood GGGPS tree inferred in IQ-Tree under the LG+C60 model from alignments with metagenomic data removed. Rooted using MAD, with lognormal relaxed molecular clock show with an asterisk

Supplementary Figure 48. Maximum likelihood reduce GGGPS tree inferred in IQ-Tree under the LG+C60 model from alignments with metagenomic data removed. Rooted using MAD, with lognormal relaxed molecular clock show with an asterisk

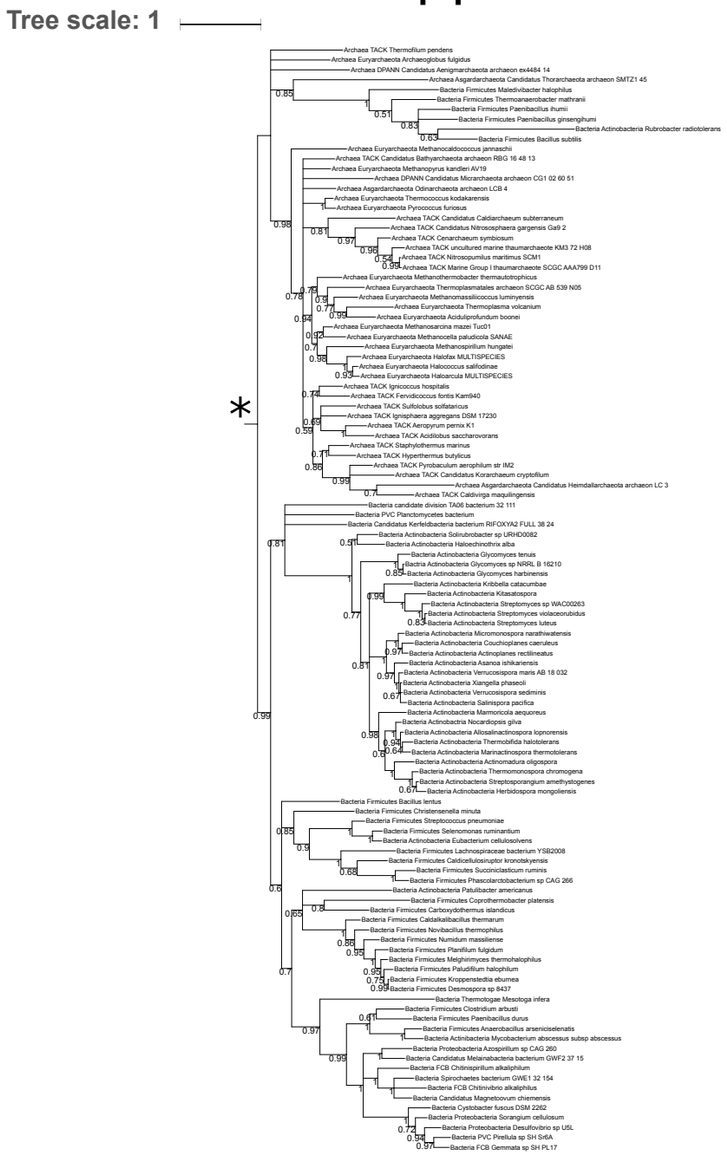
Supplementary Figure 49. Maximum likelihood DGGGPS tree inferred in IQ-Tree under the LG+C60 model from alignments with metagenomic data removed. Rooted using MAD, with lognormal relaxed molecular clock show with an asterisk

Supplementary Figure 50. Maximum likelihood GpsA tree inferred in IQ-Tree under the LG+C60 model from alignments with metagenomic data removed. Rooted using MAD, with lognormal relaxed molecular clock show with an asterisk

Supplementary Figure 51. Maximum likelihood GlpA/GlpD tree inferred in IQ-Tree under the LG+C60 model from alignments with metagenomic data removed. Rooted using MAD, with lognormal relaxed molecular clock show with an asterisk

Supplementary Figure 52. Maximum likelihood GlpK tree inferred in IQ-Tree under the LG+C60 model from alignments with metagenomic data removed. Rooted using MAD, with lognormal relaxed molecular clock show with an asterisk

Supplementary Figure 1

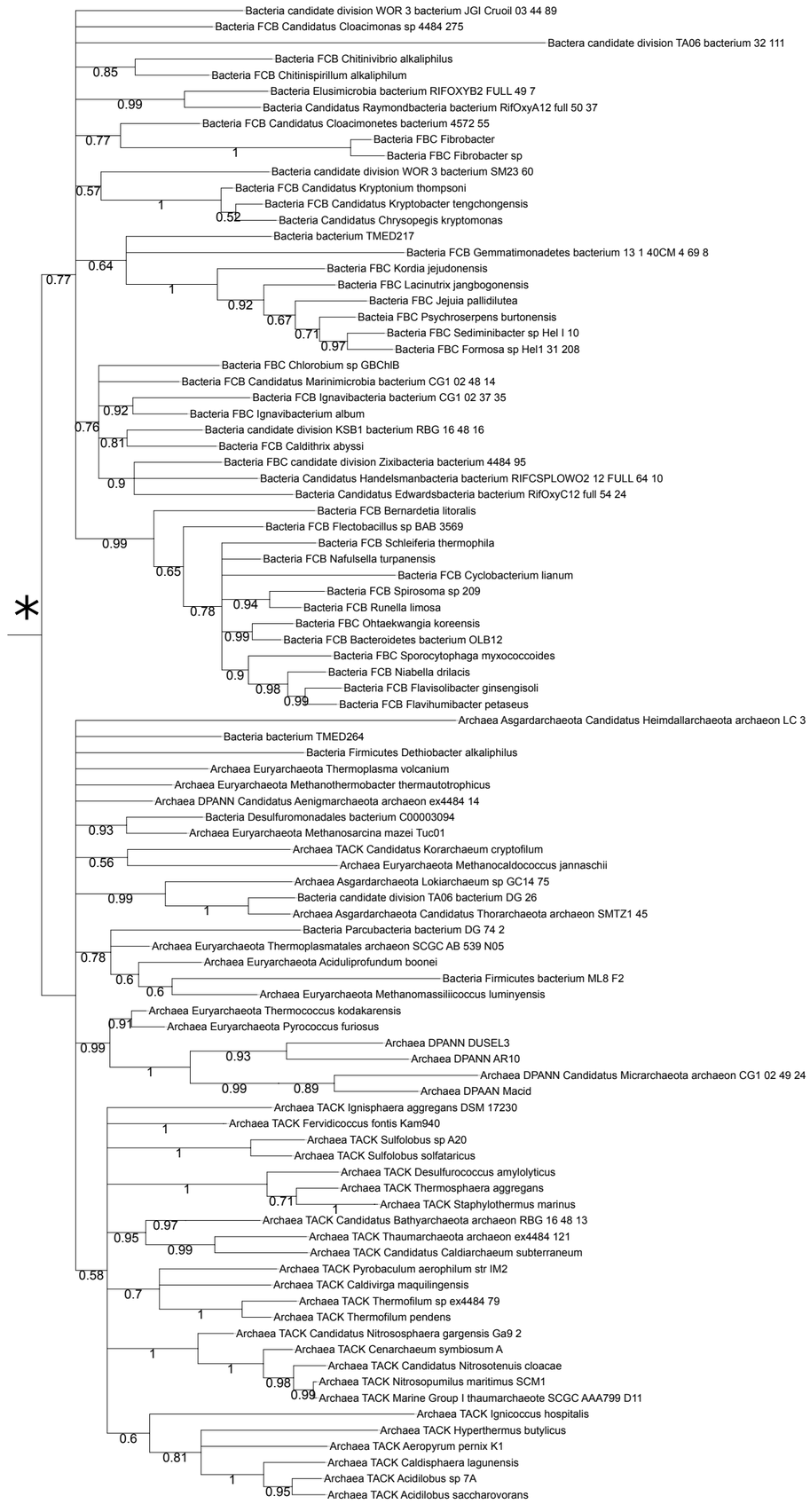


Supplementary Figure 2

Tree scale: 1
Archaea Euryarchaeota Haloferax volcanii
Archaea Euryarchaeota Halococcus saltifodinae
Archaea Euryarchaeota Haloarcula marismortui
Archaea Euryarchaeota Archaeoglobus fulgidus
Archaea Euryarchaeota Methanospirillum hungatei
Archaea Euryarchaeota Methanocella paludicola SANAE
Bacteria Firmicutes Peptococcaceae acidaminBRH c8a
Bacteria Firmicutes Tepidanaerobacter acetatoxydans
Bacteria Firmicutes Caldanaerobacter subterraneus
Bacteria Firmicutes Caldalkalibacillus thermarum
Bacteria Firmicutes Paenibacillus ginsenghumi
Bacteria Firmicutes Paenibacillus humi
Bacteria Firmicutes Paenibacillus durus
Bacteria Firmicutes Numidum massiliense
Bacteria Firmicutes Planifilum fulgidum
Bacteria Firmicutes Paludifilum halophilum
Bacteria Firmicutes Marinisama mesophilum
Bacteria Firmicutes Melghirimyces thermohalophilus
Bacteria Firmicutes Kroppenstedtia eburnea
Bacteria Firmicutes Desmospora sp 8437
Bacteria Actinobacteria bacterium RBG 13 63 9
Bacteria Firmicutes Tuberibacillus calidus
Bacteria Firmicutes Streptococcus pneumoniae
Bacteria Firmicutes Geobacillus galactosidasius
Bacteria Firmicutes Bacillus subtilis
Bacteria Firmicutes Fictibacillus sp GDSW R2A3
Bacteria Firmicutes Anaerobacillus arsenicoselenatis
Bacteria Firmicutes Anaerobacillus alkalidiazotrophicus
Bacteria Firmicutes Planococcus maritimus
Bacteria Firmicutes Bhargavaea cecembensis
Bacteria Actinobacteria Streptomyces sp WAC00283
Bacteria Firmicutes Bacillus lentus
Bacteria Firmicutes Bacillus aquimaris
Bacteria Actinobacteria Mycobacterium abscessus subsp abscessus
Archaea Asgardarchaeota Candidatus Heimdallarchaeota archaeon LC 3
Archaea DPANN Candidatus Aenigmarchaeota archaeon ex4484 14
Bacteria Desulfuromonadales bacterium C00003094
Archaea Euryarchaeota Methanosarcina mazei Tuc01
Archaea Euryarchaeota Thermococcus kodakarensis
Archaea Euryarchaeota Pyrococcus furiosus
Archaea DPANN DUSEL3
Archaea DPANN AR10
Bacteria Parcubacteria bacterium DG 74 3
Archaea DPANN Candidatus Micrarchaeota archaeon CG1 02 49 24
Archaea DPANN Mack
Archaea Euryarchaeota Methanothermobacter thermautotrophicus
Archaea TACK Candidatus Korarchaeum cryptofilum
Archaea Euryarchaeota Methanocaldococcus jannaschii
Archaea Asgardarchaeota Odinarchaeota archaeon LCB 4
Archaea Asgardarchaeota Lokiarchaeum sp GC14 75
Bacteria candidate division TA06 bacterium DG 26
Archaea Asgardarchaeota Candidatus Thorarchaeota archaeon SMTZ1 45
Archaea TACK Thaumarchaeota archaeon ex4484 121
Archaea TACK Candidatus Caldiarchaeum subterraneum
Bacteria Latescibacteria bacterium DG 63
Archaea TACK Candidatus Bathyarchaeota archaeon RBG 16 48 13
Archaea TACK Thermofilum sp ex4484 79
Archaea TACK Thermofilum pendens
Archaea TACK Pyrobaculum aerophilum str IM2
Archaea TACK Caldivirga maquilingensis
Archaea TACK Ignisphaera aggregans DSM 17230
Archaea TACK Ignicoccus hospitalis
Archaea Euryarchaeota Methanopyrus kandleri AV19
Archaea TACK Candidatus Nitrosocaldus yellowstonii HL72
Archaea TACK Candidatus Nitrosocaldus cavascurensis SCGC AAA279 C10
Archaea TACK Candidatus Nitrosocaldus islandicus 3F
Archaea TACK Candidatus Nitrosocaldus sp A74
Archaea TACK Candidatus Nitrosocaldus sp A75
Archaea TACK Candidatus Nitrosocaldus sp A89
Archaea TACK Candidatus Nitrosocaldus sp A90
Archaea TACK Candidatus Nitrosocaldus sp A92
Archaea TACK Candidatus Nitrosocaldus sp A96
Archaea TACK Candidatus Nitrosocaldus sp A97
Archaea TACK Candidatus Nitrosocaldus sp A98
Archaea TACK Candidatus Nitrosocaldus sp A99
Archaea TACK Candidatus Nitrosocaldus sp A100
Archaea TACK Candidatus Nitrosocaldus sp A101
Archaea TACK Candidatus Nitrosocaldus sp

Supplementary Figure 3

Tree scale: 0.1

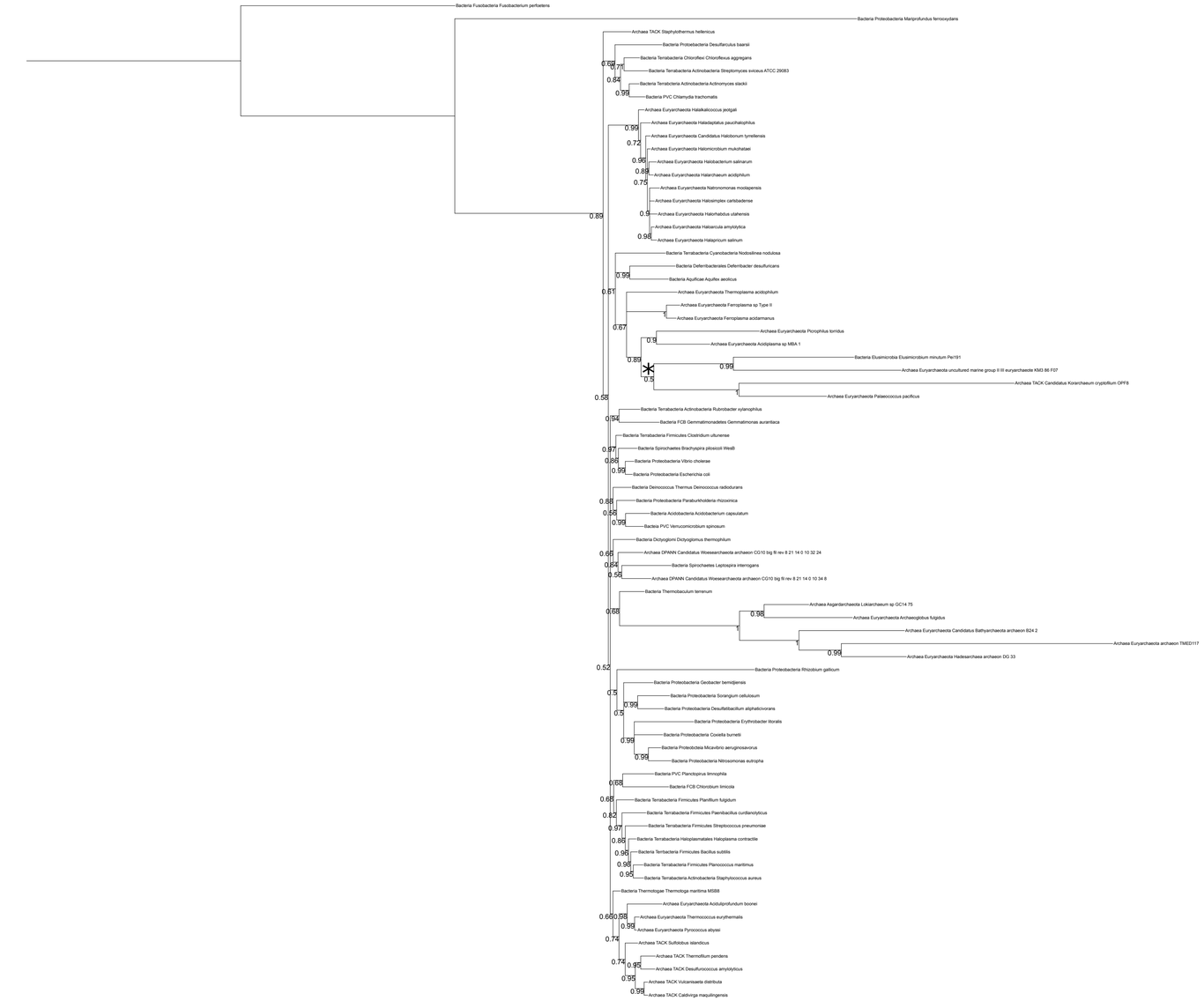


Supplementary Figure 4

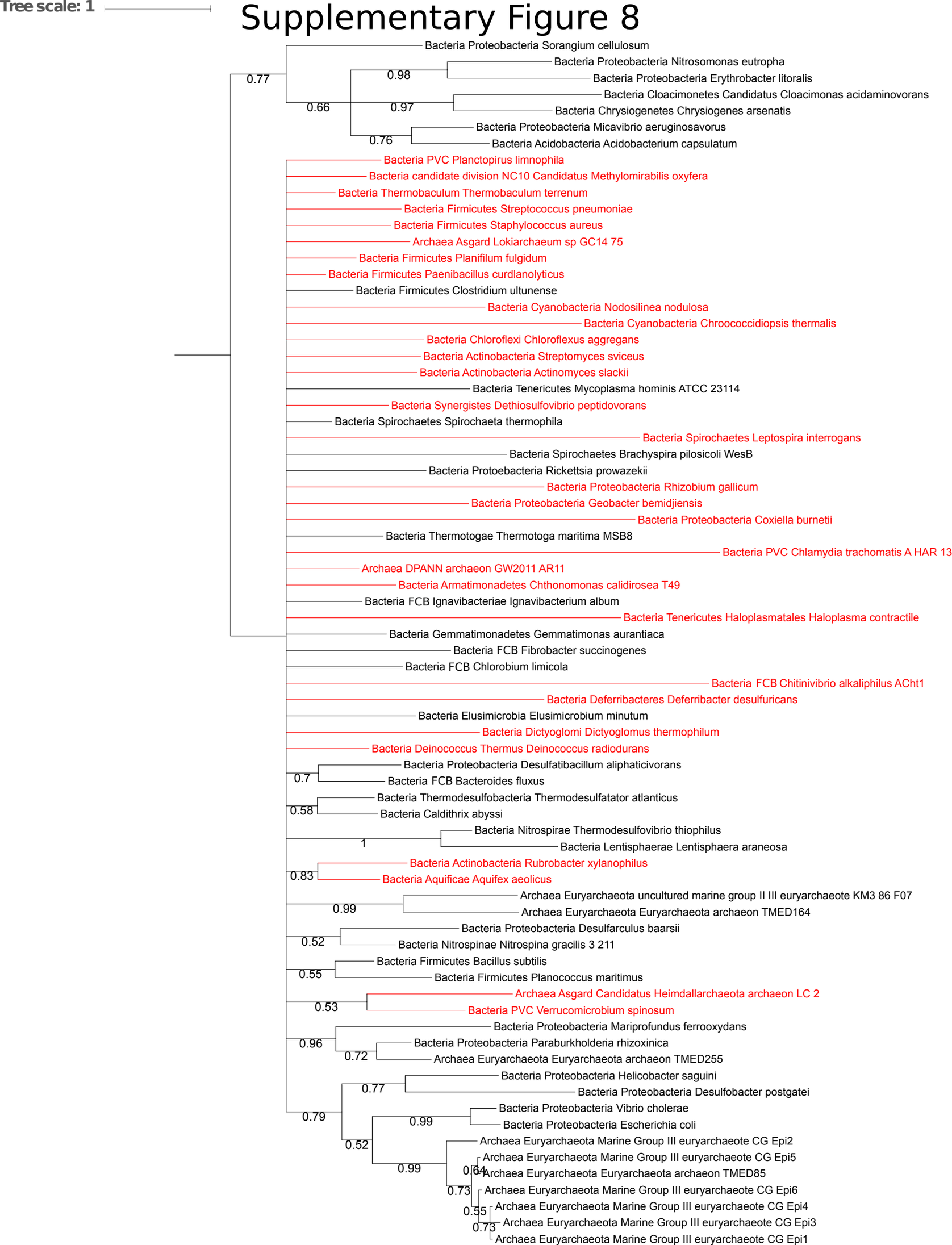
Supplementary Figure 5

Supplementary Figure 6

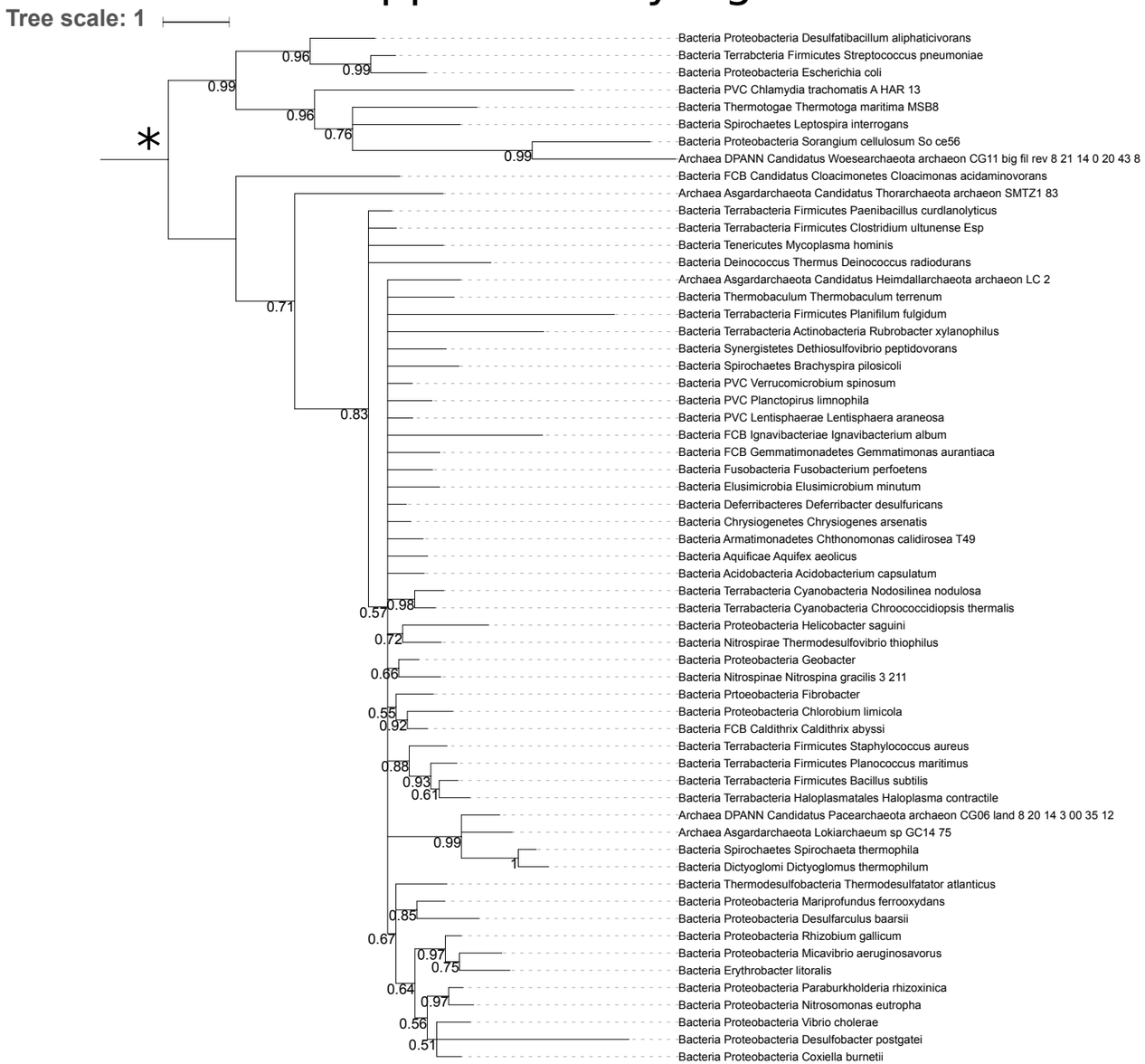
Supplementary Figure 7



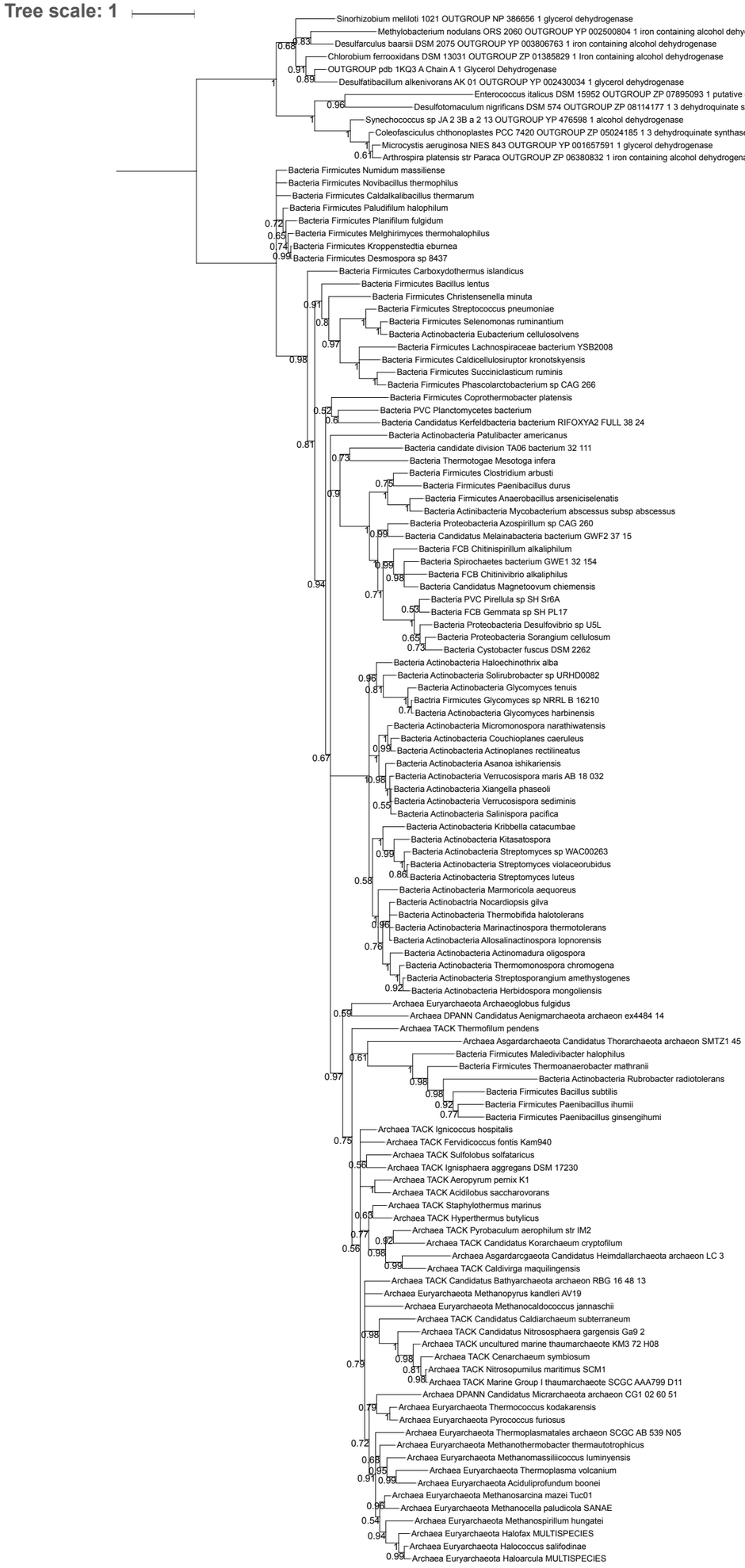
Supplementary Figure 8



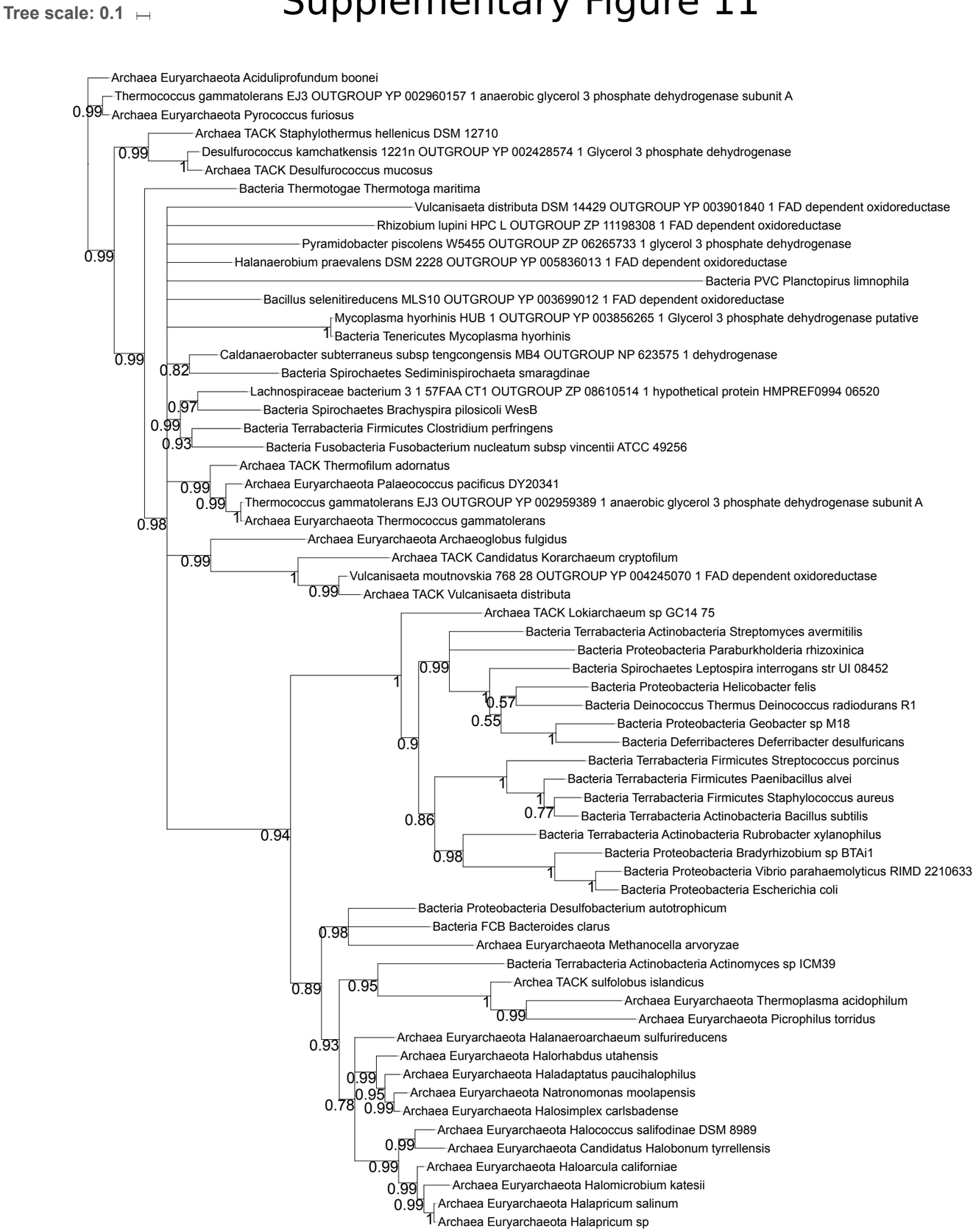
Supplementary Figure 9



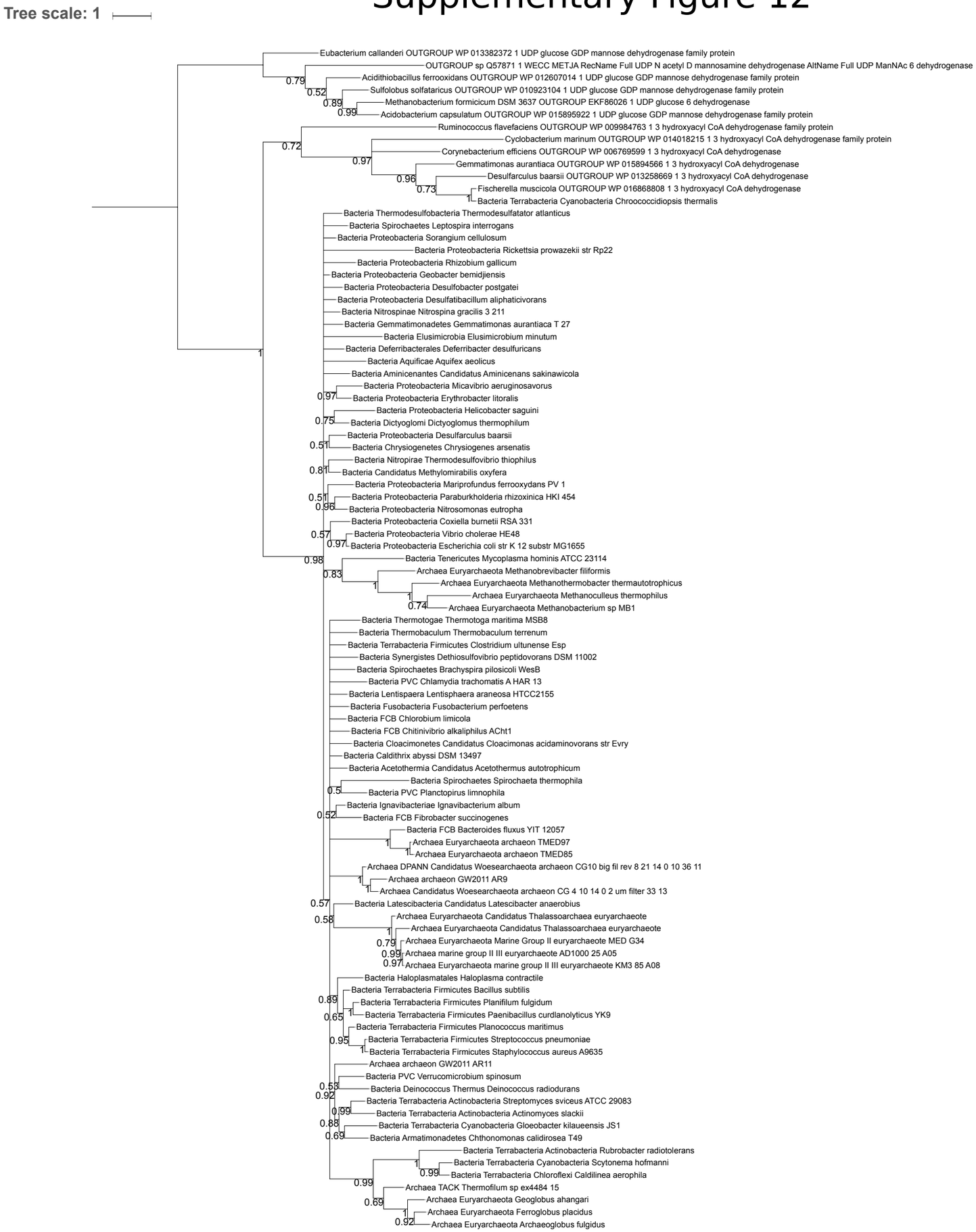
Supplementary Figure 10



Supplementary Figure 11

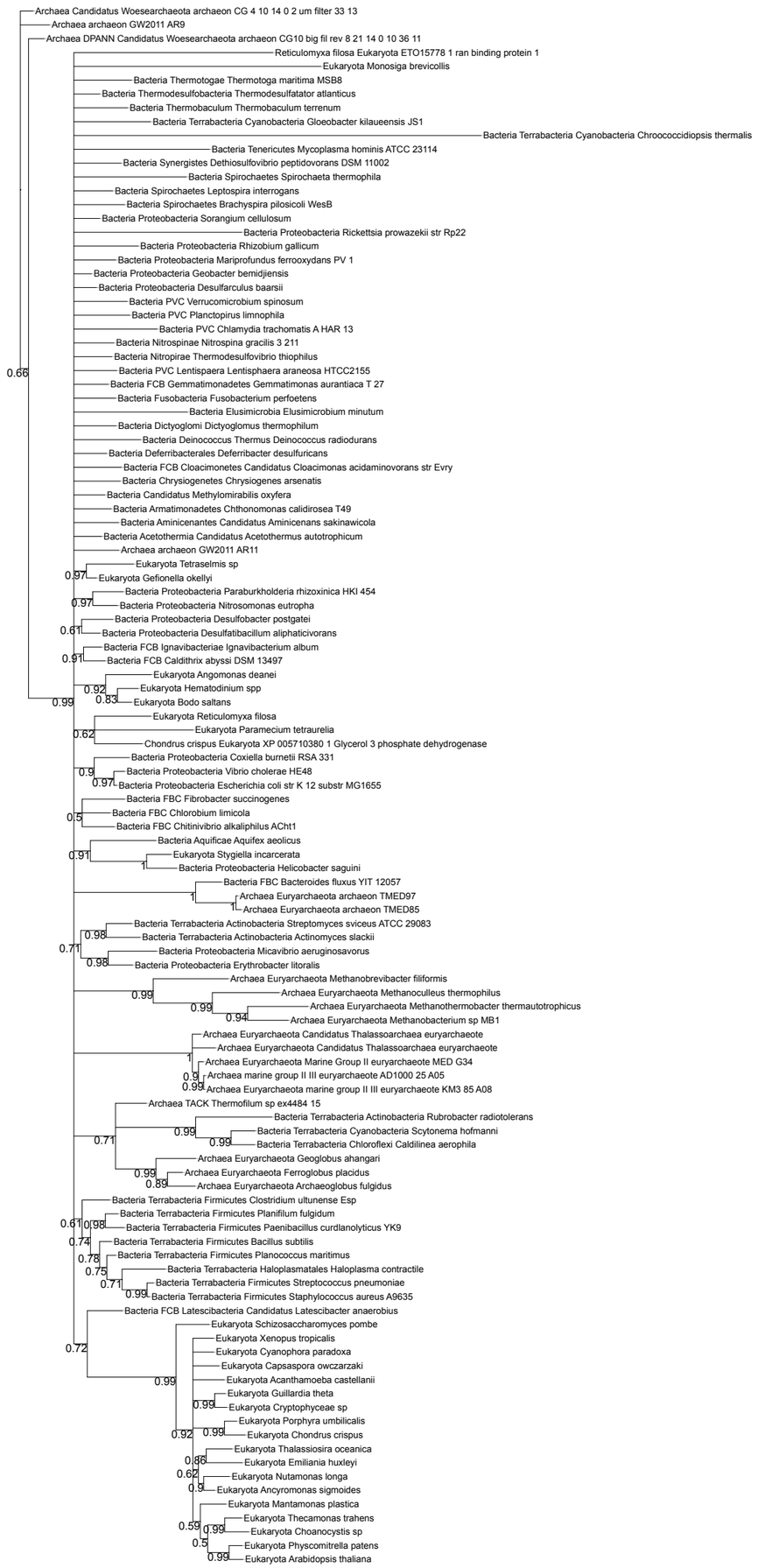


Supplementary Figure 12



Supplementary Figure 13

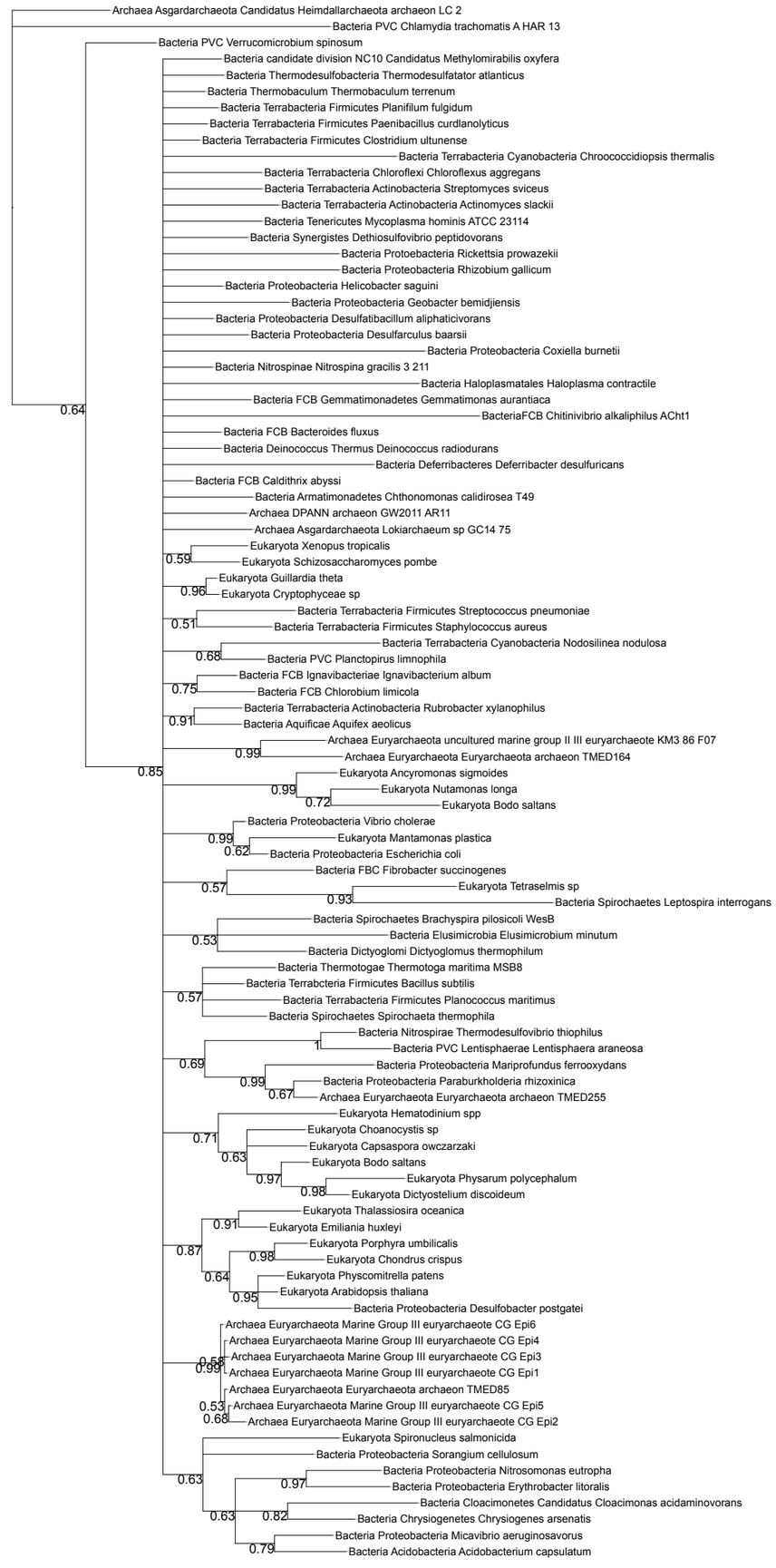
Tree scale: 1



Supplementary Figure 14

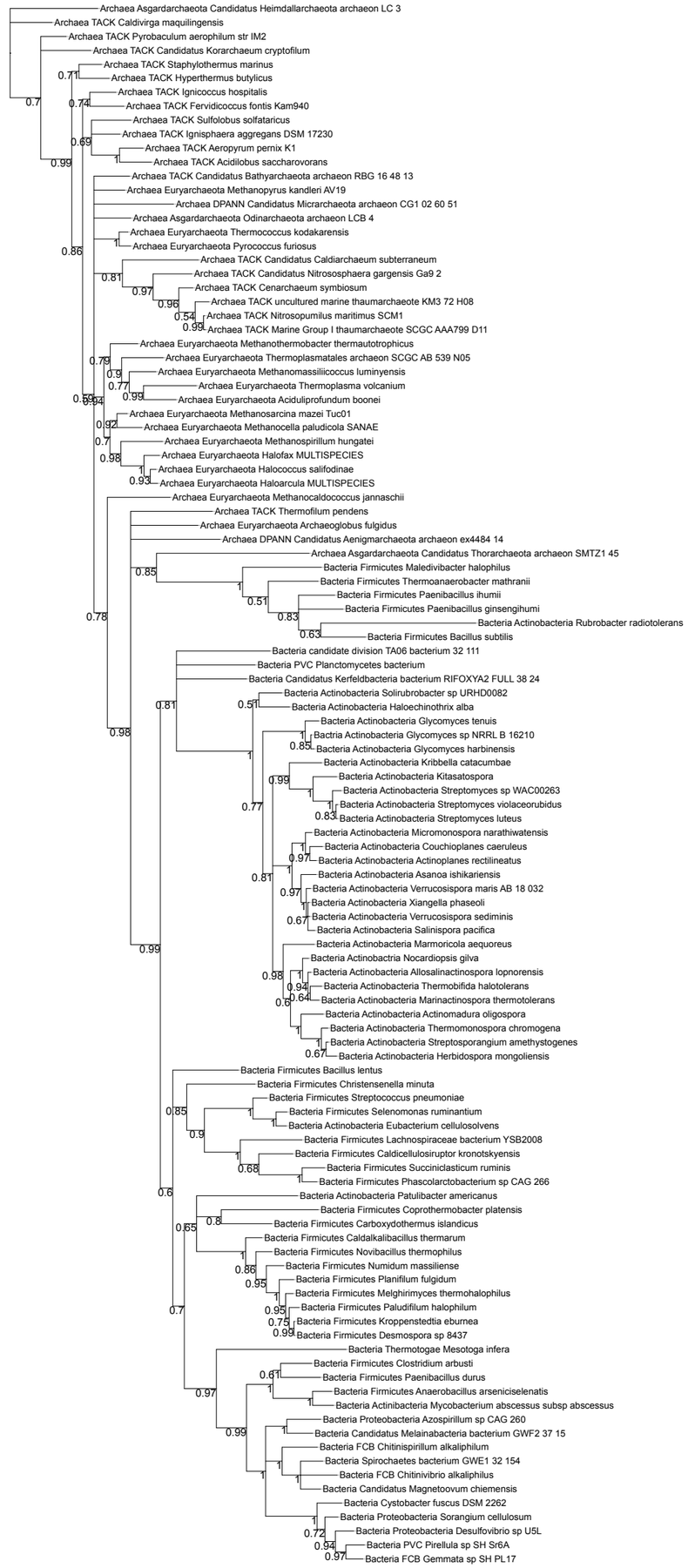
Supplementary Figure 15

Tree scale: 1



Supplementary Figure 16

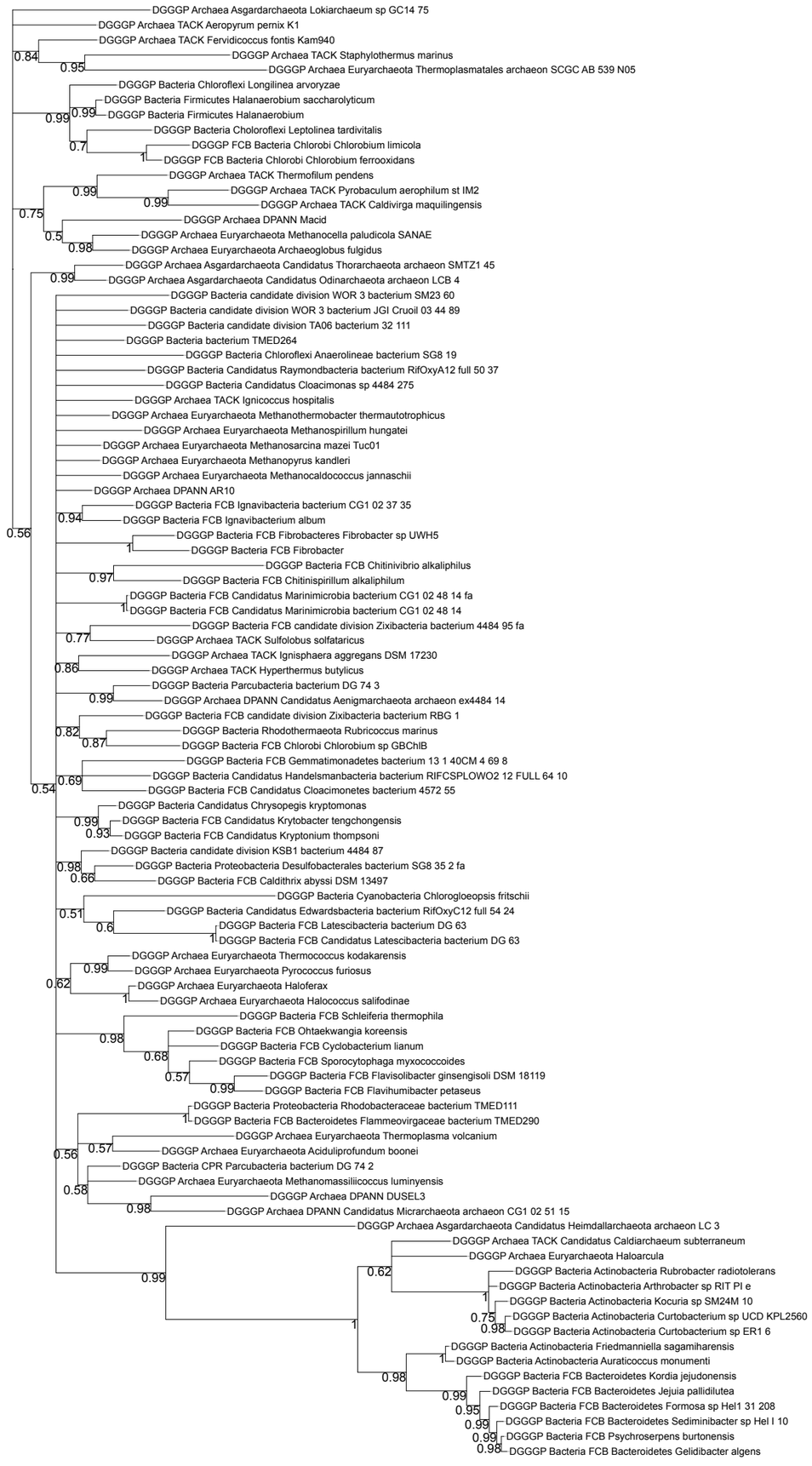
Tree scale: 1



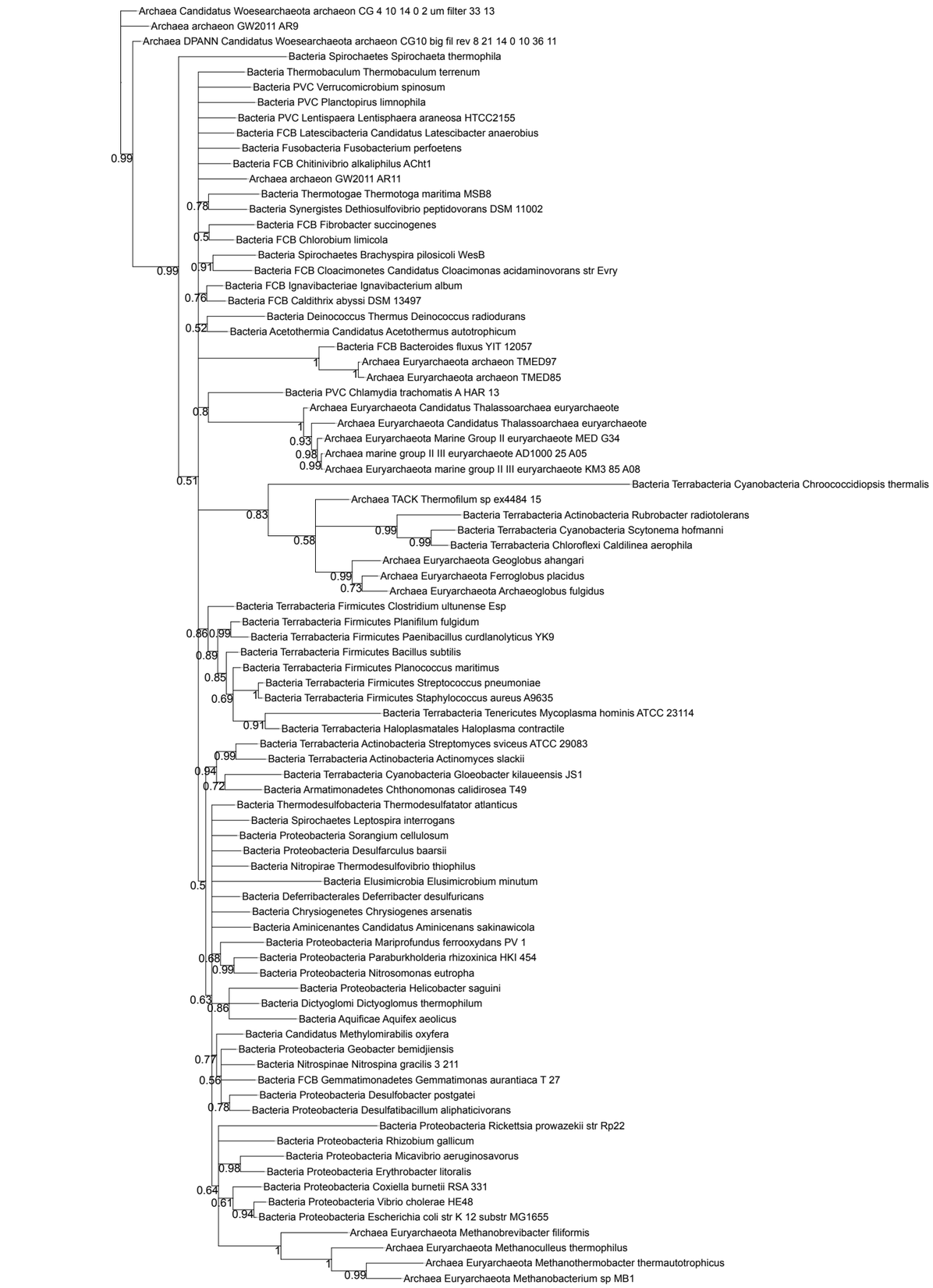
Supplementary Figure 17

Supplementary Figure 18

Tree scale: 1



Supplementary Figure 19

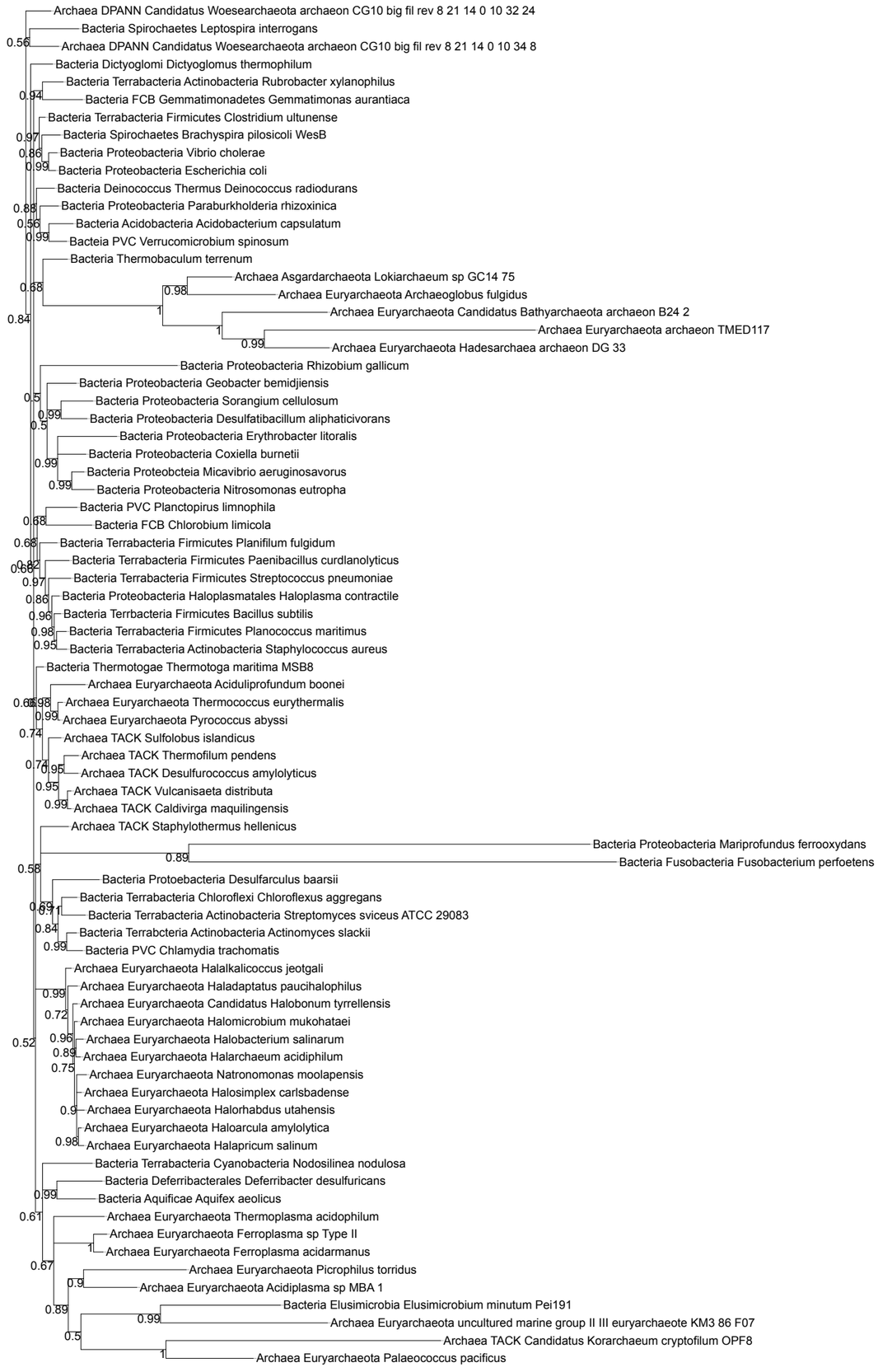


Tree scale: 1

Supplementary Figure 20

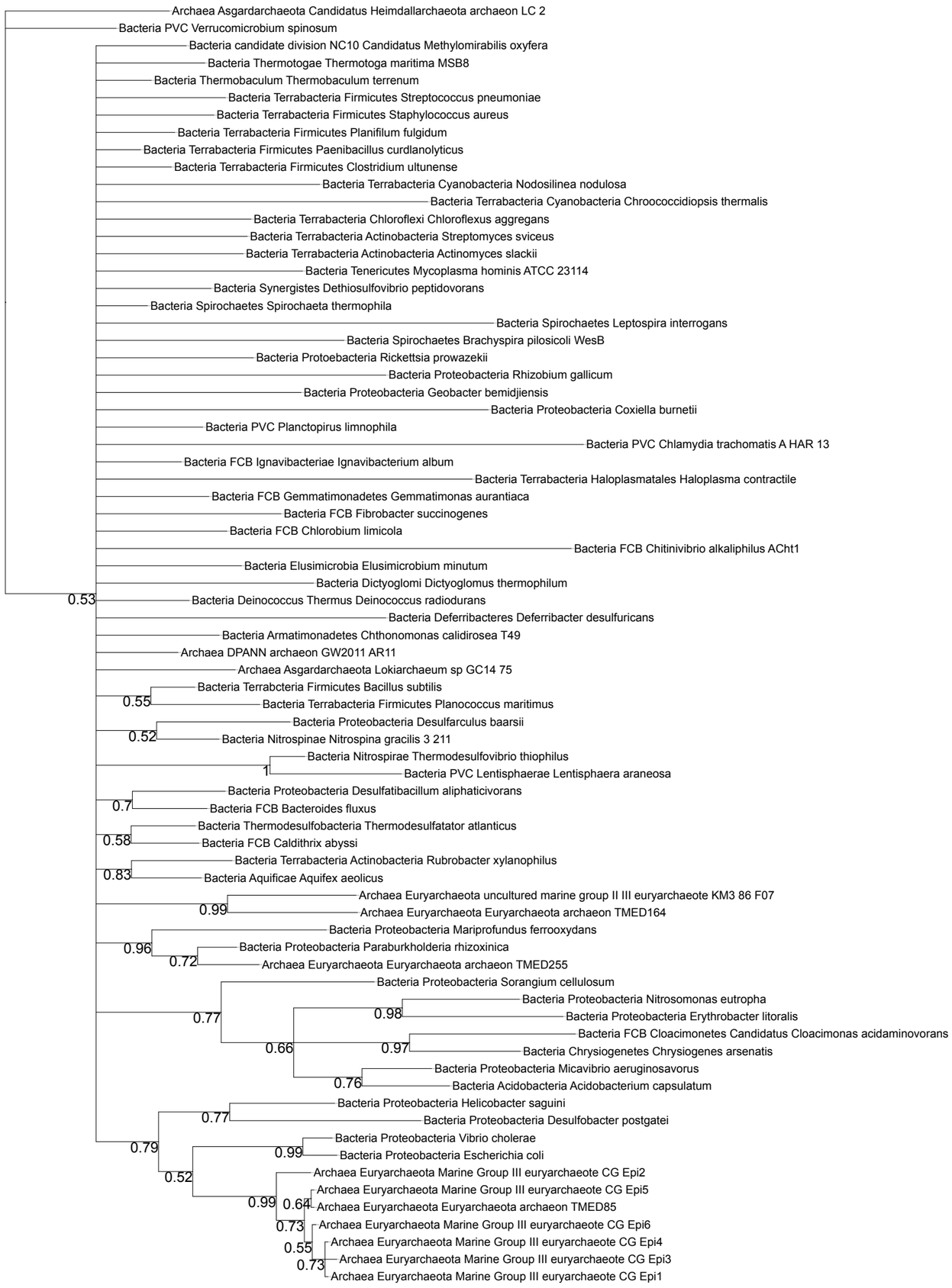
Supplementary Figure 21

Tree scale: 1

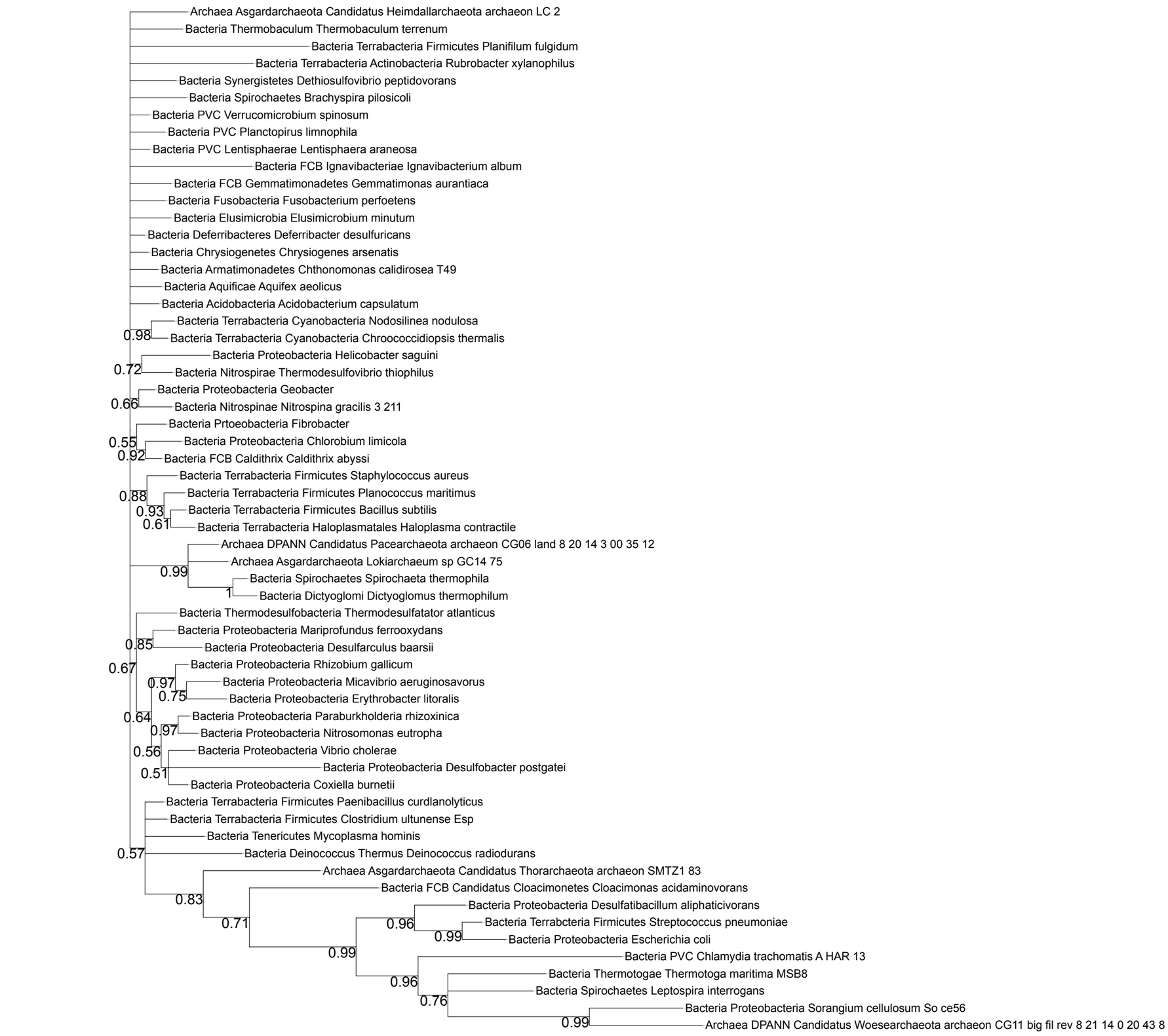


Tree scale: 1

Supplementary Figure 22



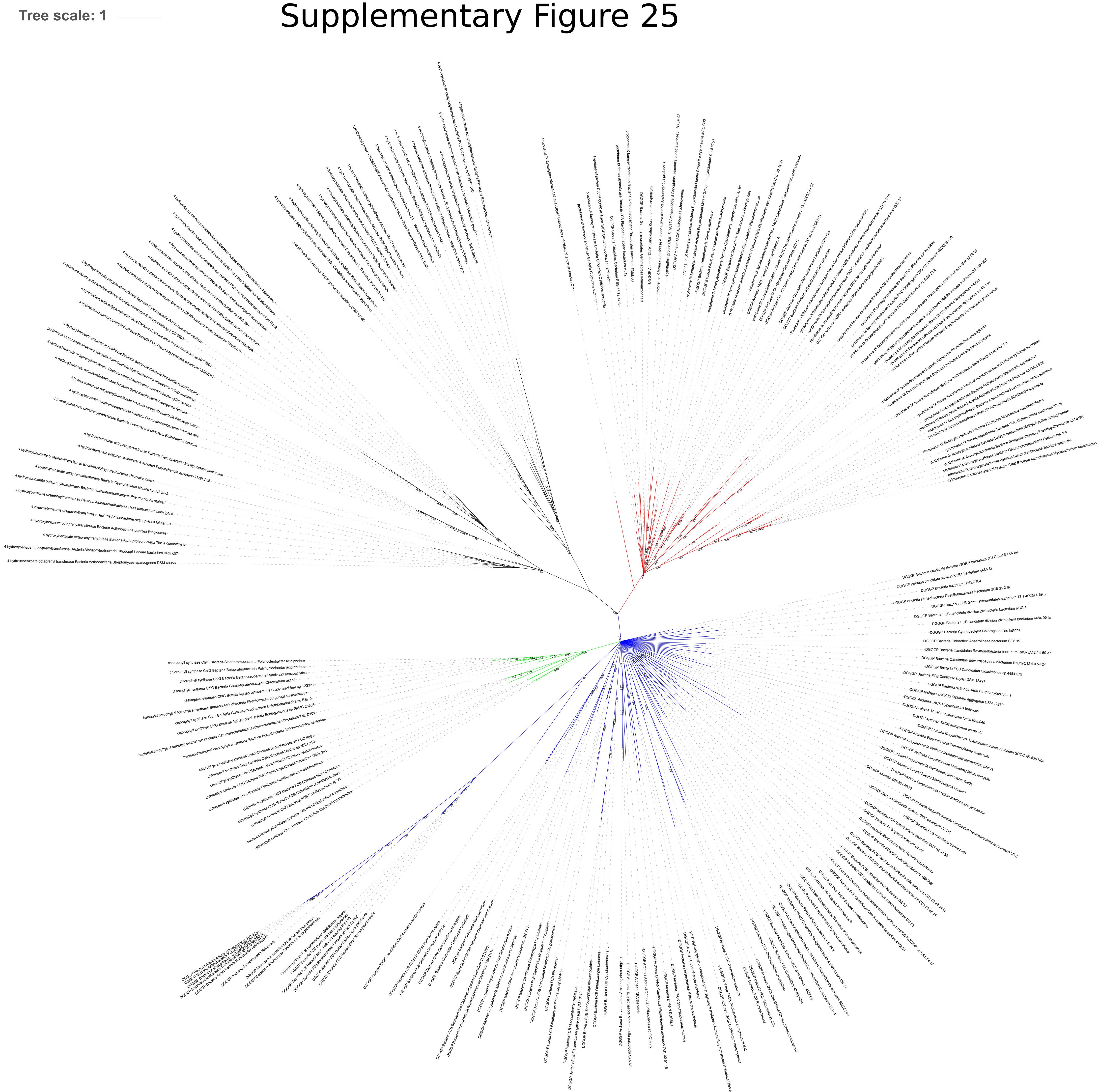
Supplementary Figure 23



Supplementary Figure 24

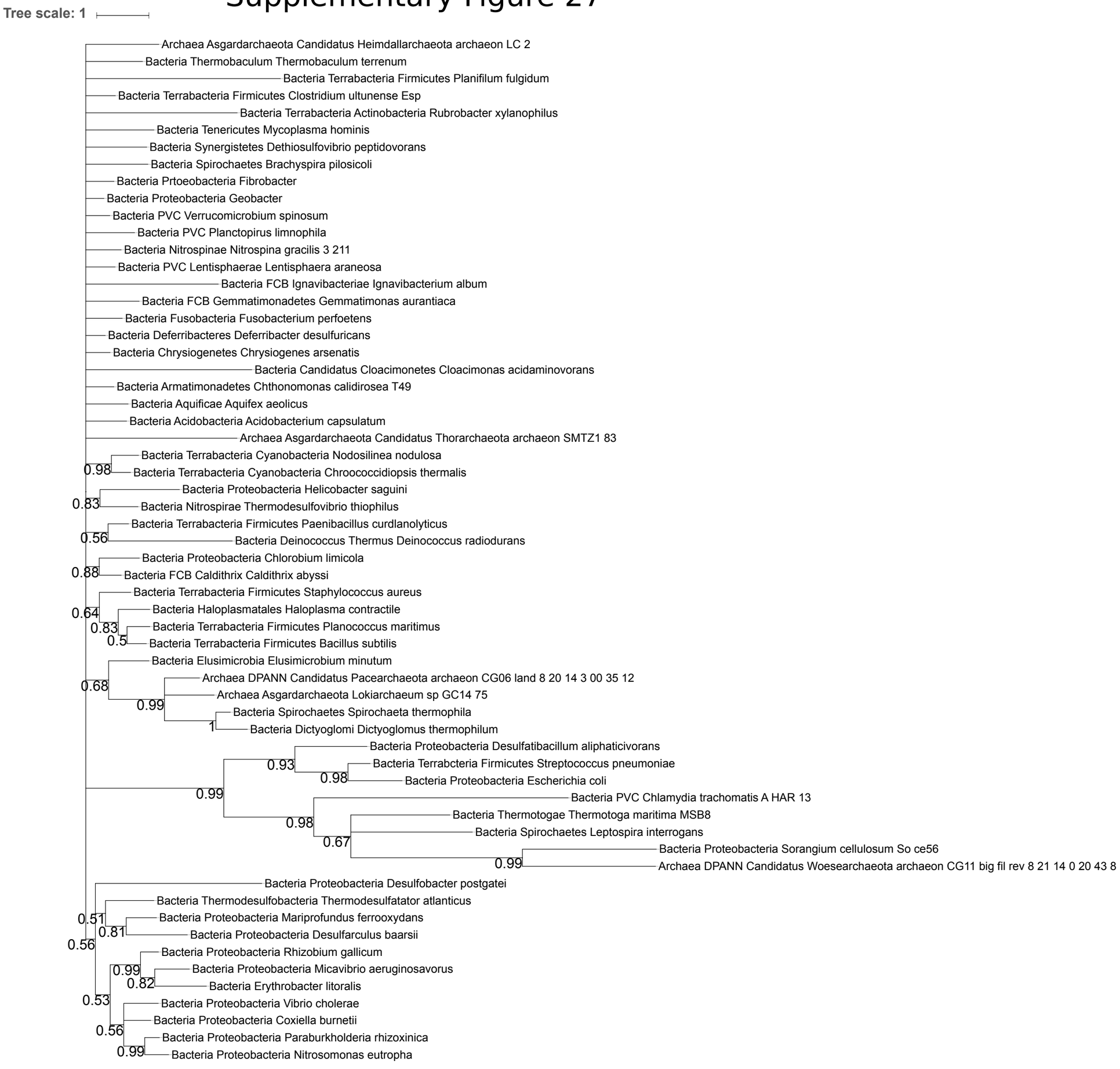
Tree scale: 1

Supplementary Figure 25



Supplementary Figure 26

Supplementary Figure 27

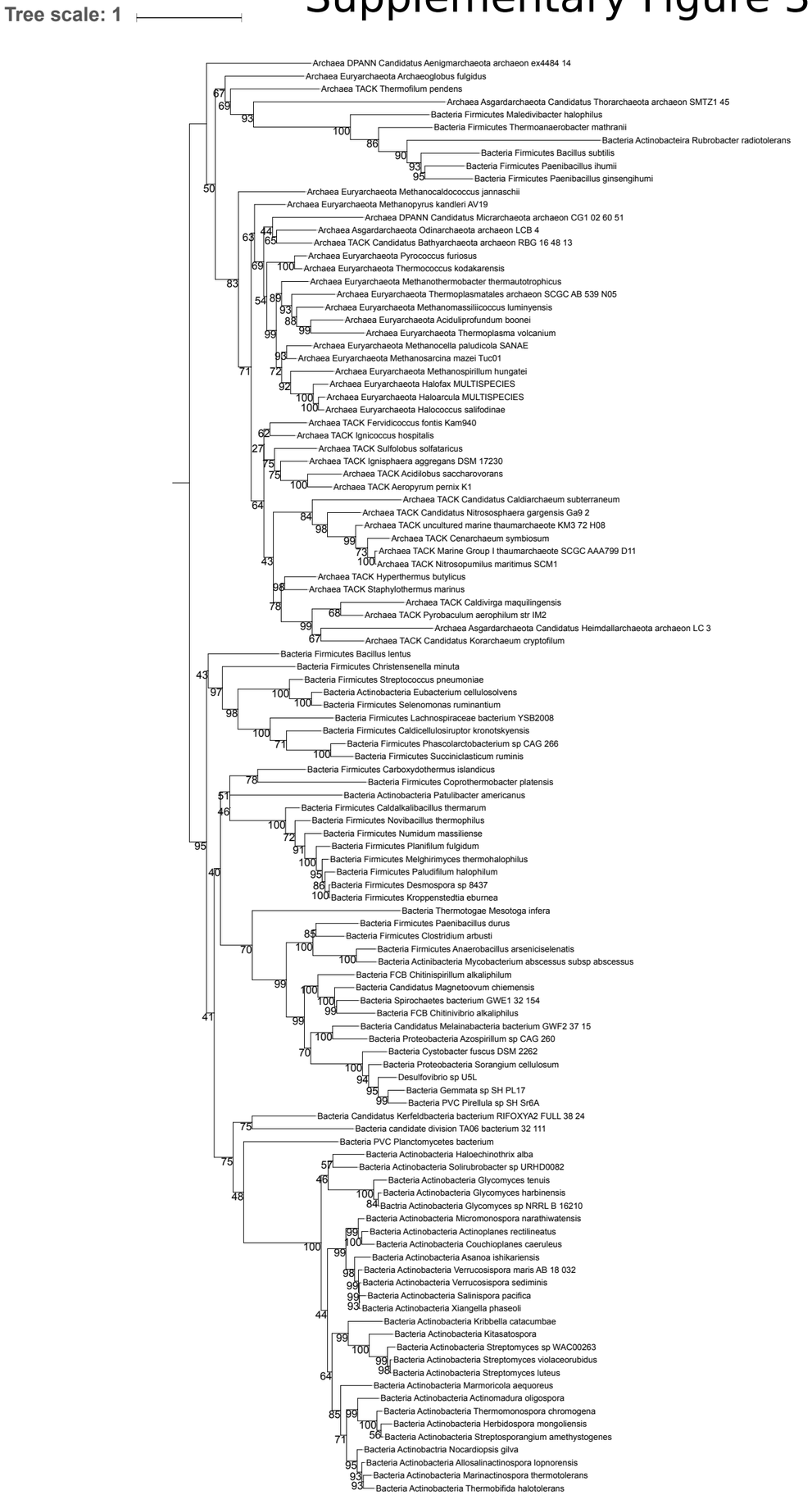


Supplementary Figure 28

Tree scale: 1

Supplementary Figure 29

Supplementary Figure 30

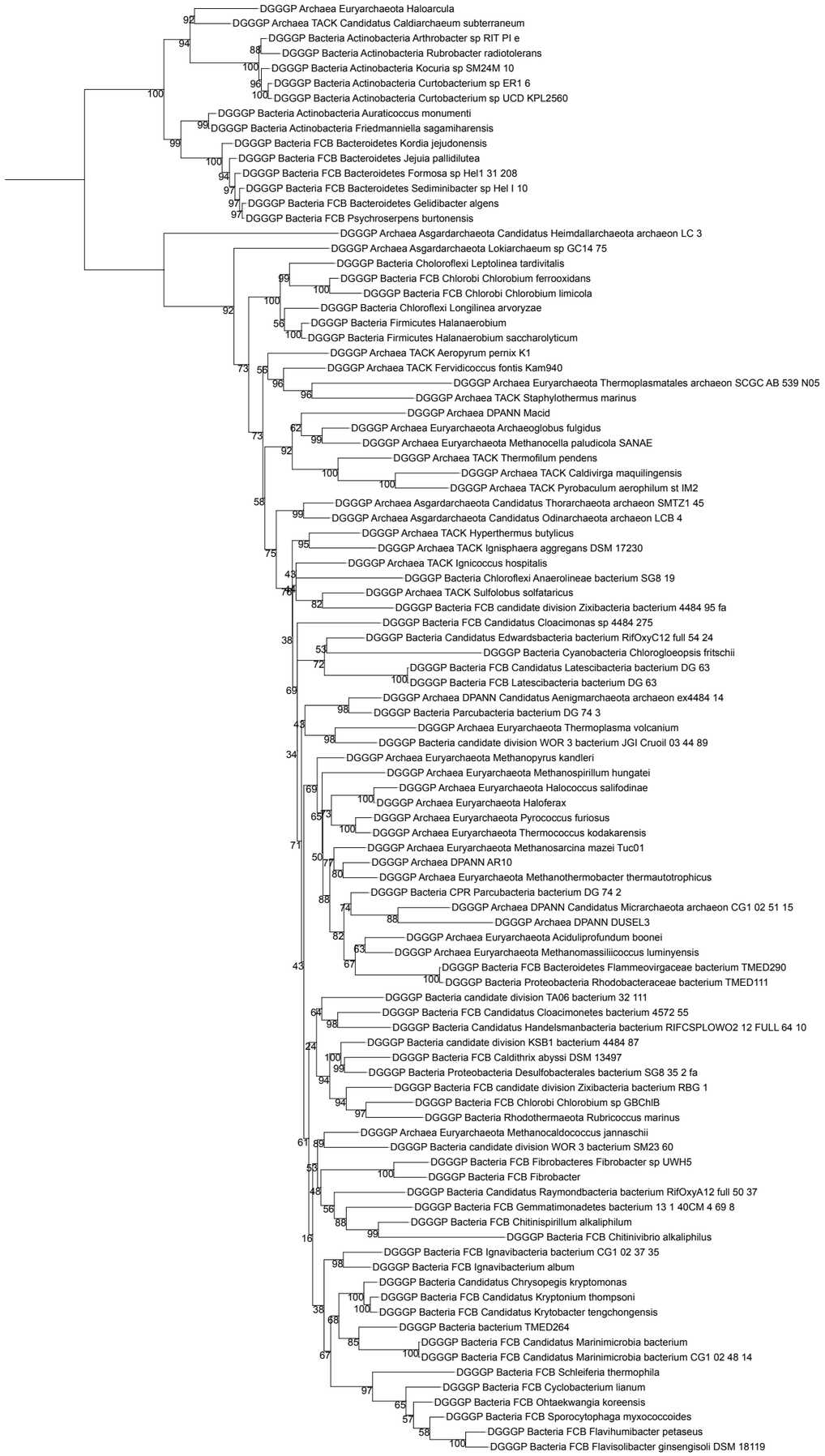


Supplementary Figure 31

Tree scale: 1

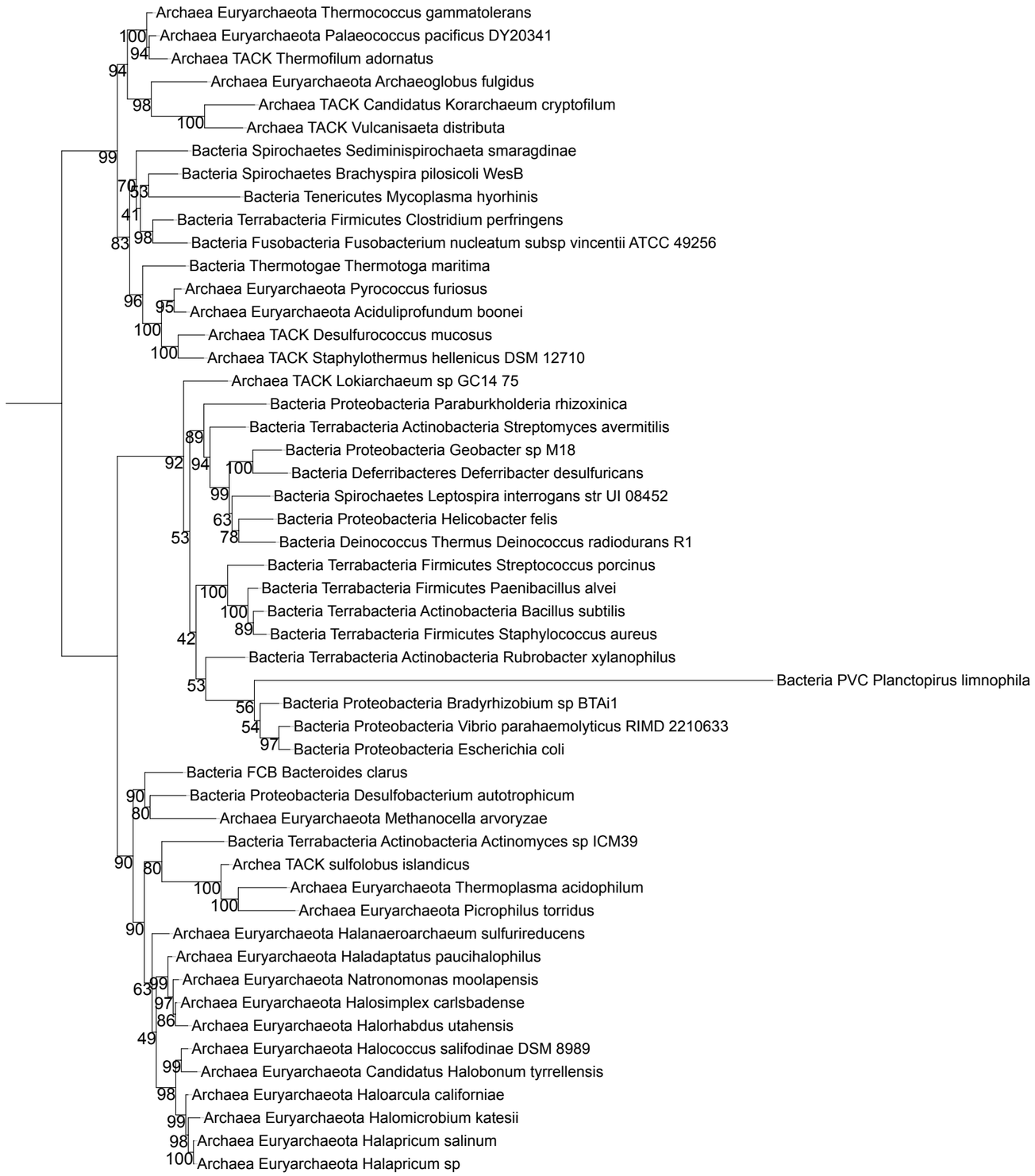
Tree scale: 1

Supplementary Figure 32

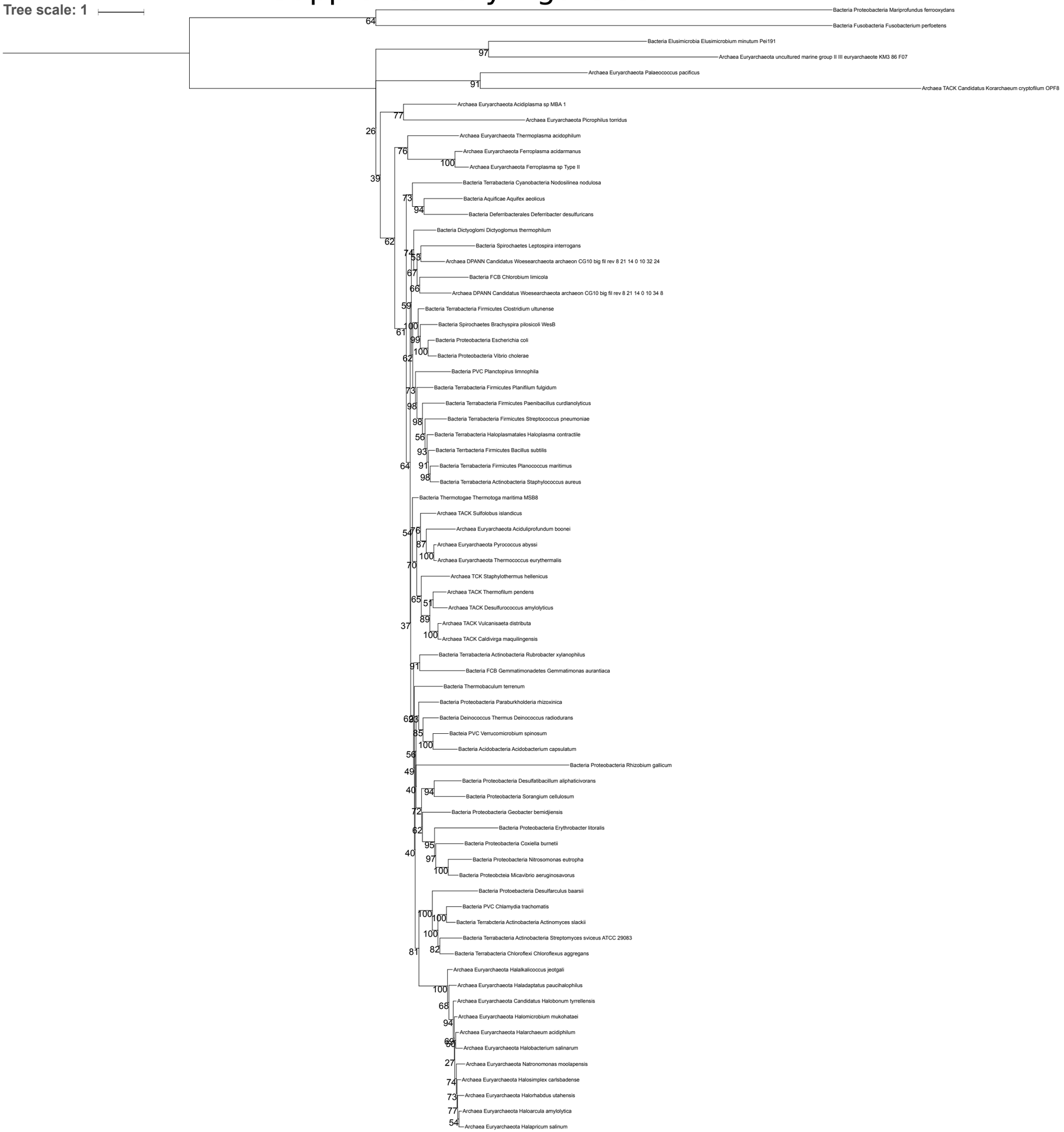


Supplementary Figure 33

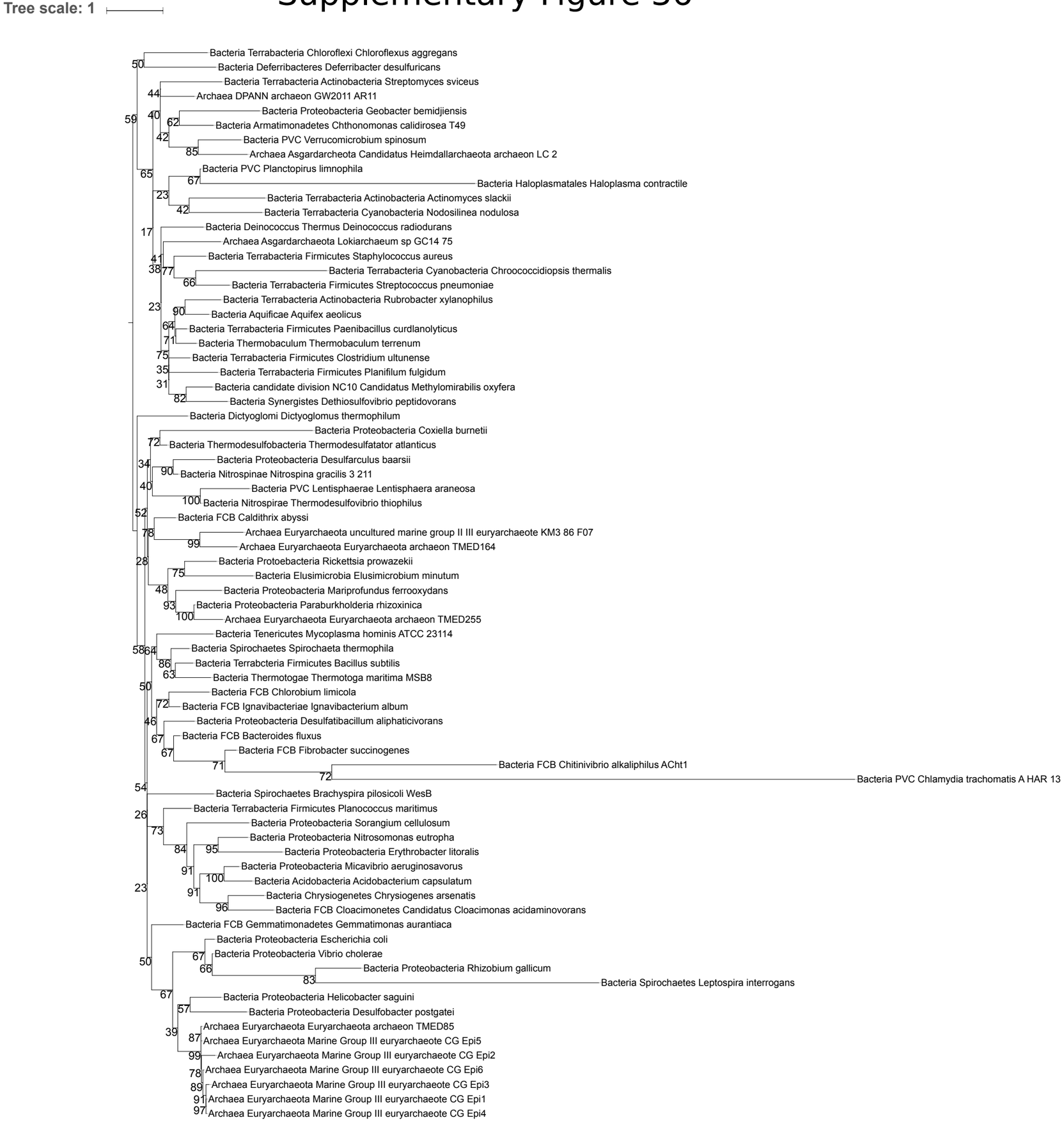
Supplementary Figure 34



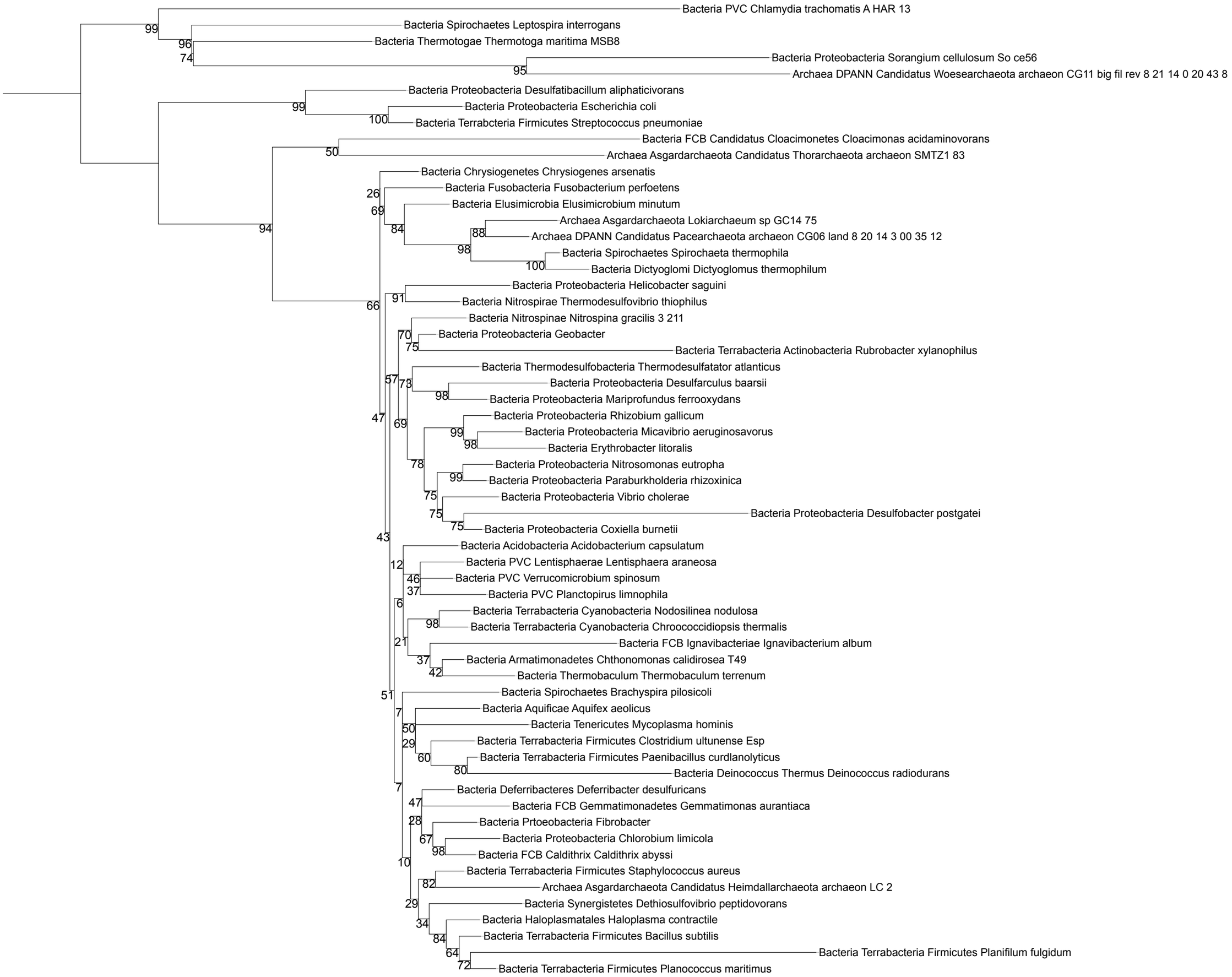
Supplementary Figure 35



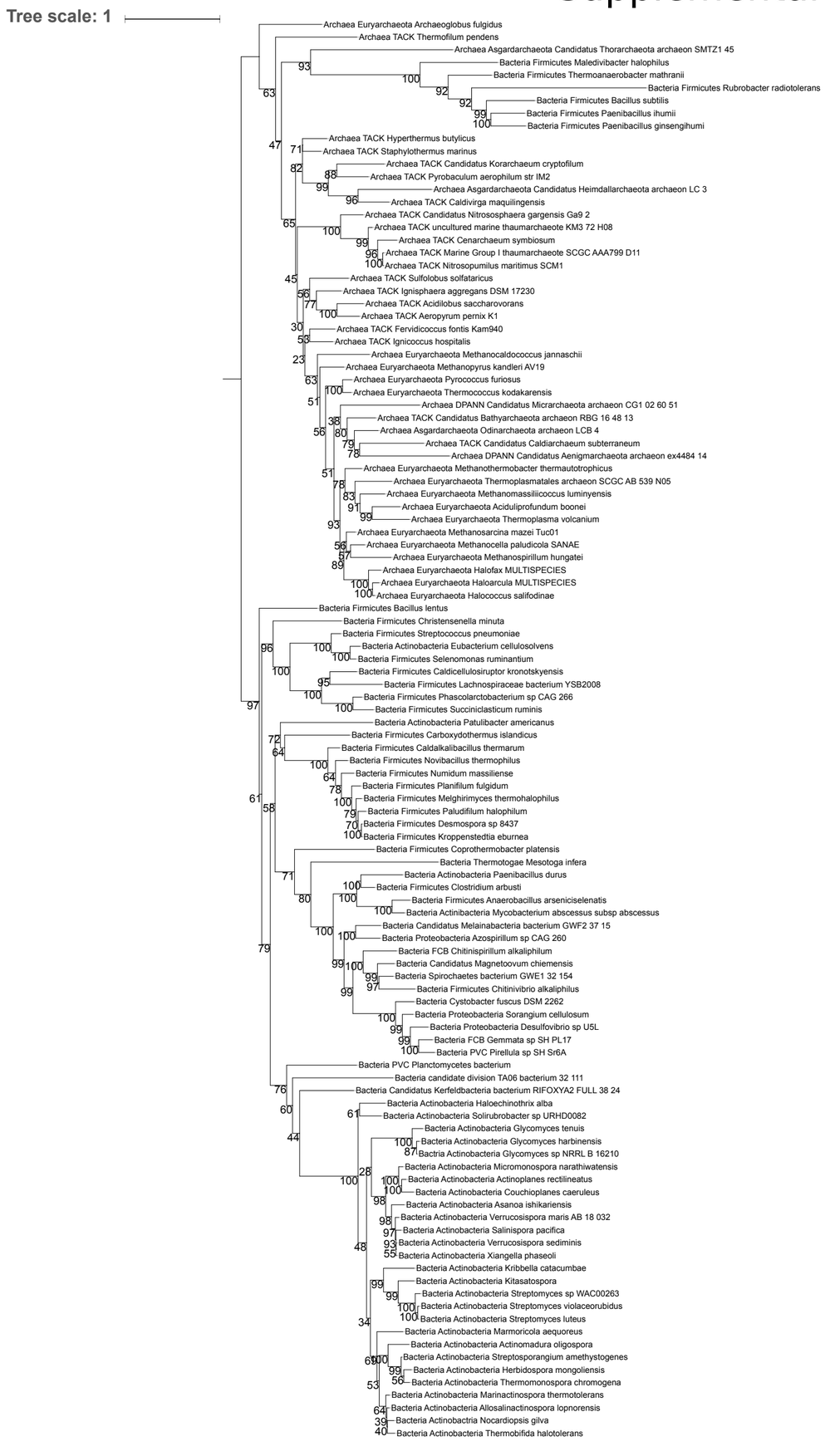
Supplementary Figure 36



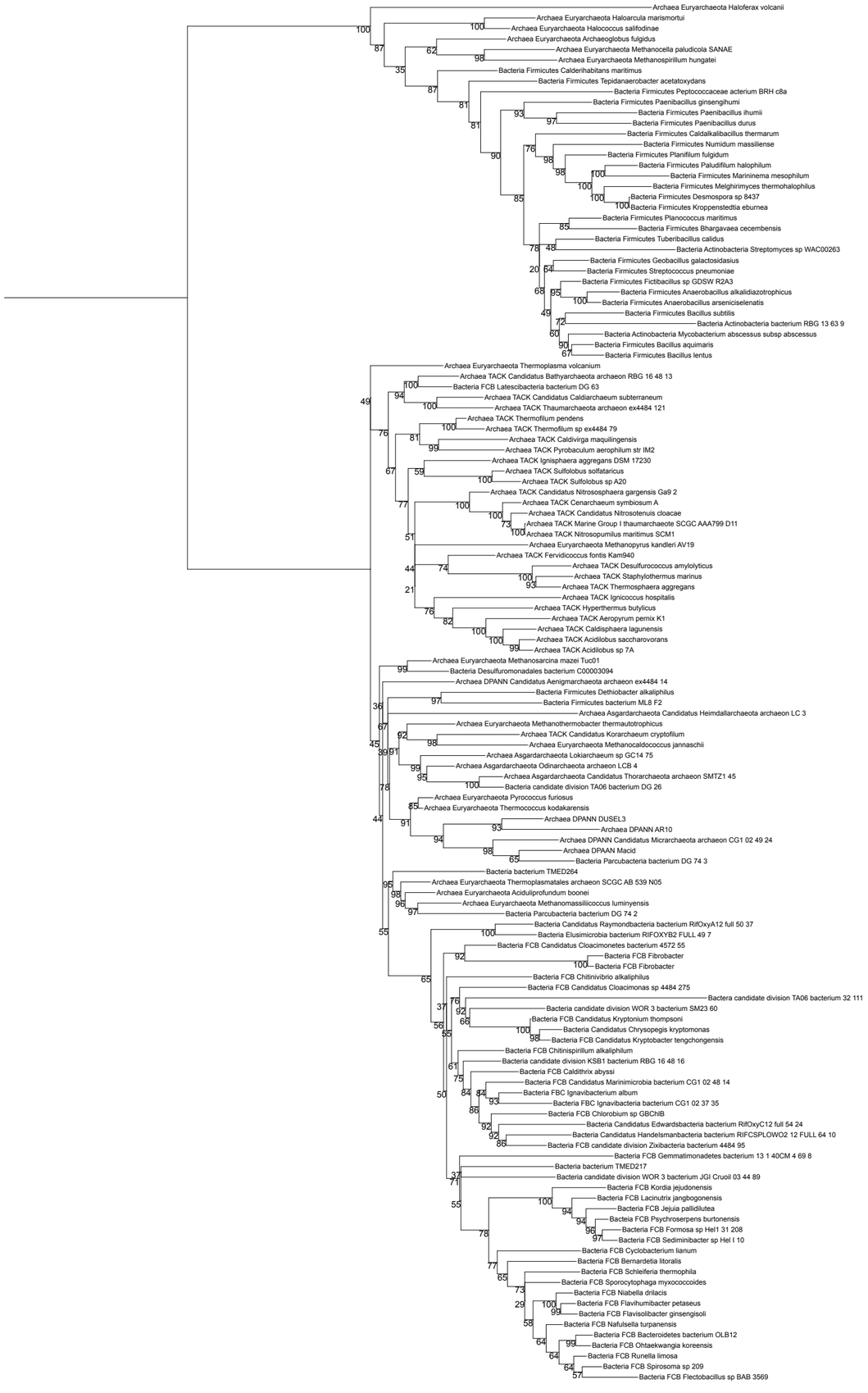
Supplementary Figure 37



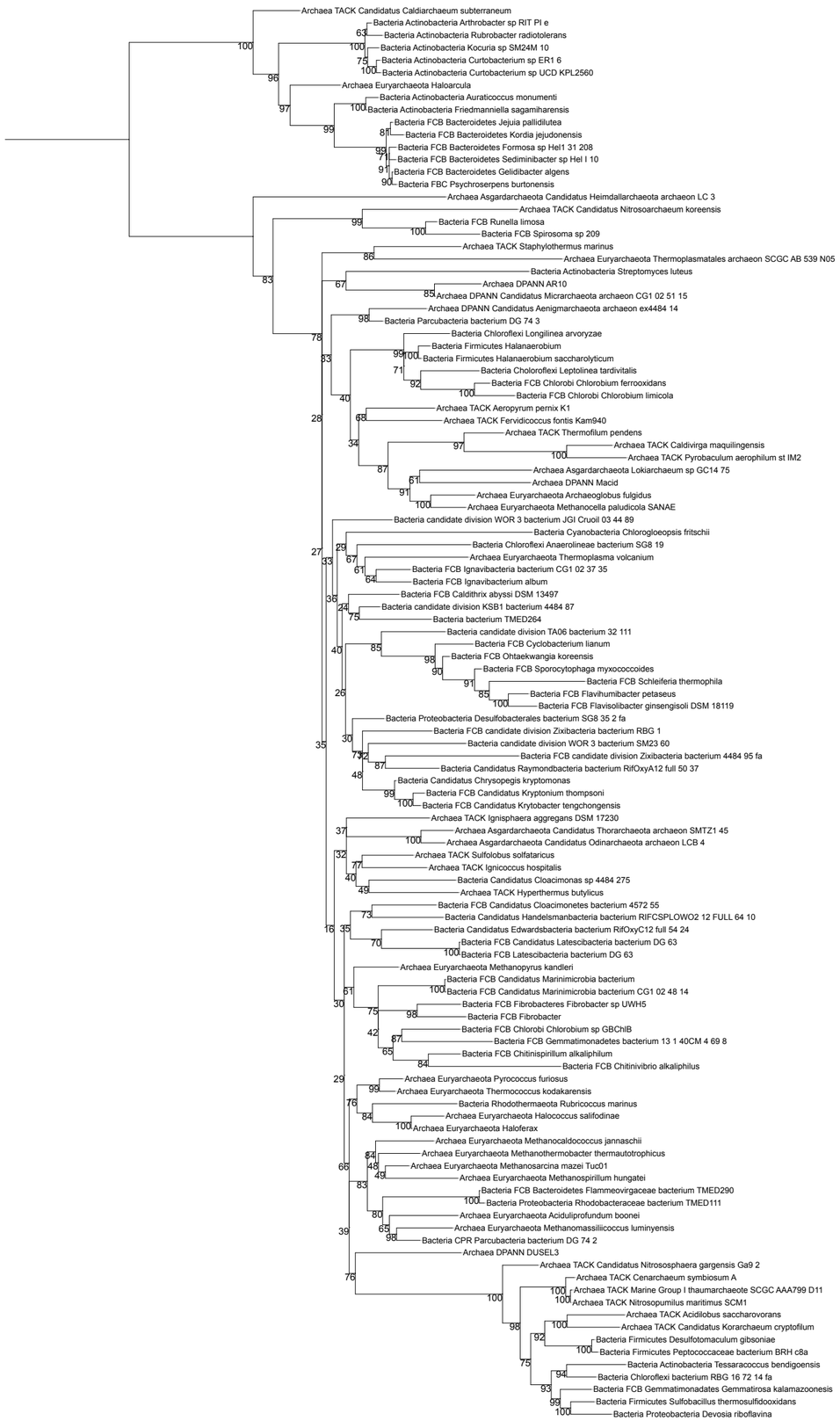
Supplementary Figure 38



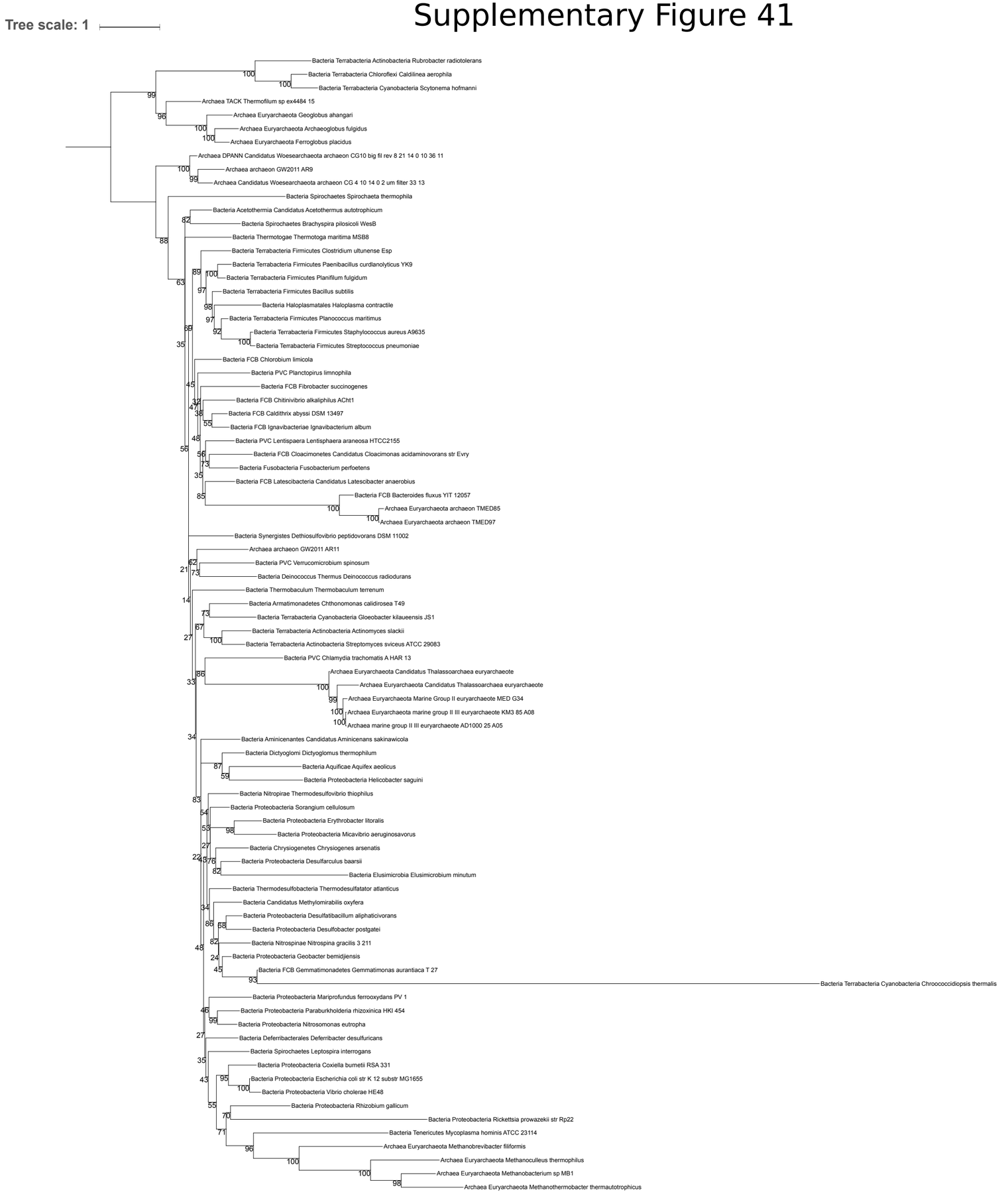
Supplementary Figure 39



Supplementary Figure 40



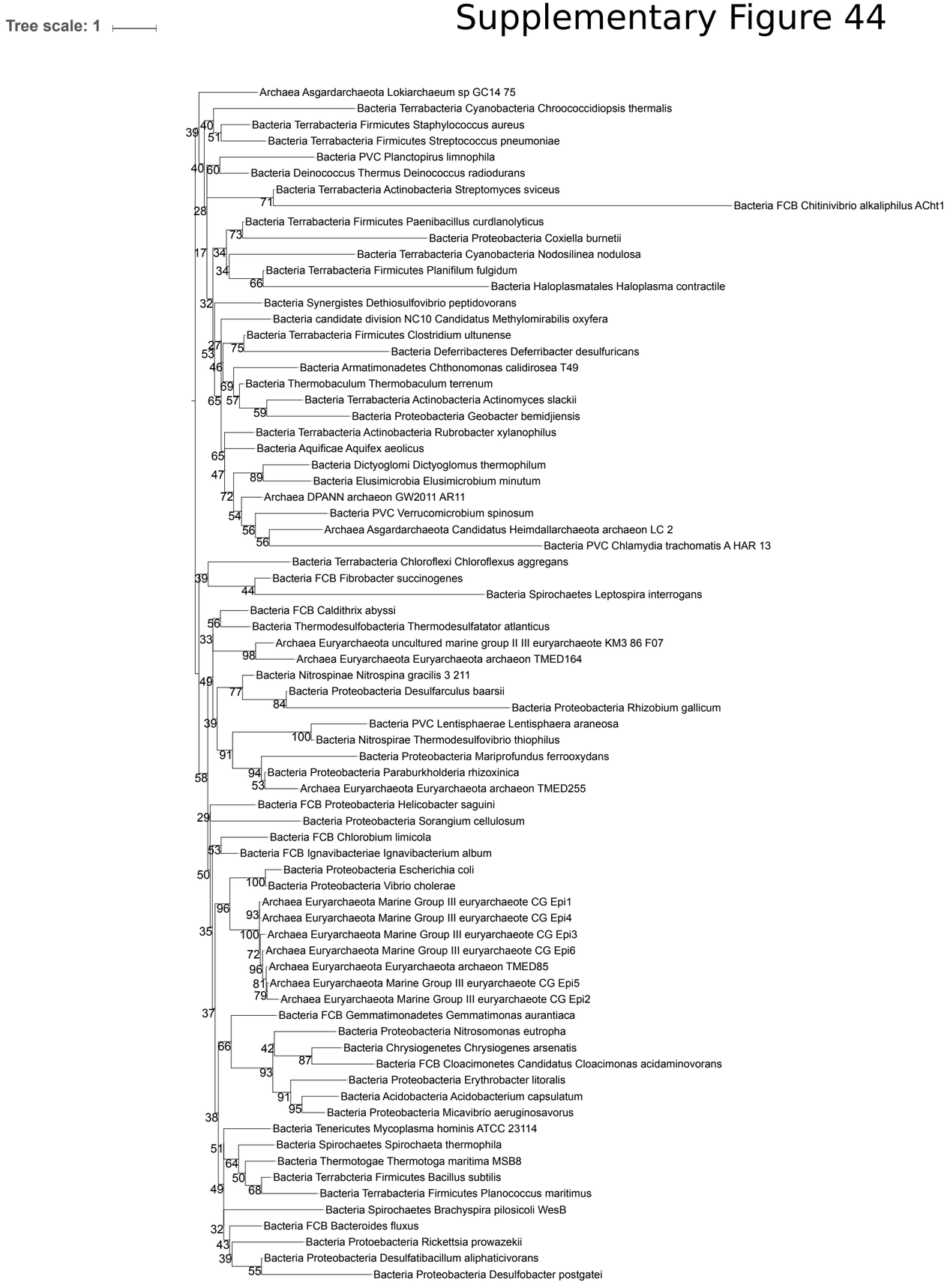
Supplementary Figure 41



Supplementary Figure 42

Supplementary Figure 43

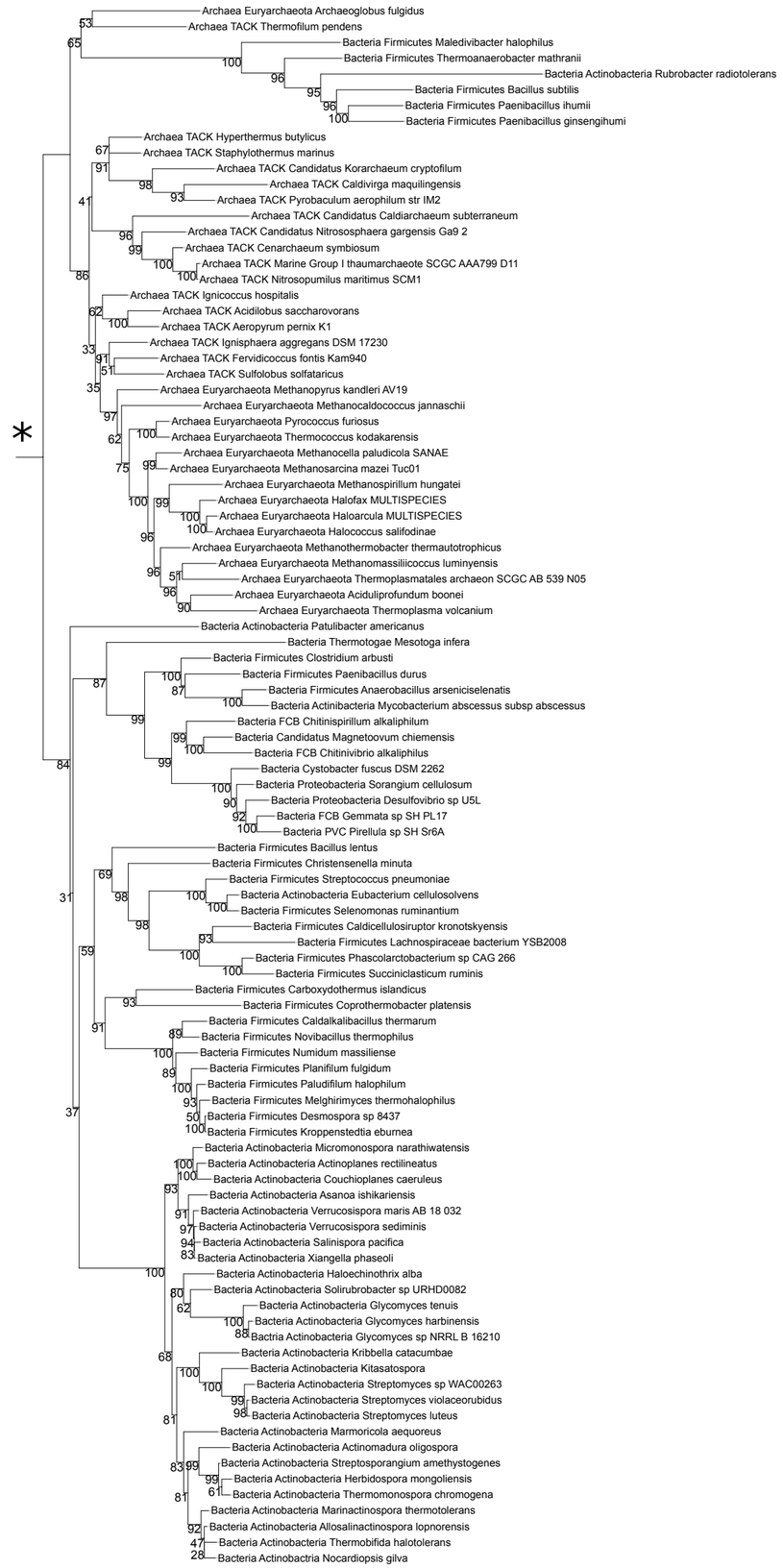
Supplementary Figure 44



Supplementary Figure 45

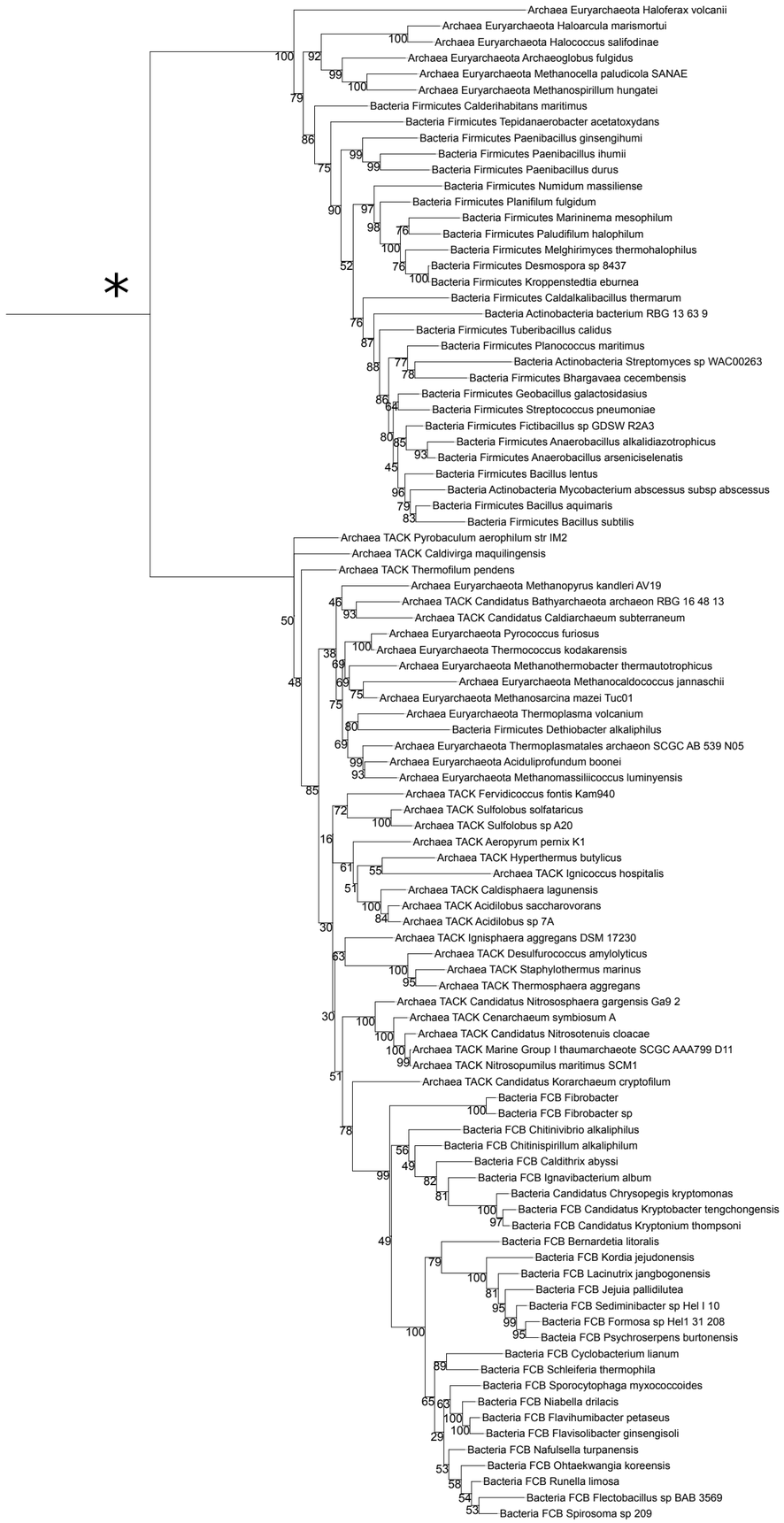
Supplementary Figure 46

Tree scale: 1



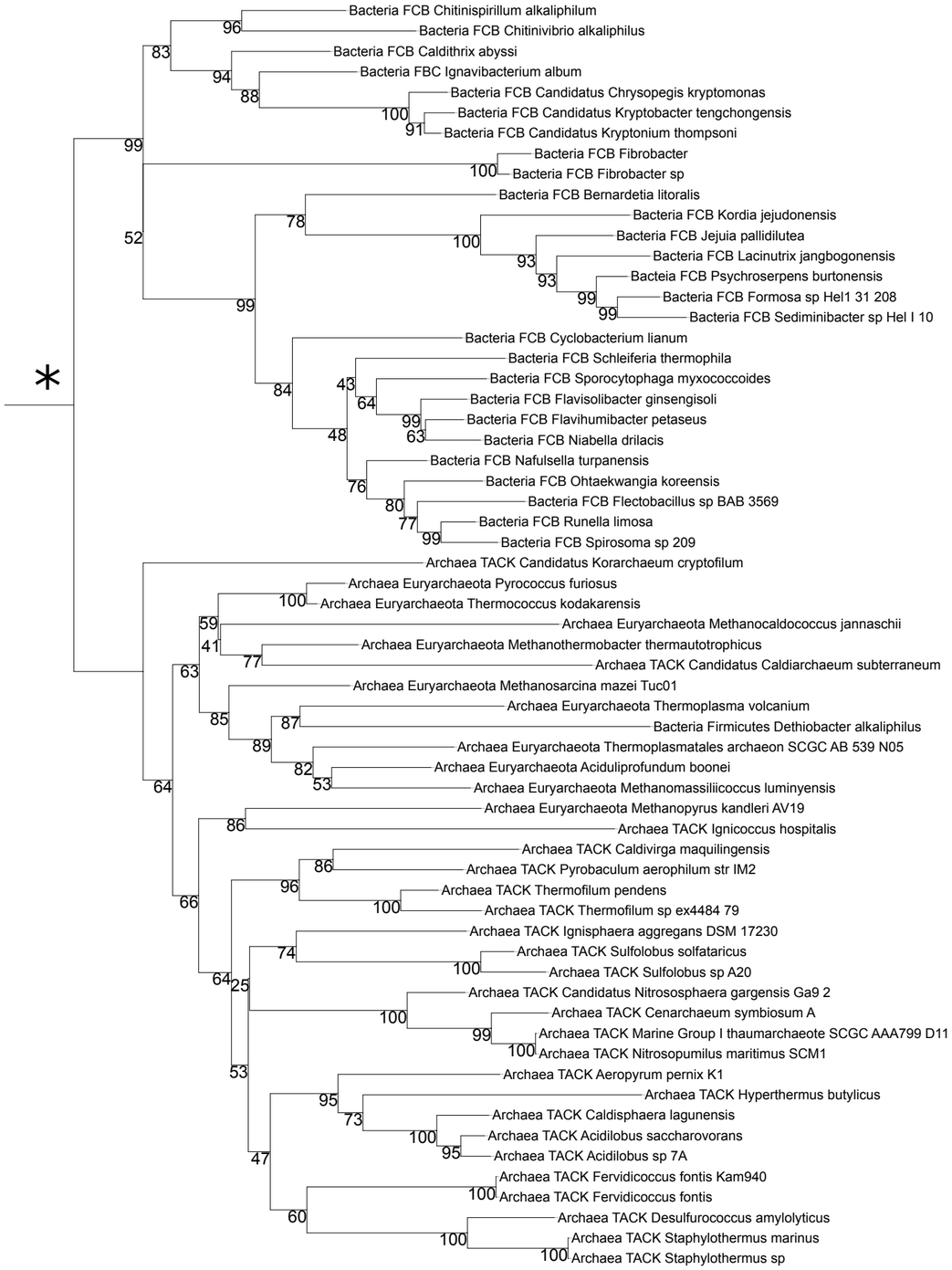
Supplementary Figure 47

Tree scale: 1



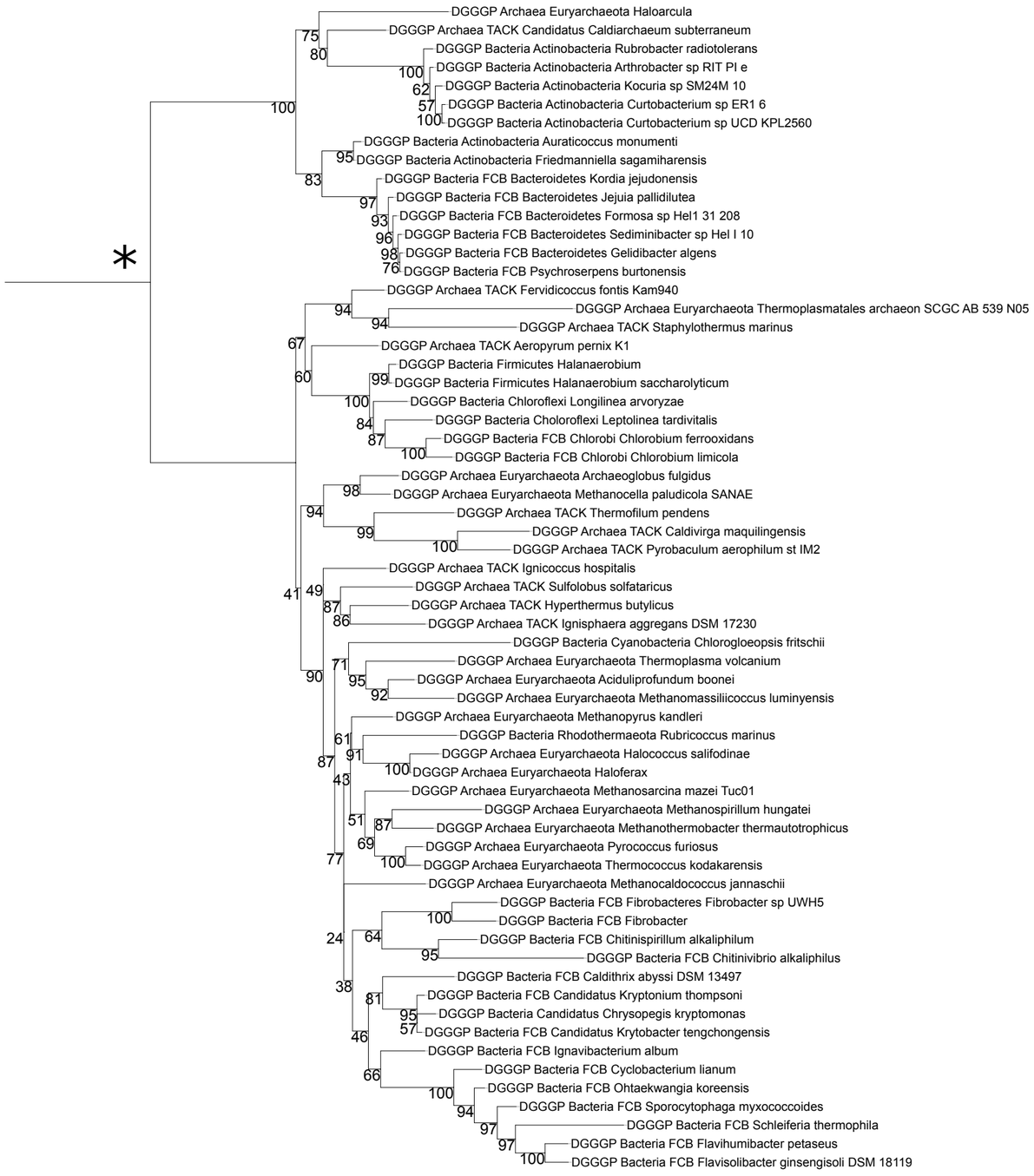
Supplementary Figure 48

Tree scale: 0.1



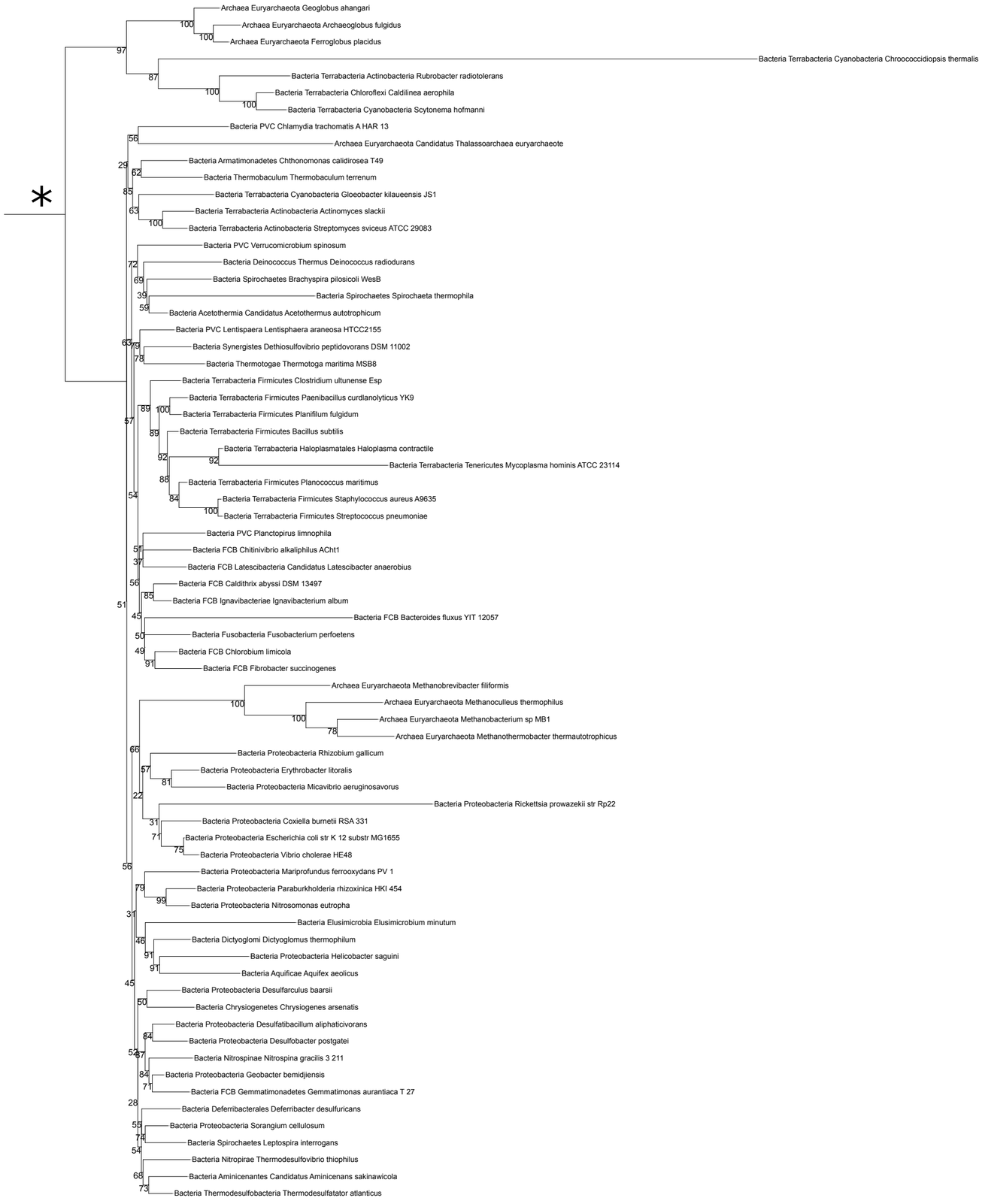
Supplementary Figure 49

Tree scale: 1

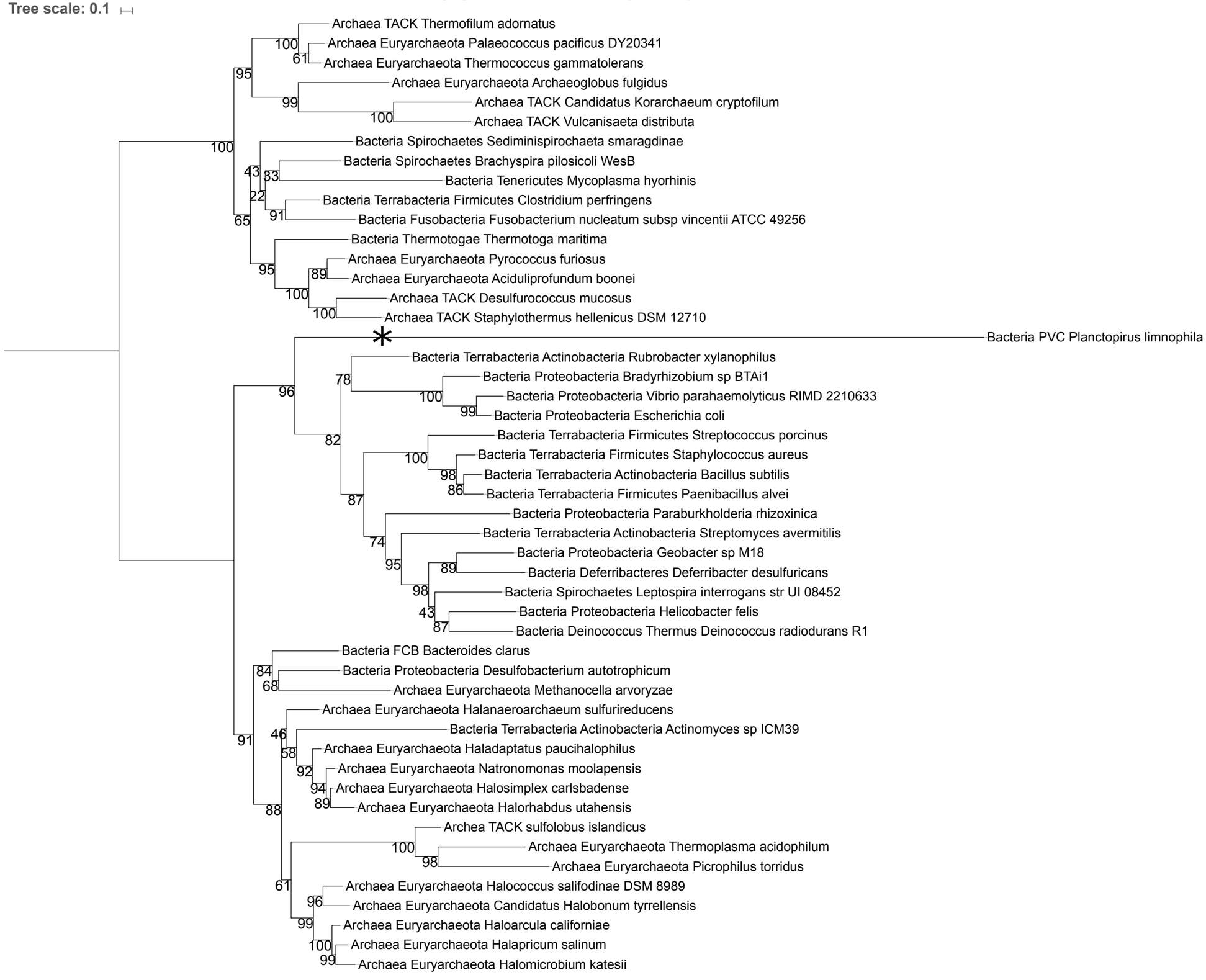


Supplementary Figure 50

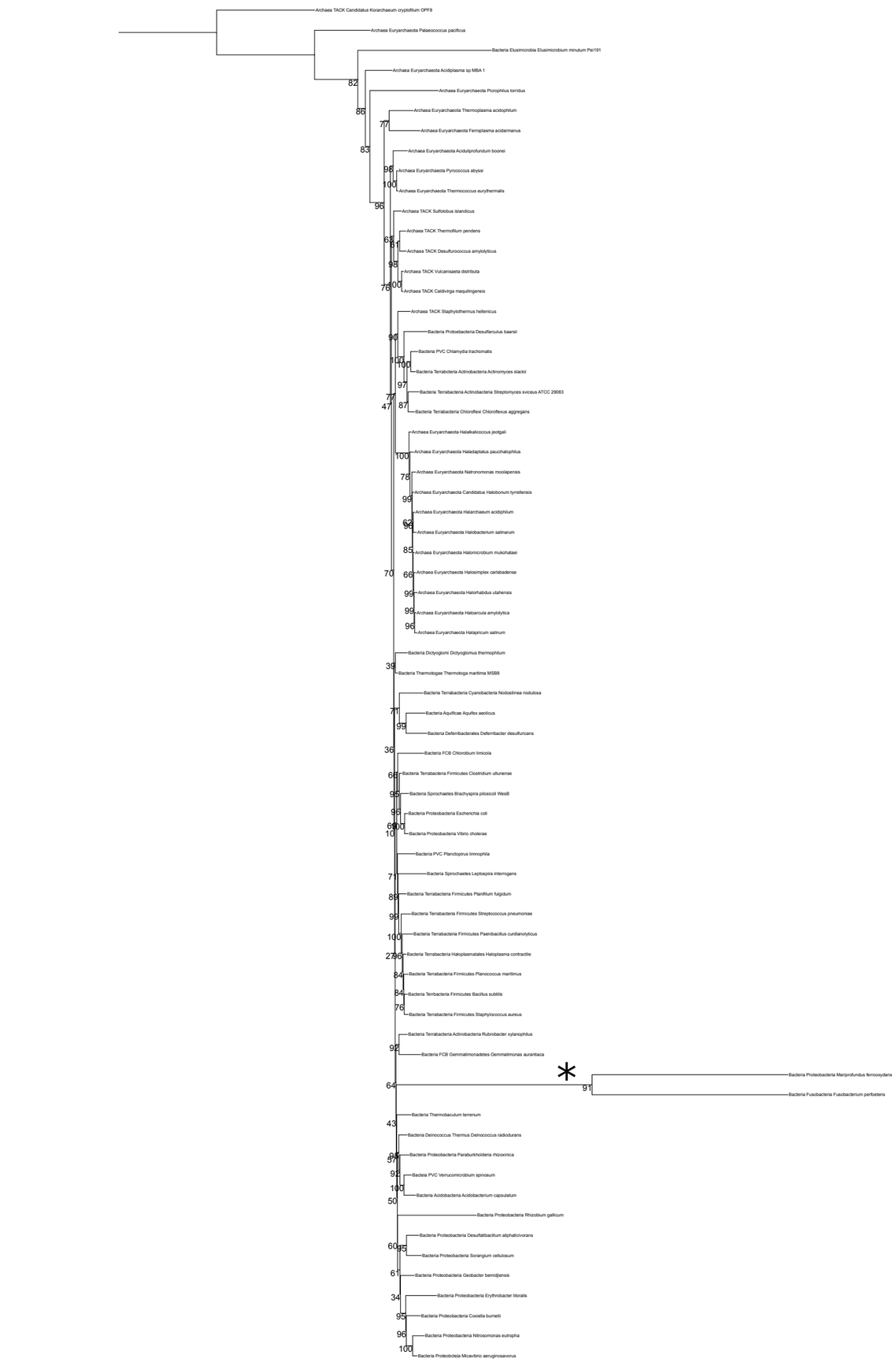
Tree scale: 1



Supplementary Figure 51



Supplementary Figure 52



Appendix D

Papers derived from work included in this thesis.

Coleman, G.A., Pancost, R.D. and Williams, T.A., 2019. Investigating the origins of membrane phospholipid biosynthesis genes using outgroup-free rooting. *Genome biology and evolution*, *11*(3), pp.883-898.

Coleman, G.A., Davín, A.A., Mahendrarajah, T., Spang, A.A., Hugenholtz, P., Szöllősi, G.J. and Williams, T.A., 2020. A rooted phylogeny resolves early bacterial evolution. *bioRxiv*.