Bacterial cell-wall architecture: from automated genome selection to evolution of genes and traits.

by

Raphaël Léonard

Jury:
President: M. Galleni (Université de Liège)
Secretary: P. Meyer (Université de Liège)
Member & Promoter: F. Kerff (Université de Liège)
Member & Copromoter: D. Baurain (Université de Liège)
Member: C. Brochier-Armanet (Université Lyon 1)
Member: D.P. Devos (Universidad Pablo de Olavide)
Member: B. Joris (Université de Liège)

1 Abstract

My aim is to produce possible scenarios for the bacterial evolution based on the bacterial phylogeny and the bacterial cell-wall. For that, we need a selection of genomes which represent the bacterial diversity and are not redundant. However, there is an overabundance of bacterial genomes and most are redundant, so a solution to remove redundant genomes while conserving the bacterial diversity was needed. Yet, none were available when I began my thesis.

I created a tool to automatically cluster genomes and select the best representative for each cluster. The clustering is based on whole genome comparison and the selection considers genome quality, annotation richness, completeness level and absence of contamination. We called my tool ToRQuEMaDA (Tool for retrieving queried Eubacteria, metadata and dereplicating assemblies) or TQMD for short. TQMD is optimized to dereplicate at high taxonomic levels (phylum) but remains competitive while compared to other programs which are optimized to dereplicate at low taxonomic levels (species).

Based on a selection of 903 genomes, we computed orthologous groups (OGs) from which we studied the synteny of the division and cell wall (*dcw*) cluster. Using a smaller selection of genomes, 85, we produced a phylogenomic tree based on the 117 most conserved (and single copy) genes in our selection of bacterial genomes. Using this tree, we reconstructed the *dcw* cluster using an ancestral gene order reconstruction tool and the last bacterial common ancestor (LBCA) cell wall using Bayesian Inference. From our results, it appears that the LBCA was a monoderm already featuring a peptidoglycan layer. We further studied genes involved with the outer membrane (OM) to validate (or invalidate) our results and did not find decisive clues to reject them.

2 Résumé

Mon objectif est de produire des scénarios possibles pour l'évolution bactérienne en me basant sur la phylogénie et la paroi bactérienne. Pour cela, nous avons besoin d'une sélection de génomes représentant la diversité bactérienne et qui ne sont pas redondants. Toutefois, il y a une surabondance de génomes bactériens et la plupart sont redondants, donc une solution pour retirer les génomes redondants tout en conservant la diversité bactérienne était nécessaire. Pourtant, aucune solution n'était disponible lorsque j'ai commencé ma thèse.

J'ai créé un outil qui regroupe automatiquement des génomes homologues et sélectionne le meilleur représentant pour chaque groupe de génomes. Le regroupement se base sur des comparaisons de génomes complets et la sélection du représentant prend en compte la qualité des génomes, la richesse de leurs annotations, leur niveau de complétude et l'absence de contamination. Mon outil est appelé ToRQuEMaDA (Tool for retrieving queried Eubacteria, metadata and dereplicating assemblies) ou TQMD pour faire plus court. TQMD est optimisé pour dérépliquer à haut niveaux taxonomiques (phylum) mais reste compétitif lorsqu'il est comparé aux autres programmes qui sont optimisés pour dérépliquer à bas niveaux taxonomiques (espèce).

En nous basant sur une sélection de 903 génomes, nous avons calculé des groupes orthologues à partir desquels nous avons étudiés la synténie du cluster de la division et de la paroi bactérienne (*dcw*). En utilisant une sélection plus petite de génomes, 85, nous avons produits un arbre phylogénomique basé sur les 117 gènes les plus conservés (et en simple copie) de notre sélection de génomes bactériens. Sur base de cet arbre, nous avons reconstruit la forme ancestrale du cluster dcw avec un outil de reconstruction de l'ordre ancestral de gènes et la paroi du dernier ancêtre commun des bactéries (Last Bacterial Common Ancestor – LBCA) en utilisant l'inférence Bayésienne. D'après nos résultats, le LBCA aurait été un monoderme possédant une couche de peptidoglycane. Nous sommes allés plus loin en étudiant des gènes en lien avec la membrane externe pour valider (ou invalider) nos résultats et nous n'avons pas trouvés d'indices décisifs permettant de rejeter nos résultats.

3 Abbreviations

AA = amino acid
AD = atypical diderm
ALE = amalgamated likelihood estimation
API = analytical profile index
ASTRAL = accurate species tree algorithm
BI = Bayesian inference
BP = bootstrap proportions
BS = bootstrap
CAM = codon aversion motif
CAT = categories
CPR = candidate phyla radiation (= Patescibacteria)
CPU = central processing unit
dcw = division and cell-wall synthesis
DIY = do it yourself
DNA = deoxyribonucleic acid
DPANN = Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaeota
EF = elongation factor
FBC = Fibrobacteres, Bacteroidetes, Chlorobi
Fts = filamentous temperature-sensitive
GlcNAc = N-acetylglucosamine
GTDB = genome taxonomy database
GTR = Generalised time-reversible
HGT = horizontal gene transfer
HMM = hidden Markov model
Hsp = heat shock protein
IGF = identical genome fraction
IM = inner membrane
JI = Jaccard index
JTT = Jones-Taylor-Thornton
LBA = long branch attraction
LBCA = last bacterial common ancestor
LCA = last common ancestor
LG = Le-Gascuel
Lol = localization of lipoprotein
LPS = lipopolysaccharides
MAG = metagenome-assembled genome
MALDI = matrix-assisted laser desorption/ionization
MBN = membrane
MCMC = Markov chain Monte Carlo
ML = maximum likelihood
MRP = matrix representation by parsimony
MSA = multiple sequence alignment
MSC = multispecies coalescent model
MT = monoderm Terrabacteria
MurNAc = N-acetylmuramic acid
NJ = neighbor joining

nm = nanometer
nt = nucleotide
OG = orthologous group
OM = outer membrane
PCA = principal component analysis
PG = peptidoglycan
POTRA = polypeptide-transport-associated
PP = posterior probabilities
PVC = Planctomycetes, Verrucomicrobia, Chlamydia
RGC = rare genomic change
RNA = ribonucleic acid
rRNA = ribosomal ribonucleic acid
S-layer = surface layer
SEDS = shape elongation division sporulation
SSU = small subunit
TDL = true diderm-LPS
TOF = time of flight
tRNA = transfer ribonucleic acid
UL = unsupervised learning
WAG = Whelan and Goldman

4 Introduction

4.1 Of prokaryotes, shapes and cell-walls

The first discovery of microorganisms dates back to the 1670s by Antony van Leeuwenhoek, a textile merchant and a builder of microscopes. His discoveries and descriptions of small organisms he called “animalcules” are considered to be the base of microbiology. Others have preceded him but he is the one history has remembered.

4.1.1 Shapes

Initially, bacteria were classified according to their shapes. **Cocci** are spherical, and their name is adapted depending on the way they agglomerate. A lone spherical organism is a **coccus**. If they form pairs, each organism is called a **diplococcus**. The **tetrad** is an agglomeration of four cells forming a square, while the **octet** (or *Sarcinae*) is made of eight cells forming a cube.

Another shape is the rod or **bacillus**, which again has different names depending if cells are agglomerated or not and how they are agglomerated. A single rod cell is called a bacillus, if they form a chain of two cells, each organism is called a **diplobacillus**. The rod-shaped organisms can also be named after the form of the rod. Four common shapes exist, including the **coccibacillus**, of which cells are too long to be considered a coccus but not long enough for a bacillus, and the **vibrio**, having cells with the shape of a comma. The other two most common shapes are the **spirillum** and the **filament**; both are single long cells, the spirilla showing regular curves and the filamentous bacteria being long but not necessarily curved.

4.1.2 Cell walls

In the XIXth century, several scientists made new discoveries about microorganisms (or microbes). Apart from Pasteur and Koch, the discovery of Gram **staining** in 1882 by Hans Christian Gram (published in 1884)[1] is one of the major advances in bacteriology. At that time, the Gram staining procedure allowed investigators to differentiate between the two known main types of cell wall (i.e., all layers surrounding the plasma membrane).

The first type of bacteria has only a thick layer of a polymer called **peptidoglycan** (**PG**) to protect their **plasma membrane**; they are designated as **Gram positive** bacteria or **monoderms**[2]. The second type has two layers of protection: a thin layer of PG outside the plasma membrane and a **second membrane** outside the PG layer; they are designated as the **Gram negative** bacteria or **diderms**[2]. The space between the plasma membrane and the PG for the monoderms and the space between the plasma membrane and the outer membrane for diderms is called in both cases the **periplasm**.

The Gram staining method[3] consists in adding a dye, crystal violet, to the bacteria to be identified. Both types of bacteria absorb the dye. Then a fixative is used to trap the dye into the cell. A mixture of ethanol and acetone is used to decolorize the bacteria and wash the excess of dye. The last step consists in adding a second dye, safranin, to the mix and again washing the excess dye. Monoderm bacteria do not lose the first dye when washed and are thus termed Gram positive. In contrast, diderm bacteria lose the first dye during the wash and are thus termed Gram

negative. The second dye serves to colorize the Gram negative bacteria again in order to facilitate their spotting.

Strikingly, this technique is still used today because it makes the bacteria easier to observe in light microscopy, is a cheap and fast test for diagnostics and is still the first step for the identification of a bacteria. However, it has limitations. First, there are more than two cell-wall types within bacteria. Second, some “monoderms” can stain negatively, whereas some ” diderms” can stain positively with the Gram staining. Third, the cell-wall architecture itself is not a reliable tool for classification. The following steps for the identification of a bacteria can be the use of an **Analytical Profile Index** (**API**)[4] gallery (a series of biochemical tests allowing for a fast identification of bacteria) or the use of the **MALDI-TOF** (Matrix-Assisted Laser Desorption/Ionization - Time-Of-Flight) mass spectrometry to compare the obtained profile to a collection of profiles[5].

4.1.3 Bacteria and Archaea

The bacteria were firmly separated from the rest of the living organisms in 1962 by Stanier and Niel[6]: bacteria belong to the **prokaryotes** and the rest of the living organisms to the **eukaryotes**. Prokaryotes are defined as **unicellular organisms without a nucleus** to house their genome and **without any organelles** (a specialized subunit within a cell delimited by a lipid bilayer; e.g., mitochondria and chloroplast), while eukaryotes are defined as organisms, uni- or multicellular, with a nucleus around their genome and also with organelles. Prokaryotes moreover divide most of the time by binary fission, instead of mitosis. Interestingly, the distinction between prokaryotes and eukaryotes had been made earlier by Chatton [1938][7] but not considered groundbreaking at the time[8].

At first, the organisms later known as methanogenic Archaea were considered to be peculiar extremophile bacteria. In 1977, based on his analysis of the **RNA** of the small subunit of the **ribosome** (**SSU rRNA**), Woese[9] proposed to classify the Archaea as being different from Bacteria and Eukaryota. Early on in the study of the cell wall, a glycan layer similar to the glycan layer of the bacteria, the **PG/murein**, had been identified in Archaea. However, further analyses revealed differences between the two macromolecules, which resulted in the naming of the archaeal layer as **pseudopeptidoglycan** or **pseudomurein**[10].

In 1990, based on molecular characters, Woese, Kandler and Wheelis[11] further separated the two domains into three domains by dividing the prokaryotes into two groups, the Bacteria and the **Archaea**. They also renamed the eukaryotes to Eucarya/Eukaryota.

4.1.4 Peptidoglycan, S-layers and pseudopeptidoglycan

The bacterial PG layer is a mesh-like molecule surrounding the plasma membrane. It is composed of glycan chains of alternating **N-acetylmuramic acid** (**MurNAc**) and **N-acetylglucosamine** (**GlcNAc**) units linked via β1-4 bonds. The glycan chains are connected by peptide bridges with

an alternance of L and D-amino acids. In Archaea pseudoPG, the glycan strands are made of **N-acetyl-L-talosaminuronic acid** units linked via β1-3 bonds to **N-acetylglucosamine** units with bridges made of L-amino acids[12]. Soon, it was realized that this structure was not a common occurrence among Archaea, instead only organisms belonging to Methanobacteriales and Methanopyrales have it[13].

Most Archaea have a proteinaceous layer that surrounds the cell (termed the **S-layer**)[14], a few have a cell wall made of polymers and some do not even have any of the two[12]. The most common cell wall consists of a plasma membrane surrounded by an S-layer. In both Bacteria and Archaea, S-layers are composed of only one or, in a few cases, two different (glyco)proteins[15]. Some Archaea have additional components in their cell wall, which can occur either above or below the S-layer. Examples of such components include the already mentioned pseudoPG (e.g., in *Methanothermus fervidus*), the methanochondritin layer (trimer of proteoglycans) of *Methanosarcina mazei* Go1, and the proteinaceous sheath of *Methanospirillum hungatei* JF-1. Illustrations of such cell-wall architectures are shown in Figure 1 taken from[12].

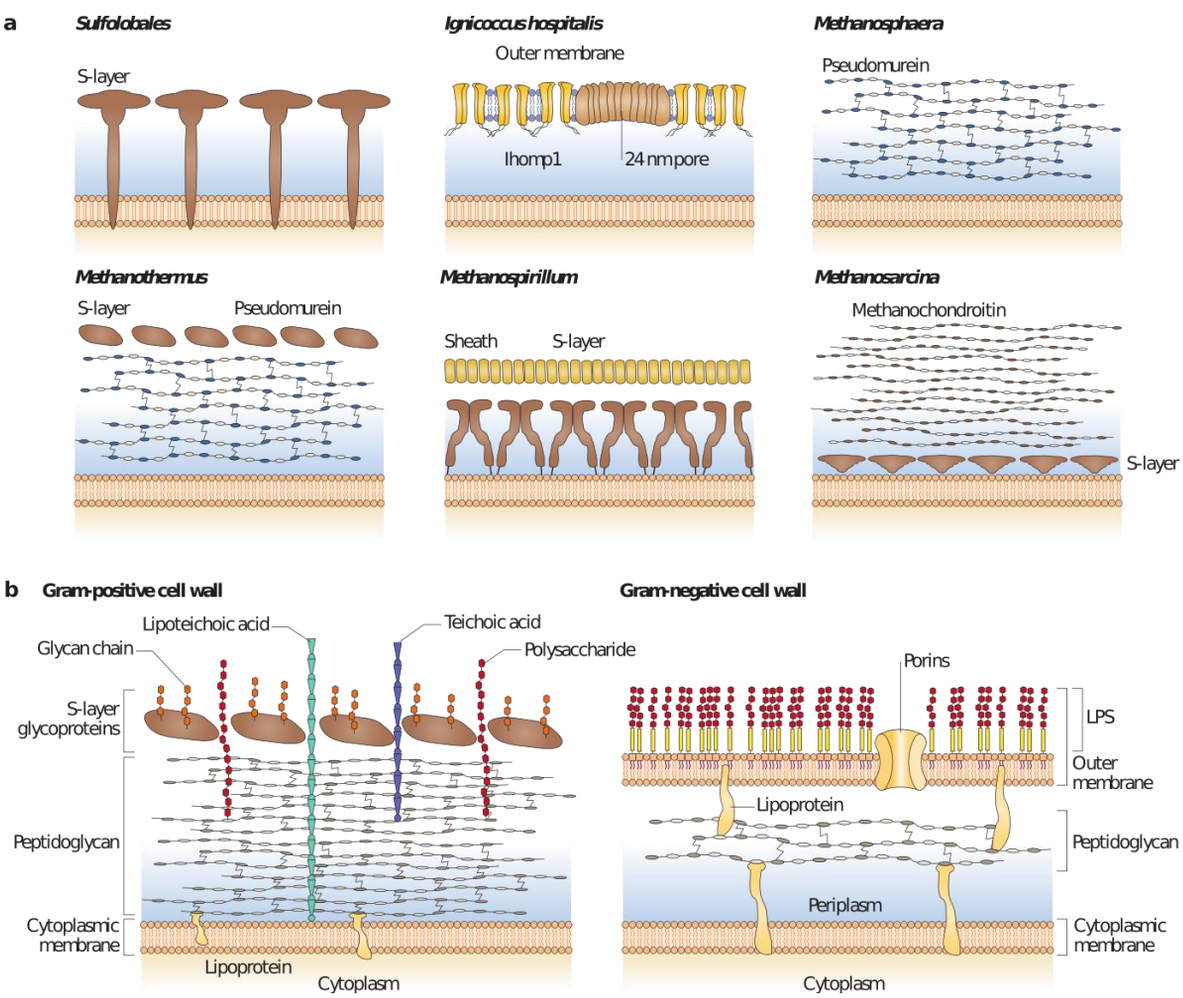


Figure 1 (adapted from Albers 2011[12], Figure 2): a | Schematic side view of cell-wall profiles from different archaea. Pseudoperiplasmic space is shown in blue. b | Schematic of bacterial cell-walls. Gram-positive bacteria have a thick multilayered coat of PG. Gram-negative bacteria have an outer asymmetric bilayer membrane and a thin PG layer. CM, cytoplasmic membrane; SL, S-layer. [Legend modified from[12]].

Another difference between Archaea and Bacteria is the composition of the cytoplasmic (or plasma) membrane. In Bacteria, this **lipid bilayer** is constituted of lipids with **two fatty acid chains** and a **hydrophilic head** usually containing a **phosphated D-glycerol** (Figure 2a). The hydrophobic fatty acid chains are buried in the inner part of the membrane, while the hydrophilic heads form the outer part. The D-glycerol and the two fatty acid chains are linked by an **ester bond**. In contrast, in Archaea, two different types of **phospholipids** co-occur in the same bilayer. The first has its fatty-acid chains replaced by **isoprenoid chains** and is linked to a **L-glycerol head** by an **ether bond** (Figure 2c). The other one consists in the fusion of two archaeal phospholipids to form a single phospholipid as long as the bilayer height (Figure 2b).

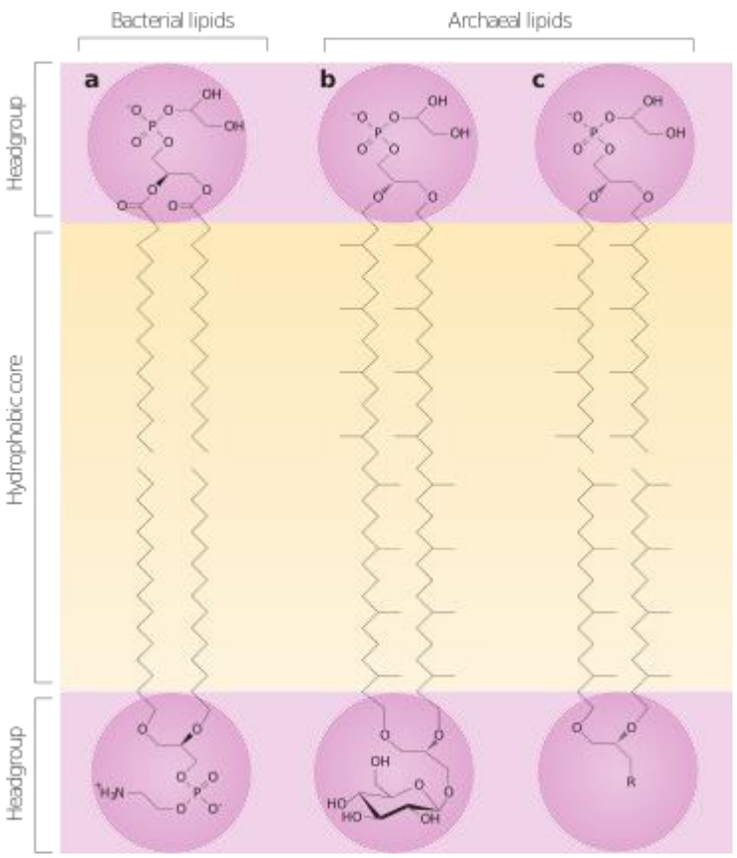


Figure 2 (from Albers 2011[12], box 1): (a) Bacterial bilayer-forming lipids are phophatidylglycerol (upper lipid) and phosphatidylethanolamine (lower lipid). (b) Structure of monolayer-forming tetraether lipids. (c) Bilayer formed of archaeal diether lipids. More details are available in [16].

4.2 Of bacterial classification approaches

Classic classification is based on morphology (such as the type of cell wall or the shape of the cell). This approach is similar to the classification of animals based on morphology. While it worked “good enough” in the case of animals, due to the diversity of possible morphologies within the same group of prokaryotes, this approach does not attain the “good enough” level of accuracy. It is due to the fact that prokaryotic morphologies cannot be compared between themselves due to the high level of diversity within the same group. Nowadays, we shifted to **molecular phylogenetics**, which rely on **genetic markers** and are more accurate.

4.2.1 Single-gene phylogenies using rRNA and orthologous proteins

The RNA from the small subunit of the ribosome (SSU rRNA 16S for the prokaryotes and SSU rRNA 18S for the eukaryotes) is a macromolecule common to every living organism. This universality, associated with a strong conservation linked to its core function, makes it appropriate for **phylogenetic** analyses aimed at resolving high-evolutionary level relationships (deep phylogenetics). This is the type of analysis that led Woese[9] to propose a division of life into three domains: Bacteria, Archaea and Eukaryota.

Woese decided to split the prokaryotes (cellular organisms without a nucleus) into two domains, due to their partition into two well-separated groups in SSU rRNA (16S) phylogenetic trees. Yet, this schism is corroborated (among other features) by morphological differences in their cell wall, as explained above. It is of note that the distance separating the two groups of prokaryotes is comparable to the distance that separates each of them from the eukaryotes, making any regrouping of higher order difficult.

SSU rRNA can be used to study these three domains more in depth, but this molecule does not provide clear answers regarding the relationships between the main subgroups within each domain. Indeed, the oldest evolutionary events are difficult to reconstruct, both because the speciations may have occurred on a short time-scale, thus preventing the examined molecule, of a limited size, to record an exploitable **phylogenetic signal**[17], and due to the accumulation of subsequent **multiple substitutions** in the same sites of the molecule, thereby erasing the signal, potentially already tenuous[18]. Such phylogenetic trees using the SSU rRNA can be found in the last version of the Bergey's Manual of Systematics of Archaea and Bacteria (edition 2015, last version as of September 2021)[19] but the inter-phyla relationships are poorly resolved (e.g. Figure 3) and these trees are thus of limited use for high-order classification purposes.

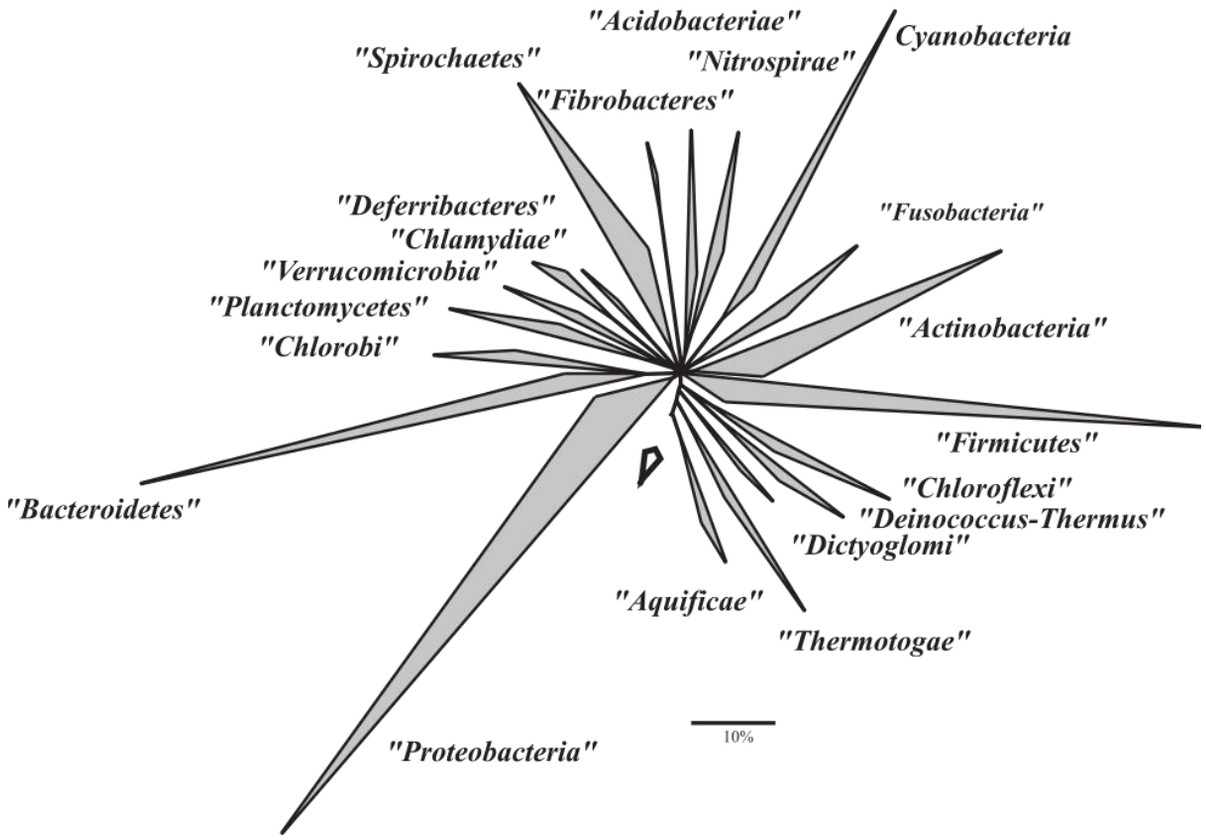


Figure 3 (from Ludwig et al. 2001[19]): 16S rRNA-based tree reconstructed with the ARB parsimony tool[20] using only sequence positions sharing identical residues in at 50% of all sequences.

Beyond SSU rRNA, it is also possible to construct phylogenetic trees using **orthologous protein** sequences (corresponding to the same gene across many organisms), as in the study by Woese[21] on the **aminoacyl-tRNA synthetases**. But these single-gene protein trees, while useful for corroboration, often suffer from the same lack of signal as rRNA. An example of these single-gene protein trees can be found in Baldauf et al. (1996)[22] and is reproduced in Figure 4.

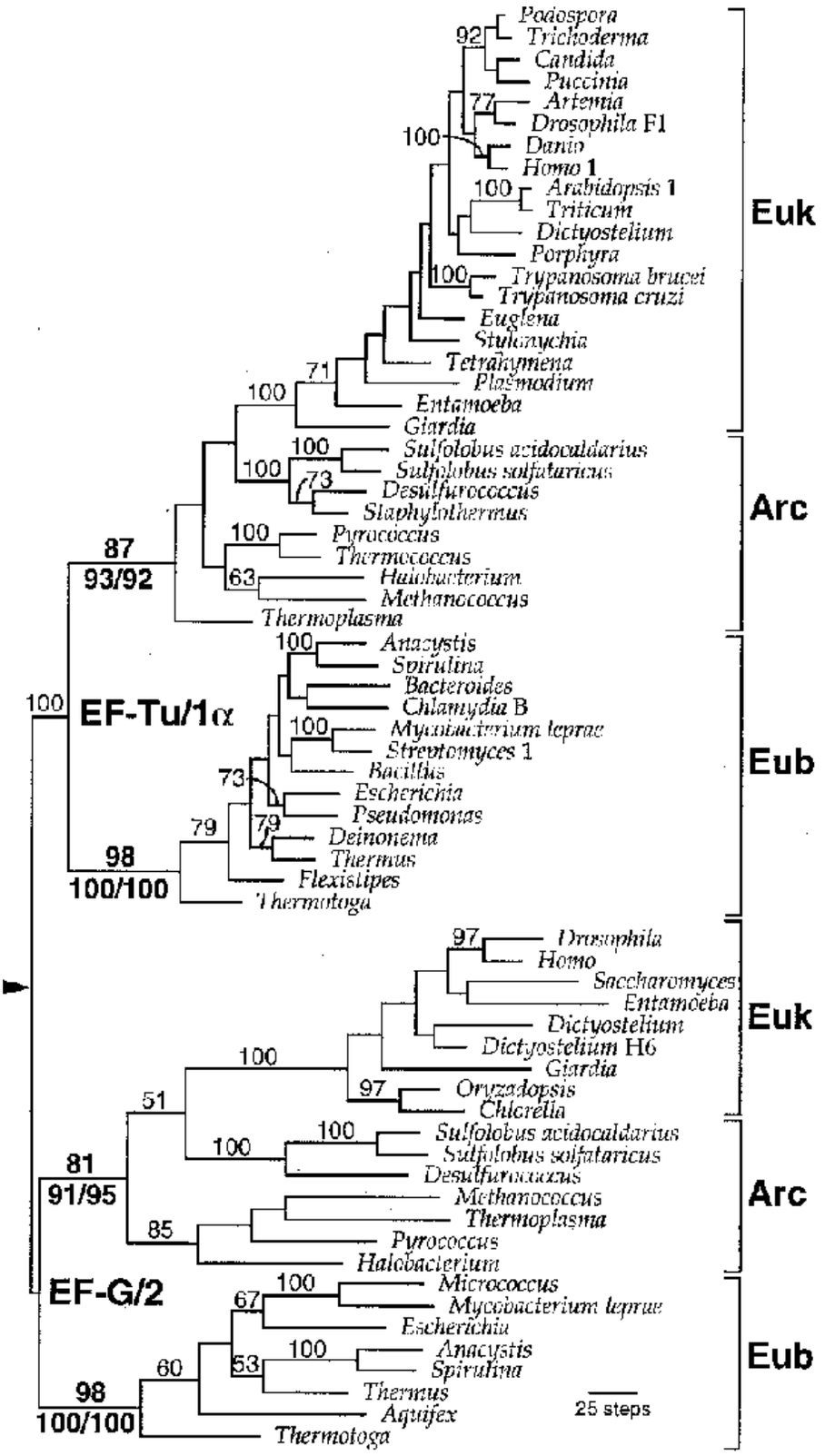


Figure 4: From Baldauf et al. 1996[22], phylogeny of two paralogs of the elongation factor, EF TU/1α (295 AAs) and EF G/2 (382 AAs), inferred by maximum parsimony. The low quality of the figure is due to its publication date, 1996.

4.2.2 Interlude – Rare genomic changes

Some authors prefer to use evolutionary events less common than the (possibly multiple) substitutions in the sequences to reconstruct phylogenetic trees. These **rare genomic changes** (**RGC**s) can be **insertion-deletion** (**indels**) of **introns**, integrations of **retroposons**, signature sequences (regions with a specific change for all members in a subset of taxa but absent outside of these taxa), alterations to the order of the genes on the chromosome (**synteny**), **duplication** of genes and variations in the genetic code coding for the proteins[23].

This type of analysis was studied by R. Gupta[24] with the indels of characteristic residuals in one or more phyla. An indel of 21-23 amino acids (AAs) (the characteristic residuals aforementioned) in the sequence of the Hsp70 protein allowed him to differentiate the **diderm-LPS** from the **atypical diderms**. Indeed, using MreB, a paralog of Hsp70, stemming from an ancestral duplication, as an outgroup to root the tree, he observed that the former lacks the insertion, as does the Hsp70 sequence of monoderms. The insertion observed in diderm-LPS Hsp70 is thus posterior to the duplication and suggests that diderm-LPS organisms are a branch emerging from the monoderms [24]. The following figure summarizes Gupta's scenario for bacterial evolution (Figure 5).

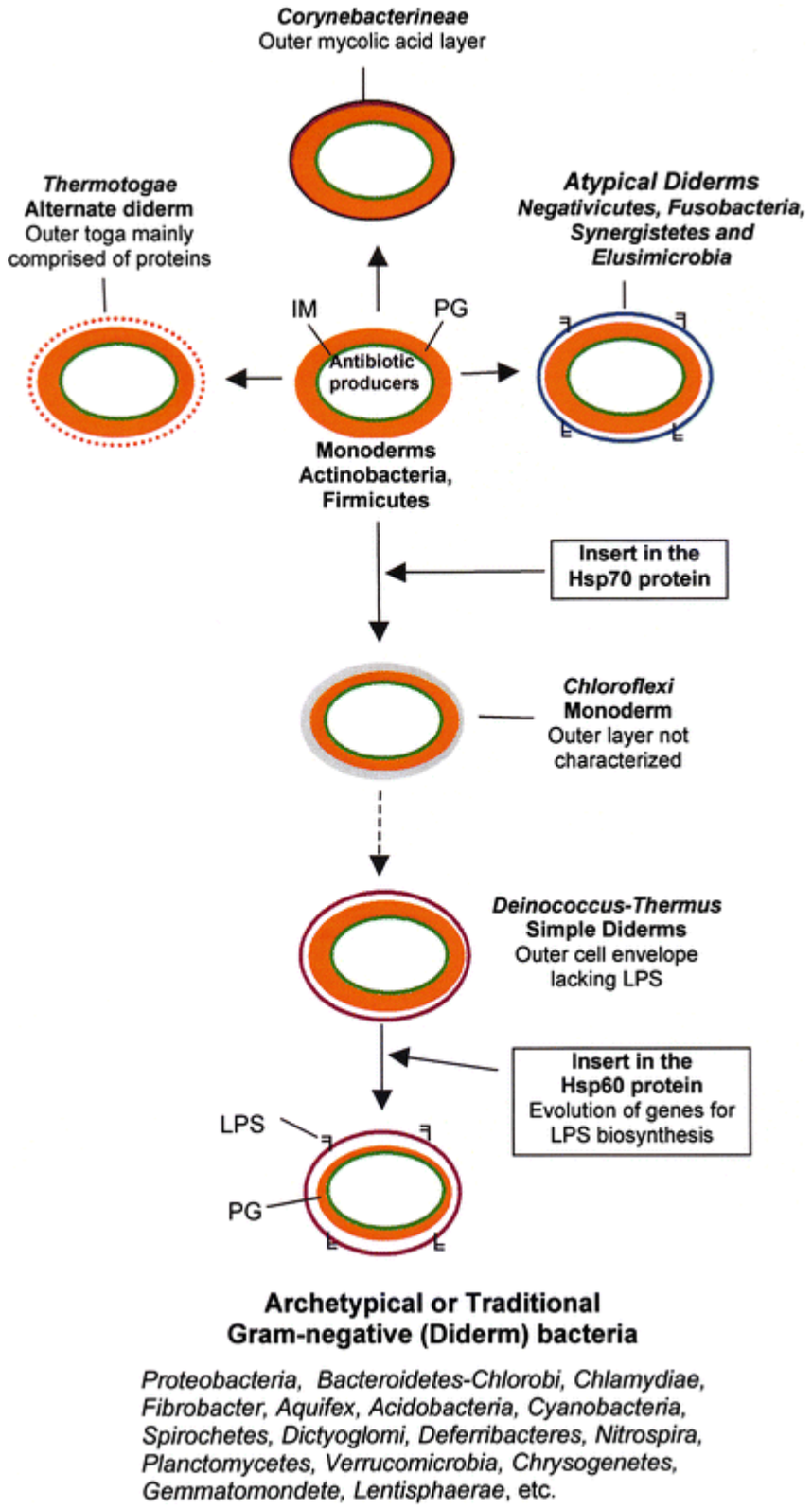


Figure 5 (from Gupta 2011[24] Figure 2): a proposition of a scenario concerning the development of outer cell envelopes in various bacterial lineages in response to antibiotic selection pressure[25]. Information regarding species distribution of Hsp70 (Heat Shock Proteins) inserts for most bacterial phyla is provided in the following works[26–28]. G+: Corynebacterineae, Actinobacteria, Firmicutes, Deinococcus-Thermus; G-: Thermotogae, Negativicutes, Fusobacteria, Synergistetes, Elusimicrobia, Chloroflexi. Abbreviations: PG peptidoglycan, IM inner membrane, LPS lipopolysaccharides.

Another insertion in Hsp60 distinguishes the diderm-LPS from the Chloroflexi and Deinococcus-Thermus and closes the gap between the other monoderms and Archaea. This argument was used by Gupta against the division of Life into three domains proposed by Woese[2]. This further led Gupta to emit a theory on the origin of diderms where they arise from the monoderms (without the Hsp70 insertion) passing by the Chloroflexi (monoderm with the Hsp70 insertion) and then the Deinococcus-Thermus (diderm without LPS with the Hsp70 insertion). The Hsp60 insertion separates further the diderm-LPS and the Chloroflexi and Deinococcus-Thermus, as shown in Figure 5 from Gupta (2011)[24]. One issue with such kinds of evolutionary scenarios is that it is only composed of extant organisms organized in a way such as some present-day groups appearing as the ancestors of other groups, in a *Scala Naturae*[29–31] way of thinking.

Cavalier-Smith[32,33] also bases his work on this type of analysis by adding the cell walls for what he calls the neomuran revolution. The Neomura clade as defined by Cavalier-Smith regroups the Archaea and the Eukarya and means “new walls” in reference to the differences in the cell wall of Bacteria on one hand and of Archaea and Eukarya on the other hand. This clade is supposed to emerge from Bacteria instead of being separate like in the “**three primary domains**” or “**two primary domains**” scenarios preferred for the moment[34,35]. An illustration of these scenarios is shown in Figure 6. Part of his work [32] is based on Gupta (1998)[2] for the vocabulary, which Cavalier-Smith reused and redefined, like monoderms and diderms. As a note, he also uses the ambiguous term “inner membrane” (IM), instead of plasma or **cytoplasmic membrane**, even though the latter terms are less misleading[36].

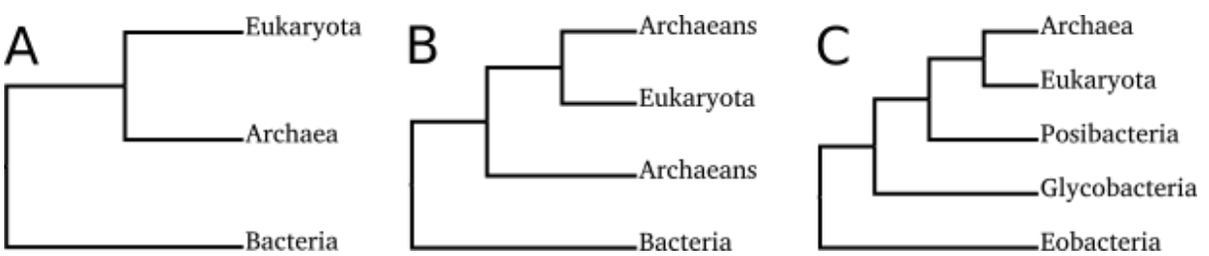


Figure 6: Three hypotheses for cellular organisms’ evolution. (A) Three-domain view (known as the Woese tree). (B) Two-domain view with the Eukaryota emerging within the Archaea. (C) Neomura hypothesis with the Archaea and Eukaryota (Neomura) emerging from within the Bacteria. Eobacteria and Glycobacteria are part of the Negibacteria (Gram negative bacteria), the Eobacteria correspond to the Chloroflexi and the Glycobacteria correspond to the Cyanobacteria, Spirochaeta, PVC (Planctomycetes, Verrucomicrobia, Chlamydia) and Proteobacteria. The Posibacteria belong to the Unibacteria (single membrane) and correspond to the Firmicutes and Actinobacteria. For more details on the bacterial groups Eobacteria, Glycobacteria and Posibacteria, see[32,33].

4.2.3 Phylogenomic supermatrices

To remedy the lack of phylogenetic signal of single markers, the **concatenation** of orthologous genes[37,38] is frequently used (Figure 7 f to g). This **phylogenomic** method, known as the **supermatrix** approach, works because the signal increases with the length of the concatenated sequences. Indeed, a longer sequence (more markers stitched together) has more chances to record a substitution per time unit. Yet, this does not solve everything[39–43].

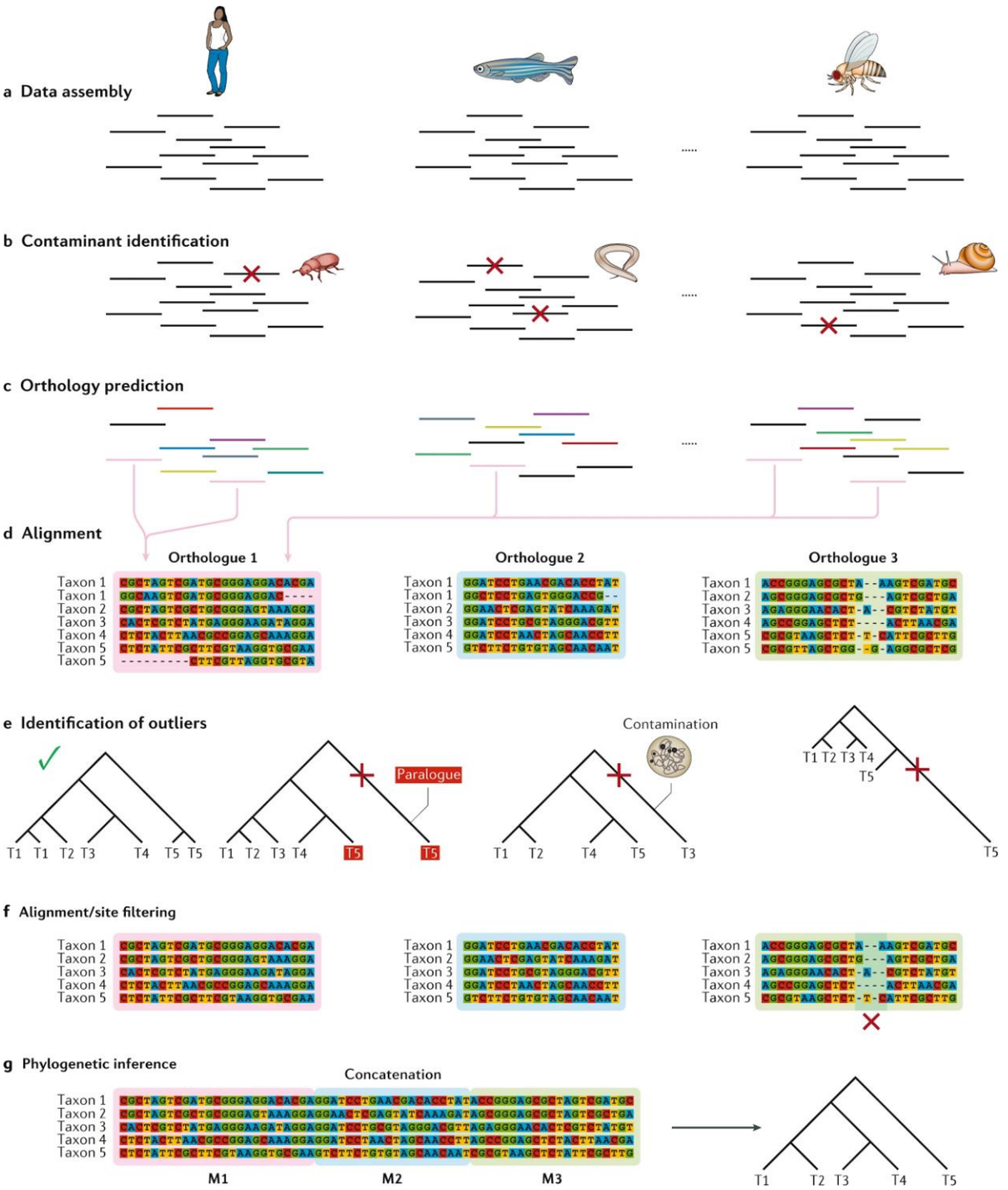


Figure 7 (from Kapli et al. 2020[44]): a | Gene sequences are selected. b | Contaminated sequences are removed. c | All-against-all comparisons are used to identify sequences that are homologous between all species of interest. d | The sequences of putative orthologues are aligned to generate a multiple sequence alignment (MSA). e | The MSA can be analysed to produce an initial phylogenetic tree for the putative orthologs, which can be used to identify remaining paralogues, contaminants and other problematic sequences indicated by unusually long branches. f | The MSA is typically filtered to remove regions of unreliable alignment. g | The orthologues are concatenated to produce a supermatrix, which is analysed to infer the species phylogeny.

The assembly of such a supermatrix requires the prior identification of every orthologous gene[43–46] for every organism studied (Figure 7 a to e). Orthologous genes are **homologous** genes but several types of homologies exist. Homologous genes are genes which share a **common ancestor**. Two genes can share a common ancestor by **speciation** or **duplication**. Speciation

refers to the process of differentiation of populations into different species. Genes originating from a single ancestral gene in the last common ancestor of the compared genomes are called **orthologs** or orthologous genes[37]. Genes are called **paralogs** if they are duplicated within a genome. Illustrations of these events are shown in Figure 8.

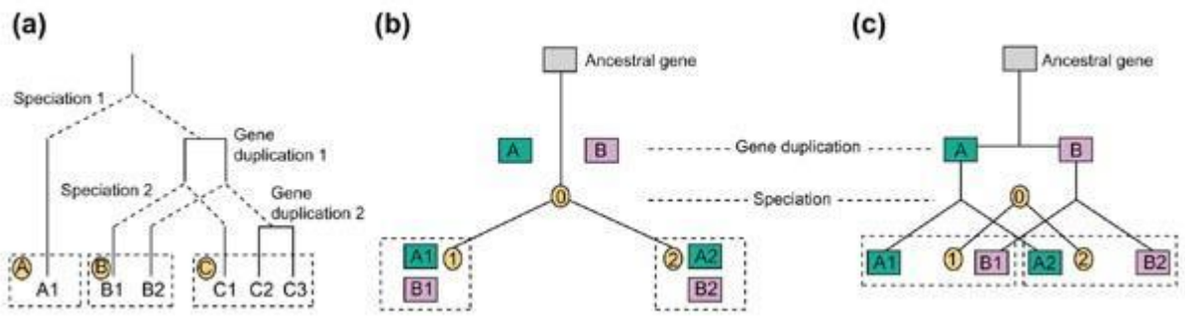


Figure 8 (from Jensen 2001[47]): (a) Simplified diagram of homology subtypes (showing orthologs like A1, B1 and C1 and paralogs like B1 and B2); adapted from[48]. (b,c) Evolutionary descent of an ancestral gene to paralogs and orthologs following gene duplication in species 0, and then speciation to yield species 1 and 2. Diagram (b) shows the resulting relationships between paralogs and orthologs, as illustrated by Koonin in his comment (2001)[49]. Diagram (c) Jensen's (2001)[47] version of Koonin's diagram using a Fitch diagram for visualization.

An example of this method is the article of Battistuzzi and Hedges (2009)[50], in which they selected 25 protein-coding genes in Bacteria, Archaea and Eukaryota[51] for 283 species and went to great lengths to improve the selection of species and the sequence **alignments**. The selection goes from 283 to 218 species and the supermatrix goes from 18,586 AA positions to 6884 AA positions. They used **Maximum Likelihood** (**ML**) and **Bayesian Inference** (**BI**) [52] for the phylogenomic trees and revealed the possible existence of two mega-groupings: the Terrabacteria (Cyanobacteria, Chloroflexi, Firmicutes, Mollicutes, Actinobacteria, Deinococcus-Thermus) and the Hydrobacteria (Proteobacteria, Bacteroidetes, Chlorobi, Chlamydiae, Planctomycetes, Spirochaetes), the ancestors of which may have been terrestrial and aquatic, respectively. These inferences are based on the environmental origin of the extant species and evidence of adaptations to desiccation and other typical stresses characteristic of terrestrial habitats for the species belonging to the Terrabacteria group[50]. A summary of the technical details is available in Table 1. The simplified topology can be found in Figure 9 and the detailed topology in Figure 10.

Supermatrices of hundreds of genes have been successfully used over the last 20 years in order to resolve the phylogenetic relationships in various regions of the Tree of Life, e.g., animals[53–56], green plants[57,58], fungi[59] and among eukaryotic lineages in general[60,61]. Here, several studies using supermatrices with Bacteria have been selected for review. The corresponding methods are summarized in Table 1 and the obtained topologies are compared in Figure 9. These trees are not all congruent and the sources of this incongruence will be discussed just below, whereas the biological conclusions will be addressed in the section after that.

study	protein type	# proteins	# species	# AAs	methods and models	panel	note
Boussau 2008 [62]	core proteins	56	94	?	ML JTT	A	
Battistuzzi 2009[50]	core proteins	25	218	6884	ML JTT+G // BI JTT+G	B	comparison with a rRNA tree; several steps for improving protein alignments (283->218; 18586->6884)
Yutin 2012[63]	ribosomal proteins	50	995	6127	ML WAG+G	C	
Rinke 2013[64]	core proteins	38	2460	~4000	ML JTT+CAT / LG+G / JTT+G	D	
Lasek-Nesselquist 2013[65]	ribosomal proteins	85	109 / 146 / 118	13,432 / 13,929 / 17,876	BI CAT // ML LGF / LG4M / LG4X / UL3	E	
Raymann 2015[66]	core proteins	46	134	10,986	BI CAT+GTR+G	F	2 others datasets used the with same models
Hug 2016[35]	ribosomal proteins	16	3083	2596	ML LG+G	G	information for BS at 85% instead of 90%
Parks 2018[67]	core proteins	120	21,943	34,744	ML WAG+G / LG(+G?)	H	comparison between three different programs, the “models” column refers only to the fasttree program; BS not available
Castelle 2018[68]	ribosomal proteins	14	3356	?	ML LG+CAT	I	number of positions not communicated; BS not shown
Zhu 2019[69]	marker genes	381	10,575	~38,000	ML LG+CAT / LG+G // ASTRAL LG+G	J	BS not shown in the main tree
TCS 2020[70]	ribosomal proteins	26	151	?	BI CAT+GTR+G	K	

this study	core proteins	117	85	19,959	BI CAT+GTR+G	L	supp: Bayesian CAT+G and ML LG+G

Table 1: Comparison of 12 phylogenomic studies of bacterial evolution based on supermatrices. AA = Amino Acid, ML = Maximum Likelihood, JTT = Jones-Taylor-Thornton[71] , G = Gamma, BI = Bayesian Inference, BS = BootStrap, WAG = Whelan And Goldman[72], CAT = CATegories[73], LG = Le-Gascuel[74], LGF = Le-Gascuel empirical frequencies, LG4M = Le-Gascuel 4-Matrices (one for each Gamma rate category), LG4X = Le-Gascuel 4 Matrices with distribution free rates, UL3[75] = Unsupervised Learning of variants EX3, EX3[76] = exposed/intermediate/buried sites, GTR[73] = Generalised time-reversible, ASTRAL[77] = Accurate Species TRee ALgorithm (which is akin to supertrees and not a supermatrix approach).

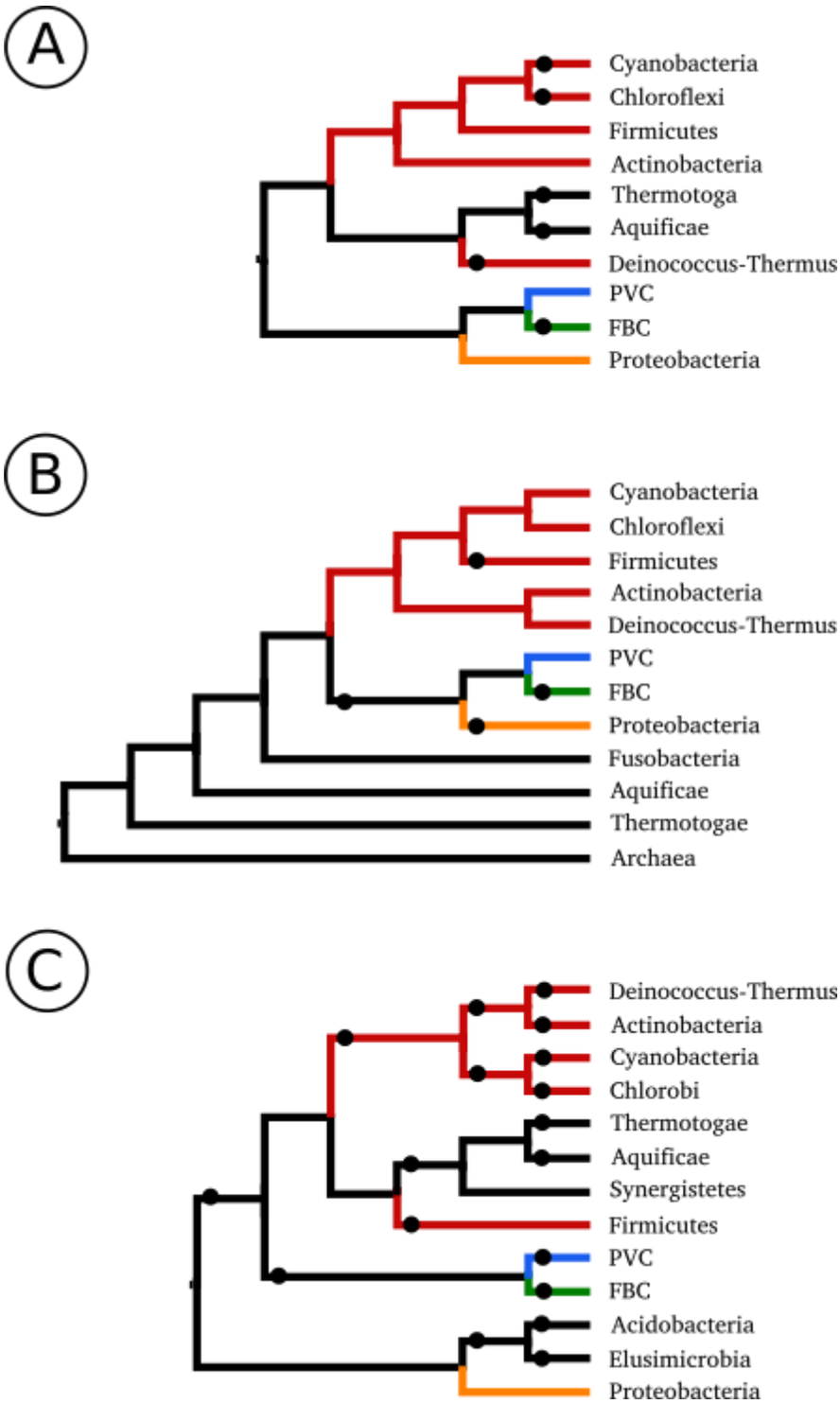


Figure 9 part 1: Illustrations of the topologies from the trees described in Table 1. The link between the article and the topology is also given in Table 1. For most trees, we were able to retrieve the values for the bootstrap proportions (BP) or posterior probabilities (PP). In these cases, a black dot represents a value of 85%/0.85 or more for the node and in the absence of a black dot a value below 85%/0.85. These topologies are not the complete topologies but simplified topologies showing only up to the phylum taxonomic level. The trees with a "*" at the root are the trees where the BP/PP are not available. FBC = Fibrobacteres, Bacteroidetes, Chlorobi, CPR = Candidate Phyla Radiation (= Patescibacteria). Red = Terrabacteria; Blue = PVC; Green = FBC; Orange = Proteobacteria and assimilated.

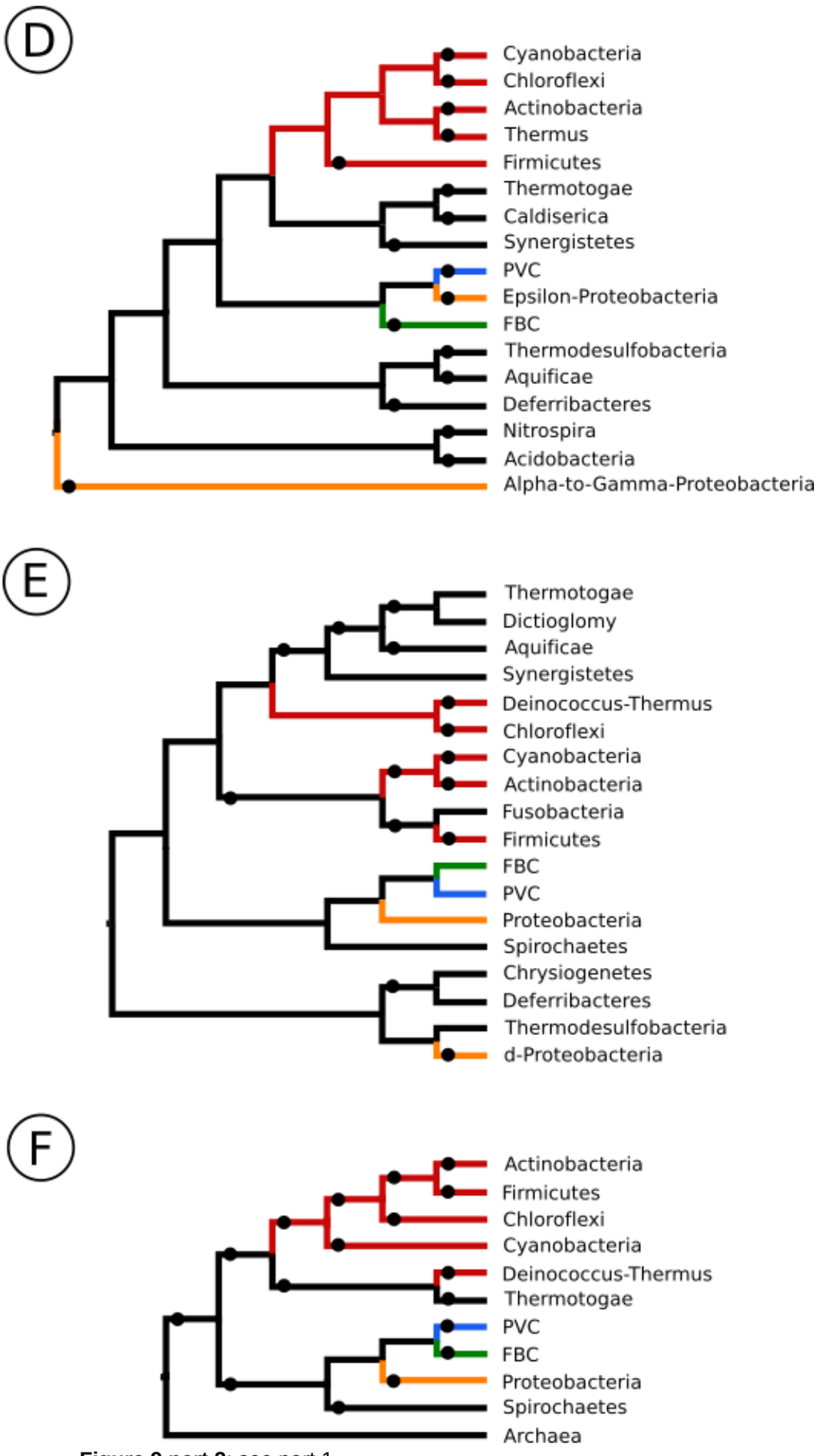


Figure 9 part 2: see part 1.

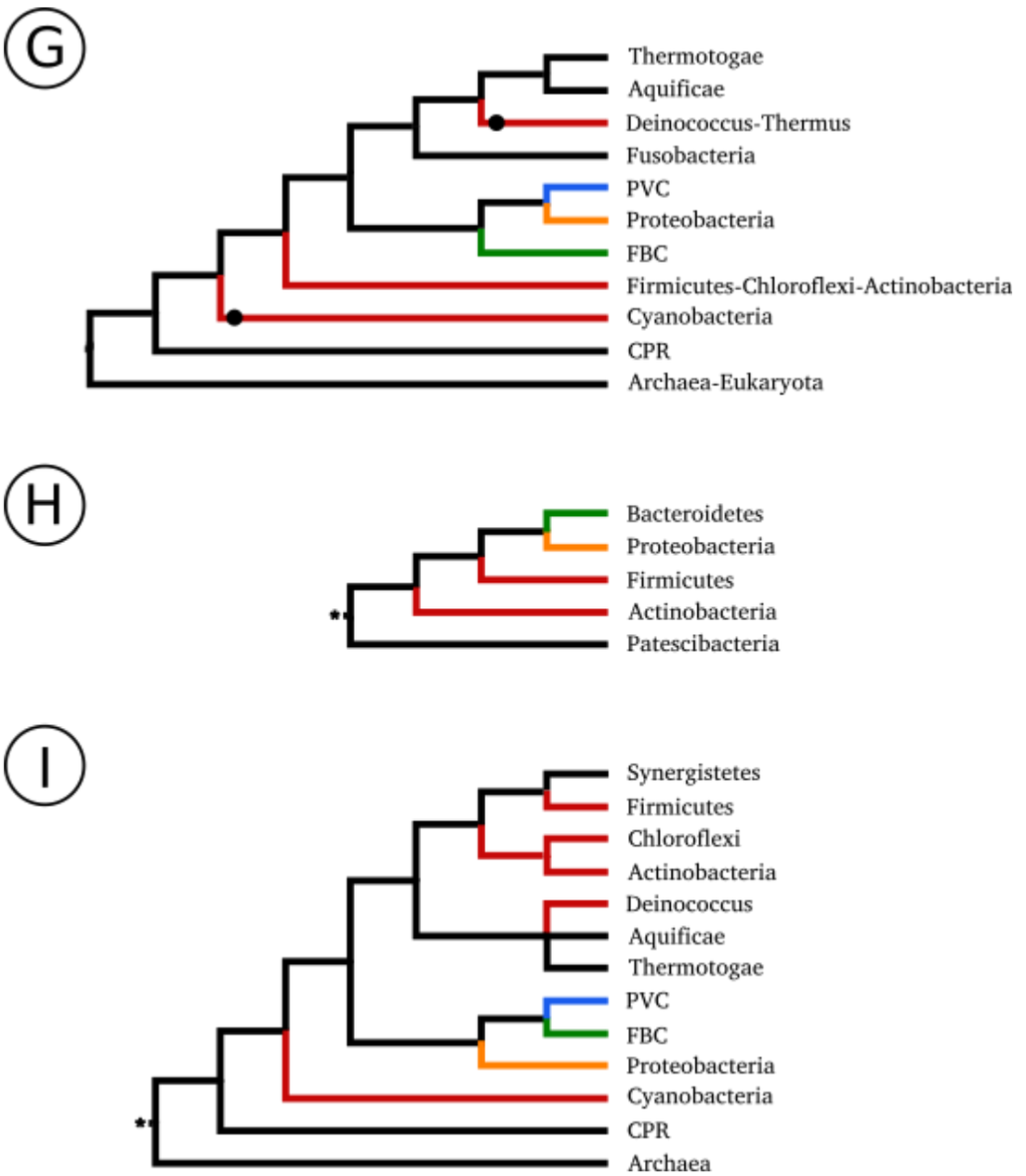


Figure 9 part 3: see part 1. In (H), Patescibacteria corresponds to the CPR[64,78].

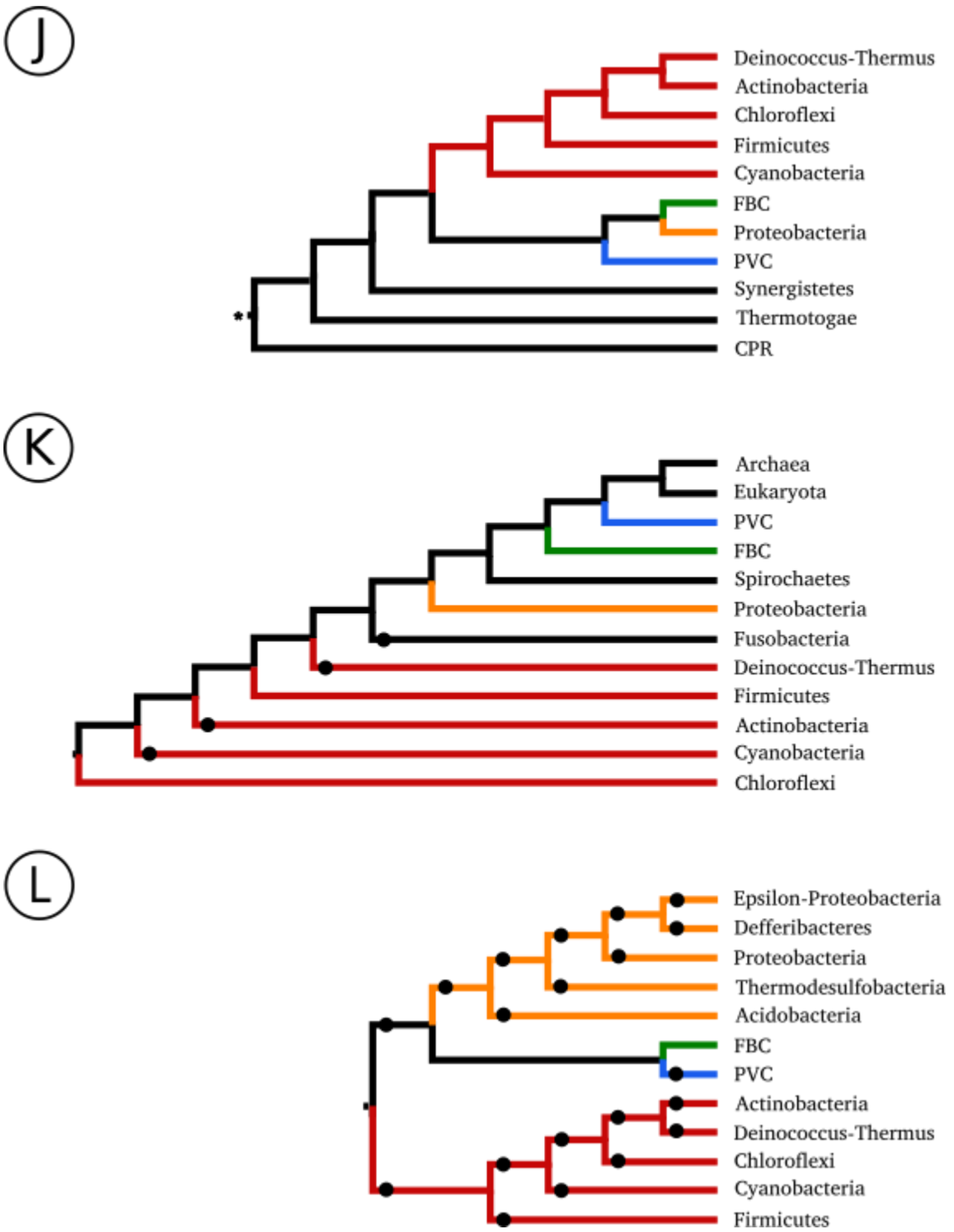


Figure 9 part 4: see part 1.

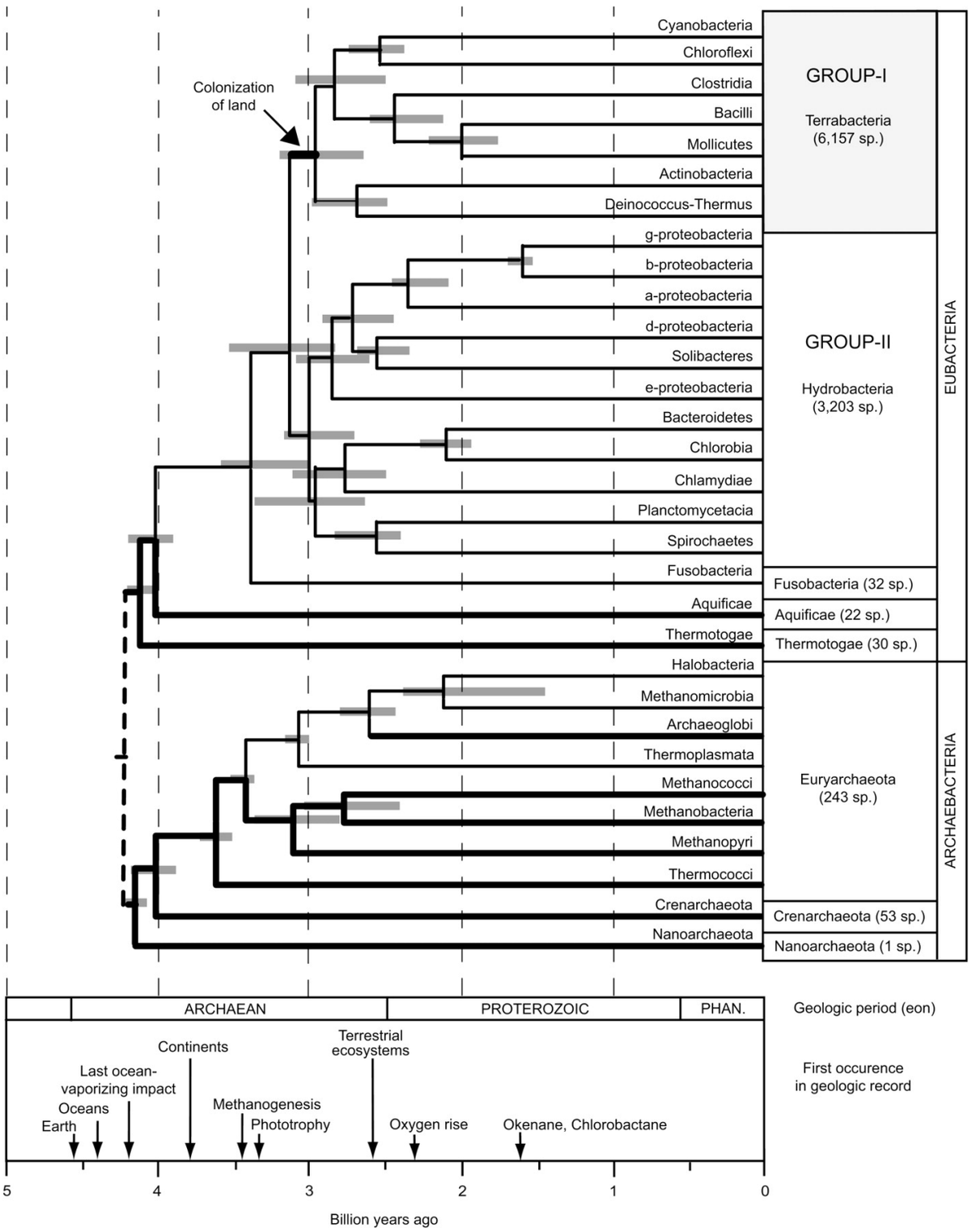


Figure 10 (from Battistuzzi and Hedges 2009[50]): Timescale of prokaryote evolutionary history. The timescale is in billion years ago. Each horizontal line represents a class; exceptions are the phylum Bacteroidetes (which includes two classes), Cyanobacteria, and Nanoarchaeota. Thicker lines are lineages that include hyperthermophilic species. Gray bars show the range of time estimates for each node, from each of the four estimation methods. The estimation was done using 21 bacterial species and 10 archaeal species with the different calibration points given in the figure. The incertitude is due to the use of several methods of which none can be excluded.

4.2.4 Limitations of phylogenomics

4.2.4.1 Imperfect evolutionary methods and models

Four main classes of methods can be used to infer a phylogeny from an alignment: the distance-matrix method, the maximum parsimony method, the ML and the BI. With distance-matrix methods, the alignment is converted into a matrix of genetic distances, whereas the maximum parsimony method tries to find a tree that explains the alignment with the least possible substitutions. ML tries to maximize the probability to observe the data, considering a model with a specific set of parameters. These parameters include interesting parameters, i.e., a tree composed of a topology and a set of branch lengths, and so-called “nuisance” parameters, to specify the model of sequence evolution, itself composed at least of compositional vector(s) and substitution matrix(ces). BI tries to maximize the probability of a hypothesis from known data, based on probabilistic models very similar to those of ML, but formalized in a Bayesian framework, hence the name Bayesian Inference.

In theory, except for clustering algorithms such as NJ suitable for distance matrices, the four methods should explore the entire realm of possible topologies, branch lengths (distance-matrix, ML and BI) and model parameters (ML and BI). However, since this is not possible outside of extremely simple cases, they rather have to rely on heuristics. These heuristics, instead of exploring the entire realm of possibilities, search for an optimum solution (local), which is not necessarily the optimal solution (global). Usually, they begin with a random starting position in the topological and parameter space then try to optimize (or sample for BI) the topology and other parameters (see Figure 11)[79].

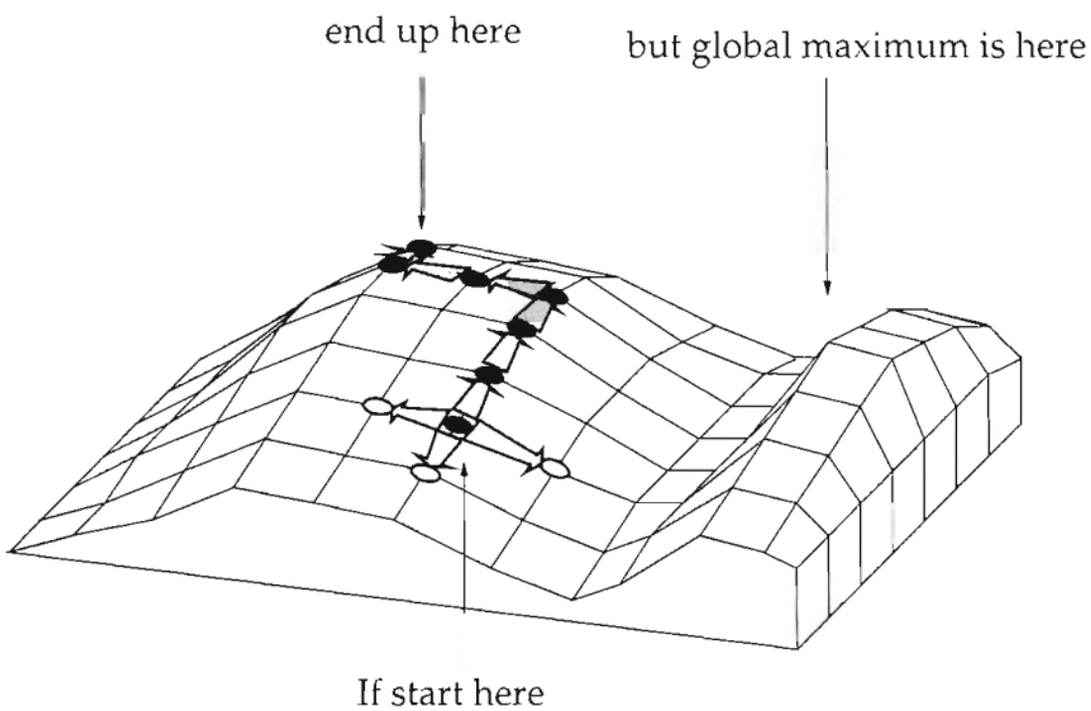


Figure 11 (from Felsenstein 2004[79]): illustration of a heuristic method exploring the realm of possibilities and setting to a local optimum instead of the global optimum.

ML and BI try to find the tree having led to the analyzed molecular sequence by a stochastic Markov process of substitution. The substitution process models the probability for a residue in a given sequence (nucleotide or amino acid) to be replaced by another residue without consideration for potential previous substitutions. In the simplest model, the substitution process is homogeneous, stationary and reversible, which is convenient from a modeling perspective but yields unrooted trees[80,81].

Models can be empiric (fixed and determined beforehand on training datasets) or parametric (determined during the inference itself). Usually, empirical models are used on amino-acid sequences due to the higher number of possible states. Such models, like WAG[72] or LG[74], are specified in Table 1. Yet, it is possible to use a parametric model like GTR on amino-acid sequences as well, when datasets are large enough (i.e., on phylogenomic supermatrices).

As mentioned above, the simplest models consider a homogeneous process for the substitutions but, in reality, the substitution process is heterogeneous, which can lead to model violations, phylogenetic artefacts and thus incongruence between trees[41]. Several types of heterogeneity exist, such as the substitution rate across sites (it changes for each residue of a sequence, leading to conserved residues or divergent residues) modelled by the Gamma distribution[82] or the substitution rate over time, known as heterotachy[83,84]. Alternatively, the heterogeneity of the substitution rate across sites can be built-in in the empirical substitution matrices, like in LG4X and LG4M variants of LG[74]. In contrast, the CAT model focuses on the heterogeneity of the amino-acid profiles (sets of amino acids actually used) across sites while assuming equal exchange probabilities between amino acids[85]. The CAT model can be complexified by adding a GTR component, leading to the CAT-GTR model. The best model for a given supermatrix can be determined by different statistical procedures. In the case of BI, cross-validation is an appropriate way of selecting the best fitting model[86]. Other types of heterogeneity exist but no model has implemented all of them for now, due to the complexity of the task from a computational point of view. The design of new models of sequence evolution is an active field and, for a recent review of amino-acid evolutionary models, see Pupko & Mayrose 2020[87].

4.2.4.2 Genomes and metagenomes

Nowadays, prokaryotic genomic data suffer from several problems. The number of **metagenomes** is exploding, leading us to struggle with the untangling and assembly of the constituting genomes. A metagenome is a collection of genomes present in a sample and sequenced together because we cannot, for the moment, separate them in **pure cultures**. For example, in 2017, Parks et al.[78] published 8000 new genomes coming from 1500 metagenomic samples (see Table 1 and Figure 9). The **MAG** (Metagenome-Assembled Genome) of a single isolated genome can be a challenge in itself if there is no "**scaffold genome**" to generate the assembly of the different contigs.

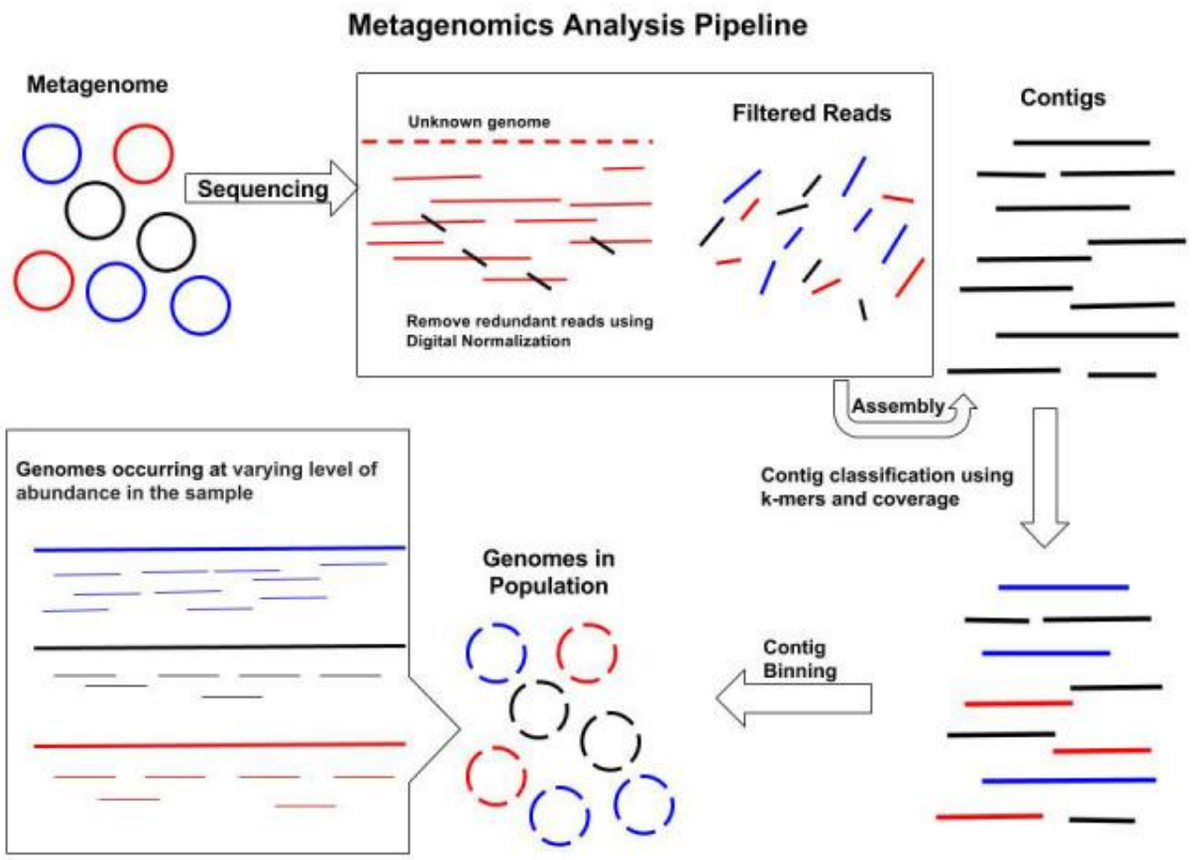


Figure 12 (from Ghurye et al. 2016[88]): Multiple bacterial genomes within a community are represented as circles of different colors (same color = same organism). After sequencing redundant reads can be removed through digital normalization, reducing the computational needs for assembly. The filtered reads are then assembled into contigs and they are classified using *k*-mers and coverage statistics. Contigs in each group are then binned to form draft genome sequences for organisms within the population. Note the different levels of sequencing coverage for individual organisms' genomes, due to the different abundance of the organisms in the original sample.

In MAGs, several non-isolated genomes (Figure 12), usually without a “scaffold genome”, are thrown in together to make some sort of mega-puzzle, multiplying the challenge and leading to more potential errors. On a side note, the description of the organisms to which these genomes belong is also problematic, due to the absence of pure cultures for most organisms composing a metagenome. This prevents researchers from describing them, since one most often needs a pure culture to describe a new organism.

4.2.4.3 Genome contamination

Even among non-metagenomic assemblies, most new genomes are barely assembled because this remains a slow and difficult process (due to the repeated regions), whereas the sequencing itself continuously gets cheaper and faster, especially with the replacement of Sanger sequencing by next-generation sequencing (e.g., Illumina)[42,89]. There is also the issue of genome **contamination**, i.e., the presence of foreign sequences in a given genome (see Figure 7b).

Contaminating sequences arise either from the **inaccurate partition** (**binning**) of sequence data of a metagenome corresponding to multiple organisms (see Figure 12) or from the assembly of a (supposedly) single organism grown as a non-**axenic** culture or due to the sequencing technology (e.g., leakage between multiple lanes or inaccurate demultiplexing)[90,91]. That is why comparative genomics and phylogenomics require the identification and removal of such contaminated genomes in favour of better ones (or at least the flagging of problematic genomes if one cannot afford discarding them)[44].

4.2.4.4 Horizontal gene transfer

The phenomenon of **horizontal gene transfer** (**HGT**), common among prokaryotes, complicates the interpretation of phylogenomic results. While HGT events can be mistaken for contaminations, they are completely different. A contamination happens during the culture of the organism, the preparation of the sample or the sequencing of a genome, but an HGT event is a foreign sequence (sometimes termed **xenolog**) which has become a genuine part of the genome before all these experimental steps.

This is possible because a bacterium is capable of acquiring foreign genetic material by three different means: (1) by t**ransformation** (Figure 13a) where a bacterium imports genetic material from the outside medium, (2) by **transduction** (Figure 13b), which implies a **phage** to transmit **DNA** and (3) by **conjugation** (Figure 13c) where a **plasmid** is transferred from a bacterium to another[92]. These exchanges can for example transfer **resistance** genes (to antibiotics or toxins) between bacteria, leading to the acute problem of multi-resistant **pathogens**[93,94].

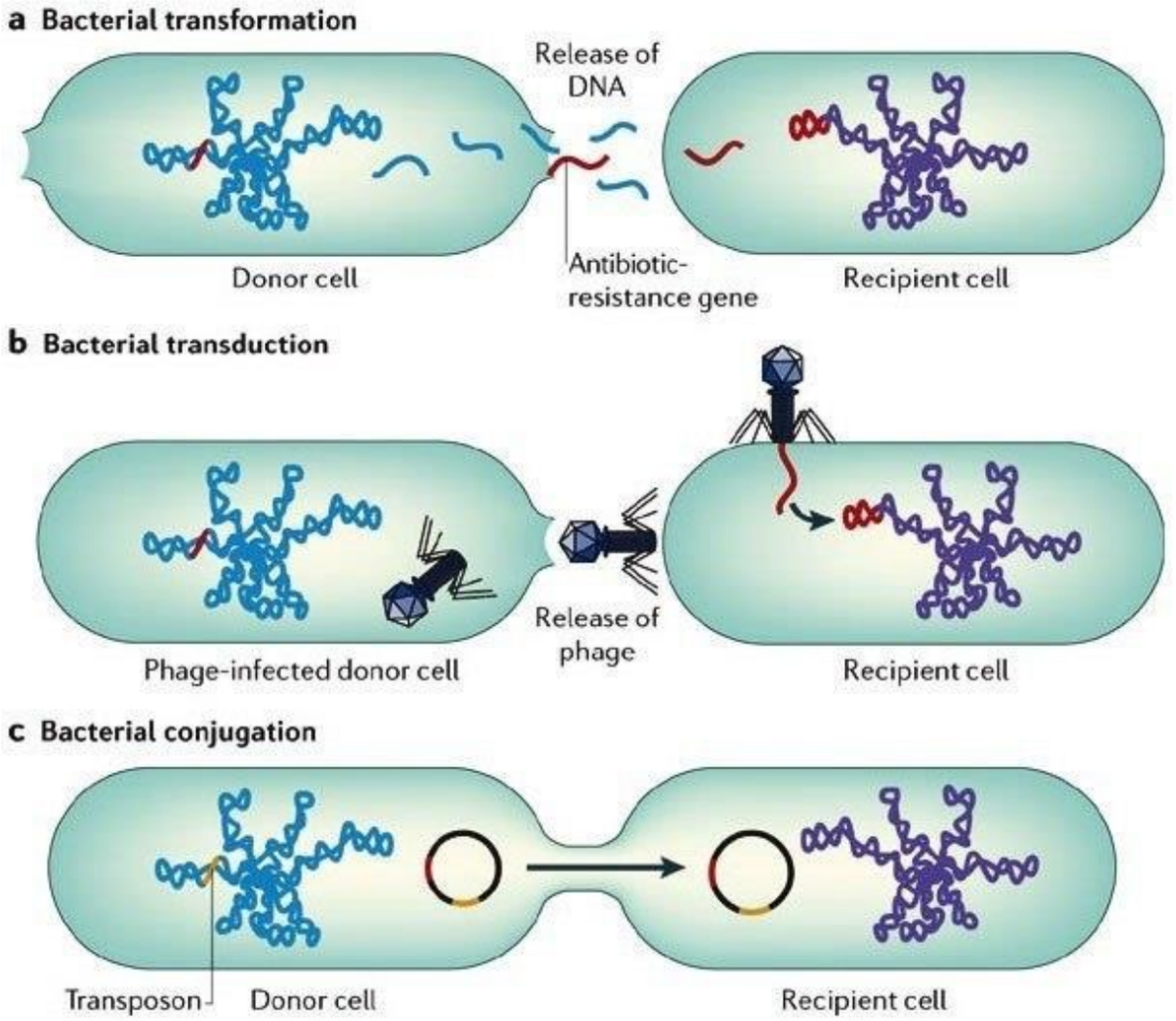


Figure 13 (from Furuya & Lowy, 2006[92]): Horizontal gene transfer between bacteria. a | Transformation occurs when naked DNA is released on lysis of an organism and is taken up by another organism. b | In transduction, genes are transferred by means of bacteriophages and can be integrated into the chromosome of the recipient cell (lysogeny). c | Conjugation occurs by direct contact between two bacteria.

4.2.5 Core proteins and case studies

HGT events can be spotted by analyzing the **GC content** along the genomes. Indeed, as the GC content is a global feature of a genome, the genes **recently** transferred present a different GC content from the rest of the genome[95]. Beware that a genome GC content is not always homogeneous, varying from part to part[96]. Thus, this reasoning is valid when comparing the immediate vicinity of a specific sequence. We can also detect HGT events due to unexpected sequence similarity to distant species instead of closely related species, incongruent gene and species trees or even anomalous genetic distribution (e.g., a gene is present in one species but is completely absent in related species)[97].

Comparisons between close species show that a part of the genome of bacteria is conserved and transmitted vertically, the **genomic core**, and another, limited to specific regions of the genome, is much more variable and made of recent HGTs[45]. The effect of HGT may be partially overcome

by using multiple aligned genes/proteins for the phylogenetic analyses. By adding more core (rarely transferred) genes, we can improve the phylogenetic signal, thus flooding the discordant signal sent by HGTs. However, it does not remove the discordant signal and, if the objective is to work with orthologous genes, removing the genes identified (or supposed to be) the result of HGT events is better[42].

4.2.5.1 The good...

The genes encoding ribosomal proteins (composing the ribosomes along with rRNAs) are amongst the less exchanged horizontally and by consequence well suited for phylogenomic studies[98]. This is why the article of Yutin and al. (2012)[63], based on the concatenation of 50 ribosomal proteins (6127 AAs) for 995 bacteria, is of interest (see Table 1 and Figure 9C). This study also suggested the existence of bacterial mega-groupings, but three instead of two, and different from those of Battistuzzi and Hedges (2009)[50] (Figure 9B): group I (Spirochaetes, Planctomyces, Chlamydiae, Verrucomicrobia, Chlorobi, Bacteroidetes, Fibrobacteres), group II (Deinococcus-Thermus, Actinobacteria, Chloroflexi, Cyanobacteria) and group III (Fusobacteria, Mollicutes, Firmicutes).

4.2.5.2 …the bad and the ugly!

Another article, by Hug et al. (2016)[35] (Figure 9G), also used ribosomal proteins to produce a phylogenomic tree of Archaea, Bacteria and Eukaryota. 16 ribosomal protein sequences were concatenated to produce a supermatrix of 2596 AA positions for 3083 genomes. Two main problems exist in this otherwise high-impact paper. First, the authors used a small supermatrix, while they study ancient events, which require far more positions to be reconstructed. In comparison, in another recent article, Parks et al. (2018)[67] (Figure 9H) used 120 proteins (34,744 AA) from 21,943 genomes, which is much more powerful in terms of phylogenetic signal. The second problem is their use of a single inference method based on a simple evolutionary model. To continue the comparison with Parks et al. (2018)[67], the latter computed their trees using three different programs and models[99–101]. As shown in the comparative Table 1, in all recent articles using phylogenomics, most used at least two different models and/or programs to produce and compare their results and also longer alignments (in terms of positions). Even if this does not solve everything nor guarantees correct results (“true tree”), it is at least a minimum for good practice in phylogeny. The topology of Hug’s tree is shown in Figure 9G.

The study from Castelle et al. (2018)[68] (Figure 9I) is another high-profile manuscript but it falls in the same pitfalls as the work from Hug et al. (2016)[35]. It uses 14 of their 16 ribosomal proteins and adds more species belonging to the **DPANN** archaeal superphylum (3356 species in total). The number of AA positions is not communicated. Moreover, they used the same simple model to compute their tree. Trying two or more models on the same datasets like in Rinke et al. (2013)[64] (Figure 9D) or trying the same model on variants of the dataset like in Raymann et al. (2015)[66] (Figure 9F) is easy to do (albeit time consuming) and allows researchers to check the **robustness** of their results. If every model tested on a dataset or if every dataset analysed with a model (or both) consistently regroup the same species together, reliable conclusions can be drawn. On the contrary, if a taxon keeps changing its position in the tree, depending on the model used and/or

the dataset, then it argues for uncertainty. All mentioned examples and details about the model(s) used can be found in Table 1 and Figure 9.

4.2.5.3 What to learn from it ?

From the twelve supermatrix-based studies described in Table 1 and Figure 9, we can learn a few things. Outside of the message just above about the use of multiple models and/or datasets as a good practice to check the robustness of the results, the topologies are the main information to remember. When comparing those topologies, without discriminating the methods used in each study, there are a few robust groups and some more fragile assemblages. The clearly robust groups are the FBC (in green in Figure 9) and the PVC (blue) superphyla, which are recovered in all trees in which they are represented. Concerning the less robust groups, let us cite the Terrabacteria (red), which are well defined most of the time, and the Thermotoga, Aquificae and Synergistetes phyla, which do not seem to prefer a particular position with respect to the backbone of the bacterial tree and are thus the phyla with the least robust positions across the different topologies of Figure 9.

If we add to the topologies the information about the application (or not) of the good practices mentioned above, the message changes slightly. In Figure 9BDJL, in which the corresponding studies can be considered to follow the good practices, the Terrabacteria group (red) is found as monophyletic whereas, in the other studies, the Terrabacteria group includes at least an additional phylum within its subtree (Figure 9ACEF), thereby transforming it into a paraphyletic group, or is exploded across the bacterial tree (polyphyletic group). Amongst the topologies with a polyphyletic Terrabacteria group, the Figure 9GIK are the worst offenders. They correspond to Hug et al. 2016[35], Castelle et al. 2018[68] and TCS 2020[70] and, as mentioned above for Figure 9G and Figure 9I, they are the “bad” and the “ugly” of our selection of phylogenomic trees, since they do not respect the good practices, i.e., they do not use enough protein sequences for their respective supermatrices and/or rely on a too simple model. The tree in Figure 9K was inferred using a good model but did not use enough protein sequences for its supermatrix. Indeed, working well is good but it does not compensate for a lack of information. For the study of Parks et al. 2018[67] in Figure 9H, despite using the good practices and enough protein sequences, the topology is as different as the worst offenders compared to the eight other topologies. In that case, it is probably due to the low number of phyla belonging to the Terrabacteria present in this study and the simple models used.

Concerning the Thermotoga, Aquificae and Synergistetes, two trends seem to exist: either an early divergence as in Figure 9BJ or belonging to the Terrabacteria, as in Figure 9ACEF. The status of chimera of the Thermotoga and Aquificae[102–106] or the possible link of the Synergistetes with the Firmicutes[107], in conjunction with the use of relatively simple models in most of these studies, might explain the difficulty to position these phyla. These three phyla are absent from the Figure 9HKL and, in the case of the studies for the Figure 9GI, they correspond to the “bad” and the “ugly” guys (i.e., “rogue taxa”[108]) from the previous section.

Considering these two pieces of information, the Terrabacteria group is likely to be a monophyletic or paraphyletic group with additional phyla nested within. If the position of these additional phyla (Thermotoga, Aquificae, Synergistetes) is indeed within the Terrabacteria, it asks the question if these taxa should be included in the Terrabacteria taxon to keep the group monophyletic or not.

4.2.6 Alternatives to supermatrices

4.2.6.1 Supertrees

As the supermatrix approach has its limitations, it is useful to resort to other methods (which also have their own advantages and shortcomings) and compare their outcome with the supermatrix trees, according to the corroboration principle. One such contender is the **supertree** approach, which consists in combining several single-gene phylogenetic trees into one single tree[109]. Each phylogenetic tree requires the use of orthologous genes like any other tree. In this instance, the better the **congruence** between the individual trees, the better the supertree will be supported. The congruence is the similarity between the topologies of different trees.

A classic method to produce supertrees is the **Matrix Representation by Parsimony** (**MRP**). It consists in converting the different phylogenetic trees in a single matrix showing the relationships between the genomes and then using an algorithm of maximum parsimony to reconstruct the supertree[110,111]. An example of a bacterial supertree built by MRP can be found in Tourasse & Kolstø 2007[112].

4.2.6.2 The MultiSpecies Coalescent model

Among these alternatives, the **MultiSpecies Coalescent** (**MSC**) model aims to improve the phylogeny when it comes to (possibly incongruent due to biological factors) **multilocus** sequence data[113,114]. The MSC could be summarized as the “upgrade” of the supertree approach[109]. The supertree uses single-gene trees to reconstruct a phylogenomic tree whereas the MSC involves the single-gene trees and the phylogenomic tree (the so-called **species tree**) to explain the history of multilocus sequences[113] (see Figure 14). These methods are recent and the subject of heated arguments between the ones who think it is an improvement compared to the concatenation of genes into a supermatrix[115] and the ones who think it is not an improvement[116]. While waiting for a definite and widely accepted conclusion on the question of the accuracy of the MSC, it might be better to either continue using the well-tested and understood supermatrix approach and trying to improve as much as we can our datasets and protocols[40–42] or identify which approach works best for the data at hand[117].

A recent example of the MSC approach can be found in Zhu et al. (2019)[69]. They used three different combinations of programs and models, and among them the ASTRAL program[77], which implements the MSC approach. You can find more details in Table 1 and Figure 9J, which belongs to a study that used both a supermatrix and a MSC supertree for corroboration of the results, a third way to overcome the limitations of supermatrices.

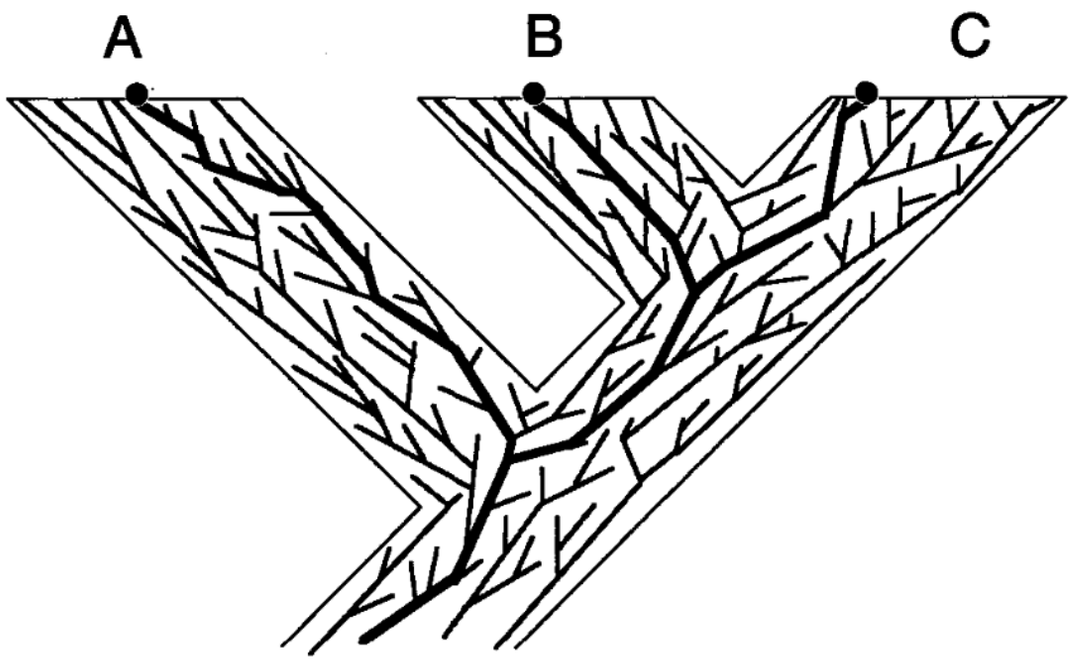


Figure 14 (from Maddison 1997[118]): A gene tree contained within a species tree leading to three extant species: A, B, and C. Bold branches of gene trees show relationships among the sampled copies of the gene (•). Sampled copies from sister species B and C are sister copies.

4.2.6.3 Reconciliation

A reconciliation model describes how the gene trees will evolve compared to a species tree. This evolution is composed of events such as the duplication, the transfers and losses of a gene since its origination[119]. The reconstruction of a species tree is a circular problem. As mentioned above, to produce a species tree using a supermatrix requires the use of orthologous genes and thus their identification (see Figure 7 a to e). To identify these orthologous genes, it requires gene trees and the identification of the events leading to these gene trees but these gene trees require themselves a species tree, hence the circularity.

A solution, proposed by Szöllősi et al. 2013[120], is to perform a joint inference of the gene trees and the species tree. They further produced a tool called **amalgamated likelihood estimation** (**ALE**)[120] for this purpose. ALE is capable of estimating the rates of the aforementioned events while taking into account the uncertainty of the gene tree topologies[119].

In Coleman et al. 2021[119], they used ALE to produce a reconciliation tree of the Bacteria, including the CPR. It is of note that in this case, the CPR is located within the Terrabacteria group instead of emerging at the base of the Bacteria like in our selected studies of Table 1 and Figure 9.

4.2.7 Alignment-free methods

Even if very common, alignment-based phylogenomics has several problems, such as the need to identify orthologous genes or the computationally heavy phylogenetic analysis of large supermatrices. These limitations can be alleviated by resorting to **alignment-free methods**.

Alignment-free methods are quantifiable ways of comparing the **similarity** of sequences without using an alignment[121]. They have several advantages over alignment-based methods: they are **less expensive computationally**, they are **resistant to shuffling** (of proteins domains or exons) and **recombination events** (meioses, transduction) and they do not depend on assumptions about the evolutionary path of sequence changes (i.e., no need to model substitutional histories)[122]. In the review of Zielezinski (2017)[121], two main categories are described, the first category includes **word-based** methods, while the second category encompasses **information theory-based** methods. But first, we will present a third method based on codon usage.

4.2.7.1 Codon aversion motifs

Several alternative approaches exist for producing phylogenetic trees without aligning the sequences, like **CAM** (**codon aversion motifs**)[123], one of the latest attempts. This method produces, supposedly, results faster than alignment-based methods and of similar accuracy than other alignment-free methods. A codon aversion motif is defined as the codons which are not part of an individual gene. For every species used in the phylogeny, the CAM is computed over all its genes. Then a pairwise distance is computed between every pair of species using their CAM for every gene. The distance is defined as one minus the proportion of shared codon aversion motifs between the species.

The idea behind this concept is that not every species uses every codon, notably due to the unequal **tRNA** expression, so the absent codons could be used to create a "**profile**" for the species usable for phylogeny[123]. According to their proponents, these alignment-free methods may be used not only for studying shallow phylogenetic issues but also to reconstruct deep phylogenies using whole genomes[124]. Yet, this remains to be demonstrated. In contrast, alignment-free methods are not restricted to generating phylogenomic trees and can be used in other applications, such as genome dereplication (see below).

4.2.7.2 Word-based methods

The idea behind word-based methods is that similar sequences share a similar set of words. The words are called ***k*-mers** and can be defined as all the words, of a given size, possible for a given alphabet. The idea is to compare the "**dictionaries**" of words between two genomes (see Table 2). If we compare a book and a nearly perfect copy of the same book, the dictionaries will be the same and thus will be considered redundant. That case corresponds to the comparison of strains from the same species with so few differences that they can be assimilated to typos in the book copy. If we compare the dictionaries of books about the same subject, say two high-fantasy novels with a usual plot, the dictionaries will be similar but with more differences as the settings likely differ (e.g., the Medieval Fantasy 1 & 2 in Table 2). That case would correspond to differences between genomes of different genera or families from a specific bacterial order. And if the subjects

are completely different, like a novel in a high-fantasy setting and a manual of macro-economy, the dictionaries will just have the basic language in common (Table 2), with the specific vocabularies completely different. That case would correspond to the difference between two bacterial phyla or even between a bacterium and an archaea.

Medieval Fantasy 1	Medieval Fantasy 2	Medieval	Macroeconomics
Elf	Elf	/	/
Orc	Orc	/	/
Sword	Sword	Sword	/
Halberd	Halberd	Halberd	/
/	Scimitar	/	/
Market	Market	Market	Market
Castle	Castle	Castle	/
/	/	/	Inflation
/	/	/	Keynesianism
is	is	is	is
have	have	have	have

Table 2: Examples of possible "dictionaries" content for four different books. The two medieval fantasy books share highly similar dictionaries whereas the medieval book does not reach the same level of similarity. The macroeconomics book is highly dissimilar compared to the other three books.

The most well-known *k*-mer is the **three-mer** also known as the **codon**. Indeed, the DNA is transcribed into RNA then RNA is translated into protein following a vocabulary of three-letter words. By studying a **longer frame**, we get a **less saturated** signal but we make the signal **more specific**. This is the reason we use the AA sequences instead of the nucleotide sequences in deep phylogenetics. The AAs give us more information by being an "alphabet of twenty characters" instead of an alphabet of four characters[121]. (By working with codons as character states, we would have a 64-character alphabet.) And since the cells work with codons, the information gained is more precise[125,126]. It also lowers the chance of matching by pure luck between sequences, it passes from 25% (nucleotides) to 5% (AAs) and 1.56% (codons). Due to the redundancy of the genetic code, the last nucleotide of a codon is rarely conserved and can still translate into the same AA. Plus, due to the similarity of the biochemical properties some AAs are sharing, they can be substituted more easily during the evolutionary process. These exchange rates can be captured into substitution matrices for AAs (the empirical model introduced above) whereas these cannot be computed for nucleotides.

If we apply this reasoning from alignment-based methods to word-based methods, longer *k*-mers should be better, shouldn't they? Yes, but only up to a certain point because if the words are too long, they will become too specific. By specific, we mean that for *k*-mers longer than a codon, the words quickly begin to be unique to specific taxa. Hence, if the *k*-mer size is too long, it will only differentiate between low taxonomic levels (species, genus) and be unusable for higher taxonomic

levels (phyla, orders), being too different to be clustered together. To differentiate isolates belonging to the same species, one can use a *k*-mer size between 20 and 25 nucleotides, for example[121].

4.2.7.3 Information theory-based methods

Information theory-based methods compute the amount of information shared between two analysed sequences, and we will discuss two of them briefly. The complexity of a sequence, as defined by Kolmogorov (1965)[127], can be measured by the length of its shortest description. This measure is commonly approximated with **compression algorithms**. The idea is to concatenate two sequences to be compared and then compress them. If the two are the same, their compressed size will be equal, or almost equal, to the size of a single compressed sequence. In contrast, the more different they are from each other, the more their compressed size will grow until they are so different that the compressed size of the concatenation is equal to the sum of the compressed size of the two sequences compressed separately.

You can compare it to the plan in each box of the famous construction game from a danish company or the plan in each box of DIY furniture from a famous swedish company. In both cases, the first part of the plan shows you every type of piece you have in the box and the number of each piece present in the box (akin to dictionaries discussed above). The second part is how to assemble the pieces until you reach your goal, the toy or the furniture, depending on which box you have. If the box contains two different sets of things to build, then depending on how close/similar they are, the size of the plan will differ. If the box contains two sets of the same object to build, then the number of pieces will be just multiplied by two in the first part and a line like the following, “Redo every step from the first to the last again”, will be present in the second part after the instructions. However, if the two sets are for completely different objects, then the first part will likely be longer due to the presence of different types of pieces and the second part will consist of two different sets of steps without common parts.

Another example is the Shannon entropy[128–130], where the idea is that some words are **common** and thus their presence is unsurprising but the presence of **rarer** words is meaningful. The **uncertainty** to find a word in the text/sequence is computed then the "index" of two different sequences/texts are compared. The Shannon entropy can be used to identify interesting parts in genes in order to focus on them. The **entropy score** being a measure of the level of variability in the sequence, a high score of entropy is an indication of the presence of common words. By removing the parts in the genes with common words, one can enhance the variability and thus the phylogenetic signal, reducing in the same time the computational time[131]. In Table 2, the common parts which will be removed would be the verbs “is” and “have” for example, also known as **stop words** in natural language processing.

4.2.8 Other applications of alignment-free methods

4.2.8.1 Genome dereplication

For example, the fast-growing number of available prokaryotic genomes, along with their uneven taxonomic distribution, is a problem when trying to assemble broadly sampled genome sets for phylogenomics and comparative genomics. Indeed, most of the new genomes belong to the same subset of **hyper-sampled** phyla (Figure 15), such as Proteobacteria and Firmicutes, or even to

single species, such as *Escherichia coli* (e.g., 105,081 out of 939,798 genomes in GenBank as of January 2021), while the continuous flow of newly discovered phyla prompts for regular updates of in-house databases. This situation makes it difficult to maintain sets of representative genomes combining lesser-known phyla, for which only few species are available (e.g., Lokiarchaeota), and sound subsets of highly abundant phyla (e.g., Cyanobacteria). A straightforward approach for automated selection would be useful but far too slow if alignment-based methods are used. Thus, an alternative is required.

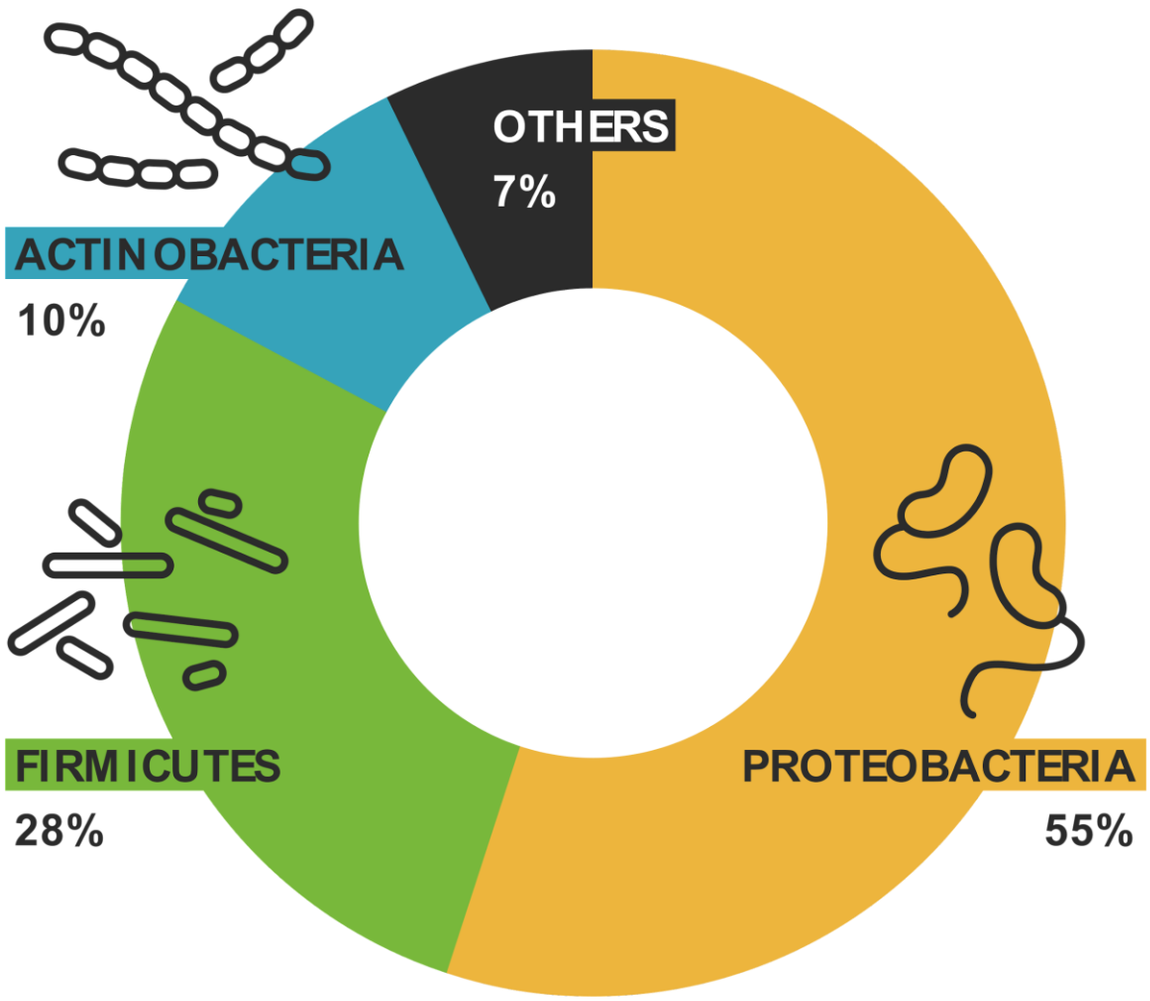


Figure 15 (from Léonard et al. 2021[132]): proportions of the top three phyla, in terms of number of genomes, in NCBI RefSeq Prokaryotes (March 2021) compared to the 50 other phyla.

4.2.8.2 Genome decontamination

As hinted above, alignment-free methods can be used for other applications than phylogenetics. For example, the program Kraken[133] is used to assign taxonomic labels to short DNA sequences. First it creates a database with every *k*-mer of the chosen genomes and assigns a **last common ancestor** (**LCA**) to every *k*-mer. If a *k*-mer is present only in a single species, then the LCA will be the species. If a *k*-mer is present in all genomes of a particular phylum or superphylum, then the LCA will be the phylum/superphylum, and this works for any taxonomic level. Once the database is built, one can in principle use Kraken with a genome not in the database and tag each

of its short DNA sequences with a taxonomic level[133]. In Cornet et al. (2018)[90], Kraken was also diverted from its primary use (along with other programs) to check if a genome is contaminated. Indeed, if a sequence is tagged as a cyanobacterium for example and that sequence belongs to a Planctomyces, then it could be considered as a clue that the Planctomyces genome is contaminated by Cyanobacteria. However, even if this is rarely acknowledged in the literature, the issue about long *k*-mers being too specific actually prevents Kraken from labeling genomes that are evolutionarily too distant from those used to build its database[90].

One of the other programs used for checking if a genome is contaminated is also based on *k*-mers, CONCOCT[134]. CONCOCT clusters assembly sequences into **non-hierarchical groups** based on a **Principal Component Analysis** (**PCA**) of short (4–6 nt) DNA *k*-mer frequencies. It was made to bin genome fragments from metagenomes. It clusters the reads into groups of reads belonging to the same species, making the separate assembly of the multiple genomes possible. By using this program on already assembled genomes split again in **pseudoreads**, the user can check if indeed all the pseudoreads really belong to a single species. The largest cluster (bin) of pseudoreads is considered to be the non-contaminated part of the genome and all other clusters are considered to be the contaminated parts of the genome. This diverted use of CONCOCT allows the user to estimate a level of contamination of a genome (but it is advised to use at least two different methods to check if a genome is contaminated or not) [90].

4.3 Towards an evolutionary synthesis for Bacteria

All these methods, except Gram staining and API, are based on the study of the phylogenetic signal and ignore everything else for the construction of evolutionary scenarios. While this “everything else” is not usable by itself, it should not be considered useless. In conjunction with phylogenetics, Cavalier-Smith, with his Neomuran hypothesis[32,33], further used the cell-wall architecture, one of the most important parts of a bacterial cell. Hence, using this information might be a solution to complete the phylogenetic signal brought by genetic sequences and produce even more accurate scenarios of evolution.

4.3.1 Cell-walls of monoderms, diderms and others

The traditional classification of bacteria is in two categories, the Gram positive bacteria (Gram+, Figure 16 left) possessing only a lipid bilayer and the Gram negative bacteria (Gram-, Figure 16 right) surrounded by two lipid bilayers. For the typical Gram+, the PG layer is the most external layer and has a thickness of 20 to 80 **nanometers** (**nm**), while the innermost layer is the plasma membrane. These two elements are separated by a thin periplasmic space. Moreover, the Gram+ cell wall contains teichoic acids. Contrary to the Gram+, the typical Gram- bacteria have an asymmetric external/outer membrane, with **lipopolysaccharides** (**LPS**) on its external side and phospholipids on its internal side. In Gram- bacteria, the PG layer is between the plasma membrane and the external/outer membrane (periplasm of 7-8 nm) and is 1 to 3 nm thick.

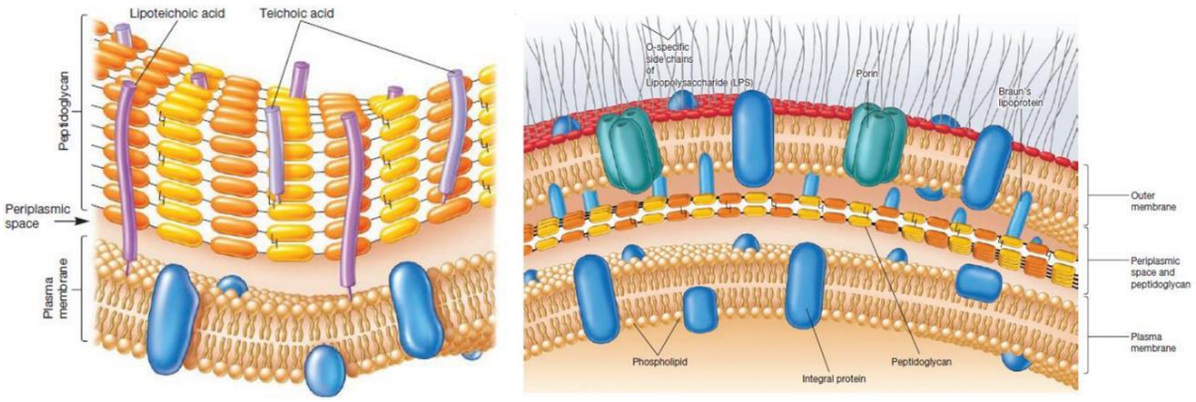


Figure 16 (Prescott, 2007[135]): illustration of the two classic architectures, Gram + (left) and Gram - (right).

But the reality is far from being that simple. Exceptions exist, for example, Tenericutes do not have PG and yet are identified as Gram-. They only have a single membrane, while phylogenetic analyses have demonstrated that they are derived from Firmicutes, which are Gram+ bacteria[136]. From an ultrastructural point of view, it would thus be wiser to refer to the terms monoderm and diderm[2,32] to distinguish bacteria with one or two membranes, even if this modified nomenclature is not satisfactory either from a phylogenetic point of view.

At this time (September 2021), bacteria are classified into 167 phyla, including 115 "**Candidatus**" phyla. A Candidatus taxon level is for prokaryotes that could not be described sufficiently for the creation of a new taxon[137]. The true number of bacterial phyla is currently unknown, due to challenges in the culture of bacteria, most being uncultured[138,139]. A solution is to sequence the metagenomes as demonstrated in the 2017 article by Parks et al.[78] where they manage to publish nearly 8000 new genomes (MAGs) by sequencing approximately 1500 sample metagenomes.

As for the Archaea, there are 38 phyla including 33 Candidatus phyla. Like the bacteria, the true number of archaeal phyla is currently unknown for the same reasons, and the solution to uncover even more phyla would also be to mine ever more metagenomes[78,140] while others painfully try to find experimental procedures allowing us to grow pure culture of these organisms[141].

The monoderm group is essentially composed of bacteria belonging to the Actinobacteria and Firmicutes phyla. However, these two phyla are not exclusively composed of monoderms. Indeed, Negativicutes possess two membranes combined to a thick PG. Phylogenetic analysis of their SSU rRNA 16S and of their orthologous proteins have shown that they belong to Firmicutes[142]. They are thus considered now as a class of Firmicutes, just like Bacilli and Clostridia. Diderms represent the largest share in the number of fully sequenced bacterial genomes. As of January 2021, there are 211,001 fully sequenced genomes, and 117,617 of them belong to Proteobacteria, a phylum of diderms. The diderm organisms thus represent well over the majority of the bacterial genomes (in terms of cell-wall architecture)[14].

Proteobacteria represent the archetype of the diderms-LPS, the "true Gram negative"[143] or the Glycobacteria[144,145]. These true Gram- are the only bacteria with LPS in their outer membrane, the other diderms lacking it. For example, bacteria of the Deinococcus-Thermus phylum have an external membrane with different glycolipids than the LPS. The Thermotogae have a proteic envelope instead of a lipidic one called the **toga**[146]. But being part of the Proteobacteria does not mean that the presence of the LPS is certain, as demonstrated by *Sphingomonas* with their

glycosphingolipid-based outer membrane without LPS[14,143,147]. A characteristic of the diderm-LPS, already mentioned above in the “Interlude - Rare genomic changes” section, is an insertion in the sequence of the proteins Hsp70 and Hsp60. These two molecular synapomorphies are only present in the diderms-LPS and differentiate the latter from the monoderms[24].

There are also differences between the PG of monoderms and diderms-LPS. Monoderms have additional polymers covalently attached to their PG, such as **teichoic acid**. Their membrane also has different lipids anchored, like **lypoglycans** and **lipoteichoic acids**[143].

4.3.2 Proteins for cell division

When dividing, the “**mother**” cell accumulates biomass and a second copy of its **chromosome(s)** to be separated into two “**daughter**” cells. In *E. coli*, more than 30 proteins are involved in the assembly of the **divisome**, a protein complex responsible for the formation of the **septum** that will separate the two “daughter” cells[148,149]. In this introduction, we will provide a simplified version of the division process for *E. coli*.

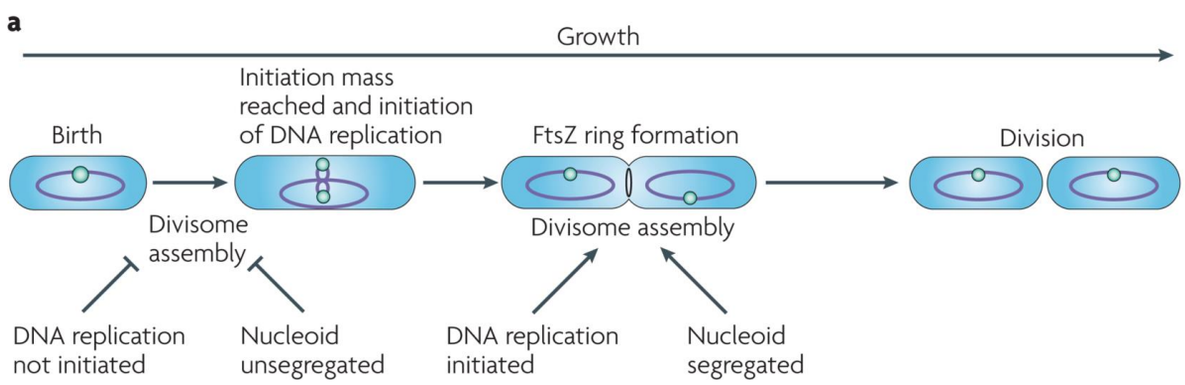


Figure 17: Division in rod shaped bacteria (adapted from Wang & Levin, 2009[150]). The green sphere represents the origin of replication and the purple ellipsis the chromosome.

The divisome assembly begins with the polymerisation of FtsZ into filaments forming the Z-ring and its anchorage in the plasma membrane by the FtsA and ZipA proteins (Figure 17). The abbreviation “**Fts**", attributed to many divisome related proteins, is the acronym for **filamentous temperature-sensitive** and comes from *E. coli* mutants that show a filamentous **phenotype** because of their incapacity to divide outside a certain range of temperature[151]. This ring recruits FtsK, which then recruits the FtsQ/FtsB/FtsL triplet. This subcomplex allows the recruitment and participates in the regulation of the FtsW/FtsI subcomplex. FtsI (also named **PBP3** for **Penicilin-Binding Protein**) is responsible for the cross-linking of the PG[148]. At first, FtsW was thought to be responsible for the translocation of the precursors of the PG for their integration to the existing PG (**flippase**) but an article from Mohammadi et al. (2011)[152] contested this role and attributed it to MurJ instead. Now it is established that MurJ indeed is the flippase and that FtsW has a role of **glycosyl transferase**[153–155].

4.3.2.1 FtsA/FtsZ

FtsZ is a protein homologous to the eukaryotic **tubulin**[156]. It possesses a **GTPase** activity and is able to form rings where the concentration of MinC regulated by the ***MinCDE* system** is the lowest, the latter system existing to control the position of the division site. MinCD inhibits the division by preventing the polymerisation of FtsZ while MinE blocks the inhibition by MinCD. The oscillation of the proteins within the cell makes the concentration of MinCD lower at the centre of the cell compared to the poles[157], see Figure 18. This gradient of concentration, coupled with the phenomenon of **nucleoid occlusion**, prevents the cell to create a Z-ring at the poles of the cell or to cut the nucleoid if it is still at the center of the cell[158]. The nucleoid occlusion is caused by the protein SlmA, orbiting around the chromosome, which prevents the polymerization of FtsZ in its vicinity[159].

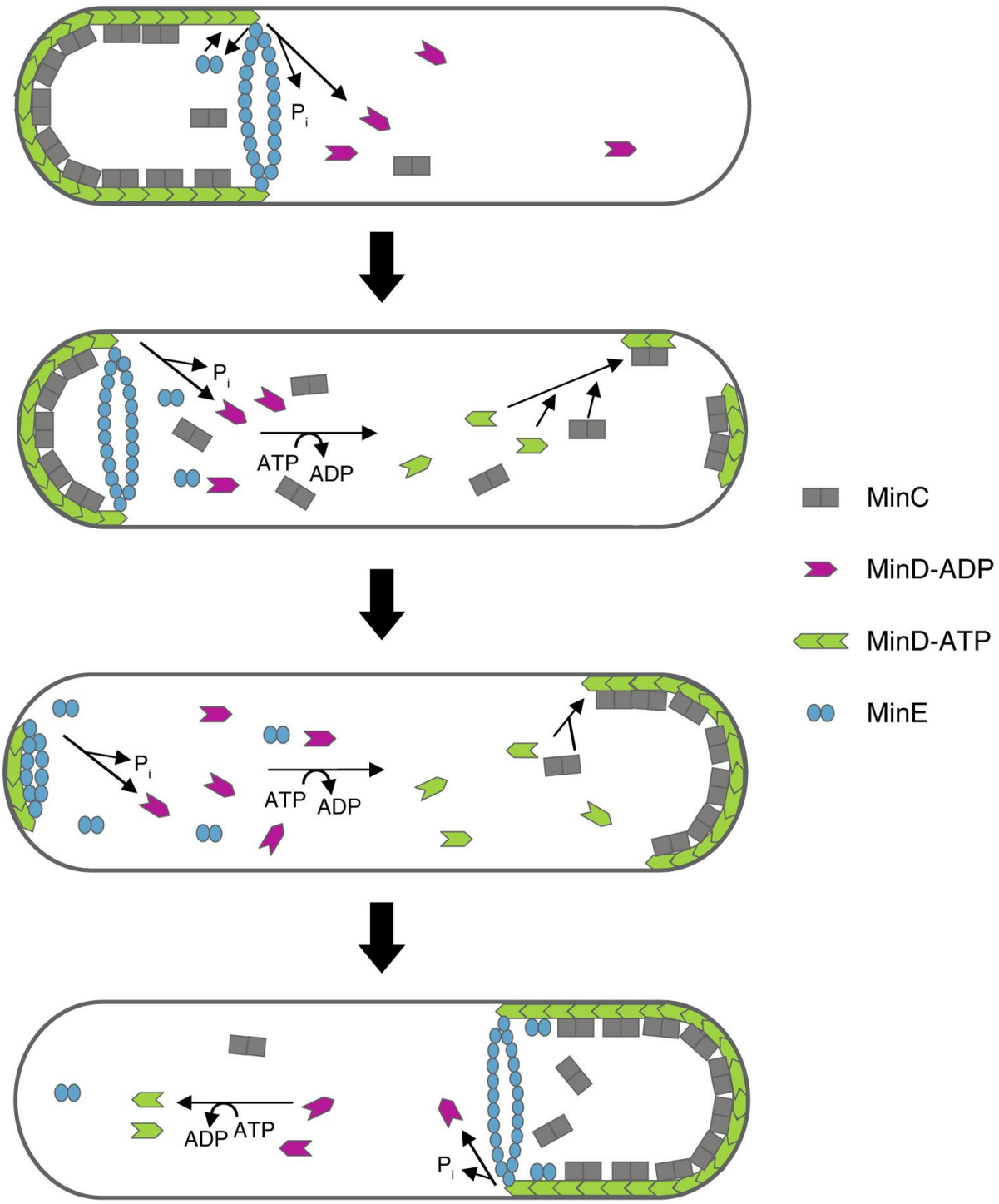


Figure 18 (from Lutkenhaus 2007[158] Figure 2): MinCD inhibits the division. The oscillation of the Min proteins in *E. coli* prevents MinCD to inhibit the division by creating a zone of lower concentration of MinCD at the center of the cell where FtsZ can form the Z-ring.

The ring made by FtsZ serves as a scaffold for the other proteins of the divisome. FtsZ is present in every bacteria studied so far, except Chlamydiae[160–163].

FtsA can hydrolyse ATP and plays a similar role to the eukaryotic ATP-binding protein from the actin Hsc70 family[164]. These two proteins, FtsA and FtsZ, act like a bacterial **cytoskeleton** assembled only during the division. The function of FtsA, with ZipA, is to link FtsZ to the plasma membrane.

4.3.2.2 FtsK

The protein FtsK is involved in the separation of the two copies of the chromosome before the separation of the two daughter cells[165]. It interacts with FtsZ, FtsQ, FtsL and FtsI, which are other proteins of the divisome[165–168].

4.3.2.3 FtsQ/FtsL/FtsB

The proteins FtsQ, FtsB and FtsL form a complex that links the events of the division in the cytoplasm with those occurring in the periplasm[148,169]. They are all **bitopic** (transmembrane proteins crossing the bilayer only once) and have most of their polypeptide chain located in the periplasm[170]. The complex of these three proteins is formed before its recruitment by FtsK on the Z-ring. Once bound to the Z-ring, the complex itself recruits the complex FtsI/FtsW. FtsL is a small protein with a **leucine-zipper motif** and a **transmembrane helix**. It also possesses a region where the tertiary structure is disordered. FtsB also possesses a leucine-zipper motif[148] whereas FtsQ is a protein with a POTRA-like domain (polypeptide-transport-associated)[171].

4.3.2.4 FtsI/FtsW

FtsI (PBP3) is a protein involved in the cross-linking of the PG at the level of the septum[164]. FtsW is, along with RodA and SpoVE, a member of the **SEDS** (**Shape Elongation Division Sporulation**) family, which is present in every bacterium with a PG cell wall[172] . FtsI is a class-B PBP, its C-terminal domain belongs to the **acyl-serine transferase family**, and its N-terminal domain has no known function. On the contrary, the N-terminal module in class-A PBPs (PBP1a and PBP1b) has a function to form the glycan chains. For both classes of PBP, the C-terminal domain binds the penicillin and possesses a **transpeptidase activity** involved in the formation of peptidic bridges between adjacent glycan chains. *In vivo*, FtsI requires FtsW to be recruited at the Z-ring[157].

4.3.3 Proteins for peptidoglycan biosynthesis

The PG, or murein, is a complex **heteropolymer** composed of long chains of **glycans** linked together by **short peptides**. The glycan chain is formed of an alternance of N-acetylglucosamine (GlcNAc) and N-acetylmuramic acid (MurNAc) linked together by β-(1-4) glycosidic bonds.

The **D-lactoyl group** of each MurNAc is substituted by an **oligopeptide L-Ala-gamma-D-Glu-meso-Dap**(or **L-Lys**)**-D-Ala-D-Ala**. The composition of this peptide can vary within the same taxonomic group. It is not synthesized by the ribosomal pathway, which allows amino acids in configuration D[173,174].

For *E. coli*, the glycan chains are constituted of 25 to 35 units of **disaccharide-pentapeptides**[175] and are linked together by interpeptidic bridges. Such a bridge is formed by the COOH group of

the D-Ala in position 4 of a pentapeptide, the last D-Ala being removed in the process, and the NH_2 group of the diamino acid of a pentapeptide of a neighboring glycan chain[176].

The biosynthesis of the PG requires 20 reactions in the cytoplasm and on the internal and external sides of the plasma membrane.

The biosynthesis takes place in three steps[157] (see Figure 19):

1. Formation of the **UDP-MurNAc-pentapeptide** is catalysed by cytoplasmic enzymes. The action of MurA and MurB, a transferase and a dehydrogenase, is to convert a UDP-GlcNAc precursor to a UDP-MurNAc precursor. MurC transfers an L-Ala on the MurNAc while MurD transfers a D-Glu to this L-Ala. MurE then attaches the meso-Dap to the D-Glu while MurF attaches a D-Ala-D-Ala, produced by DdlB, to the meso-Dap[156,177].
2. Transfer of the UDP-MurNAc-pentapeptide on the **undecaprenyl-phosphate** by MraY and the addition of GlcNAc by MurG to form the Lipid II. MraY links the UDP-MurNAc-pentapeptide on the undecaprenyl-phosphate, itself attached to the plasma membrane. Once linked to the undecaprenyl-P, the GlcNAc is added to the MurNAc to form the disaccharide-pentapeptide.
3. Transfer into the periplasm of the **disaccharide-pentapeptide** for polymerisation of the glycan chains and cross-linking of the peptides to form the PG. PBP1b, FtsI (PBP3) and FtsW act during the division, while the proteins PBP1a, PBP2 and RodA, which are equivalent to the aforementioned proteins, are in charge of this phase during the elongation[152,154,155,178].

The complete names of the genes involved in the synthesis of the PG precursor can be found in Table 3.

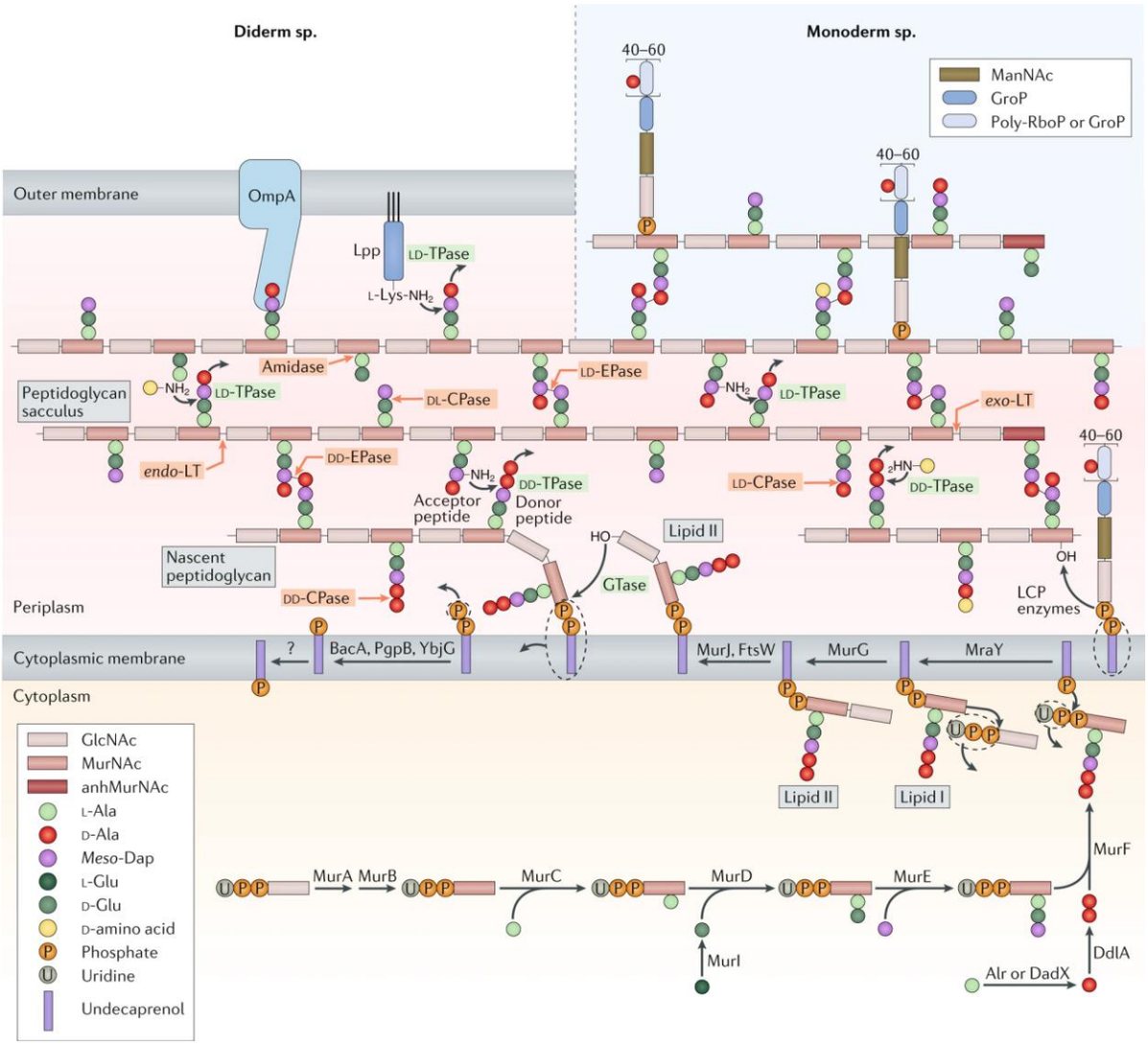


Figure 19 (from Egan et al. 2020[177], Figure 1): PG synthesis and cleavage. GTases = glycosyltransferases; TPases = transpeptidases; CPases = carboxypeptidases; EPases = endopeptidases; LT = lytic transglycolases; OmpA = outer membrane protein A; RboP/GroP = ribitol/glycerol phosphate; Lpp = lipoprotein. See the text above for more details.

Abbreviation	Complete name
MurA	UDP-GlcNAc enolpyruvyl transferase
MurB	UDP-MurNAc dehydrogenase
MurC	UDP-N-acetylmuramate L-alanine ligase
MurD	UDP-N-acetylmuramate-L-alanine D-glutamate ligase
MurE	UPD-N-acetylmuramate-L-alanine-D-glutamate meso-diaminopimelate ligase
DdlB	D-alanine D-alanine ligase
MurF	UDP-N-acetylmuramoyl-tripeptide D-alanyl-D-alanine ligase
MraY	phospho-N-acetylmuramoyl-pentapeptide transferase
MurG	N-acetylglucosaminyltransferase

Table 3: Names and abbreviations of proteins involved in lipid II synthesis.

4.3.4 Organisation of the cell-division and cell-wall genes – the DCW cluster

Some of these proteins, along with others, are encoded by genes located in the **division and cell-wall synthesis** (***dcw***) cluster, which is one of the most conserved clusters in bacteria[179]. It is composed of genes involved in either cell division or PG synthesis. The genes involved in cell division are *ftsA*, *ftsI*, *ftsL*, *ftsQ*, *ftsW* and *ftsZ*, whereas those involved in PG synthesis are *ddlB*, *mraY*, *murC*, *murD*, *murE*, *murF* and *murG*. The last two genes, *mraW* and *mraZ*, have functions not fully established. The following figure (20) represents the 15 gene *dcw* cluster of *E. coli*.

Figure 20 *E. coli dcw* cluster. Light and dark green: PG synthesis; orange and yellow: translocation and assembly of PG units; red: FtsQ and FtsL; purple: division; blue: MraZ and MraW.

A study from Mingorance & Tamames (2004)[179], based on around 40 genomes, has revealed that the “bacilli” and “filamentous bacteria” have a complete *dcw* cluster of 15 genes, whereas “cocci” (or unclassable bacteria like *Helicobacter* or *Spirochaeta*) have an incomplete cluster or even a non-existing one. However, from a taxonomic point of view, these designations are not operational and one would imply to re-examine the detailed list of genomes used before drawing any valid evolutionary interpretation.

Mingorance & Tamames (2004)[179] suggests that the **last bacterial common ancestor** (**LBCA**) already had the *dcw* cluster, which has been mainly transmitted vertically. Genes getting lost as the evolution unfolds, groups with a complete (or almost) cluster like *E. coli* would be considered to have a “primitive” form of the cluster. Another possibility would be that the *dcw* cluster was

assembled independently in several phyla, which would imply that the organisms with a reduced cluster have a “primitive” form.

We note here a phylogenetically incorrect reasoning implying that the organisms with the “ancestral” organization are older than those with the derived situation. Albeit common, this prejudice (also tracing back to the *Scala Naturae* mentioned above) is often wrong because losses (and other disorganization events) can very well happen in several evolutive lineages in parallel (secondary simplification). Thus, an organism with a primitive form of a characteristic is not necessarily older than every other organism compared. A bacterial example is mycoplasmas.

Mycoplasmas are the **smallest replicating bacteria known**. Discovered in 1898[180], they were even at first classified as viruses. In 1973, they were considered to be the most primitive organisms and placed at the root of the Tree of Life[181]. In the 1980s, however, following the works of Woese[182,183], mycoplasmas were re-classified as an offshoot of firmicutes[136], which makes them far more recent than the root of the Tree of Life. Their apparent simplicity is thus a secondary simplification, albeit a heavy one.

In terms of genomic predictions, if we start from a complete cluster, the general order of the genes should at least partially persist even after a partial disaggregation of the cluster. In contrast, genes clustered independently should lead to clusters with different gene orders, also called synteny, from branch to branch.

4.3.4.1 Intruders – MraW and MraZ

MraW and MraZ are not involved in cell division or synthesis of the PG but are nonetheless part of the *dcw* cluster. MraW is present in every bacterial genome and is always absent in archaeal and eukaryotic genomes. MraZ is sometimes absent but, when present, it precedes immediately MraW and is oriented like it, the two genes sharing, along with the following nine genes of the *dcw* cluster, the same promoter[184,185].

MraW, also known as RsmH, is a **methyltransferase** which, with YraI/RsmI, methylates a nucleotide located in the decoding centre of the SSU rRNA 16S, the m4Cm[186]. Concerning MraZ, its N-terminal sequence bears similarities to bacterial proteins AbrB (N-terminal) and MazE (addiction module), and is suspected to be a transcriptional regulator like them. Yet, its function is still unknown. The simultaneous deletion of MraZ and MraW does not have any visible effect, but any change in their relative proportions is toxic for the cell[185].

4.3.5 And the genes of the outer membrane?

It is of note that the *dcw* cluster is the division and cell-wall cluster but the genes involved with the cell wall concerns only the PG while the genes involved with the other less common cell-wall components, such as the **outer membrane** (OM), are located elsewhere in the genome. This is normal, since the *dcw* cluster is common (in one form or another) to all Bacteria, whether they have an OM or not. So, what are (some) the genes involved with the OM (when there is one) and what do they do?

As mentioned above, the OM is asymmetrical and thus the proteins involved in its creation and maintenance are bound to be different than for the plasma membrane. The OM would need proteins to add the different constituting elements of the outer part of the OM, like the proteins with the Bam pathway or the LPS with the Lpt pathway. It would also need proteins to add the constituting elements to the inner part of the OM, like the lipoproteins with the Lol pathway. In addition, the Tol-Pal system has been associated with multiple roles during cell division, e.g., ensuring that OM invagination and cell-wall processing are properly coordinated[187].

4.3.5.1 Bam pathway

Two types of integral membrane proteins exist, the α-helical and the β-barrel proteins. The Bam pathway is responsible for the folding and insertion of the β-barrel proteins into the OM[188].

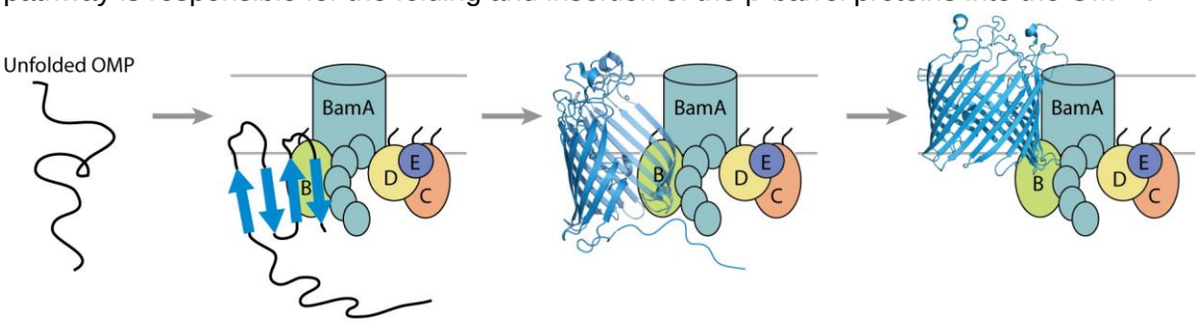


Figure 21: Bam pathway from Hagan et al. 2011[188]. By binding an unfolded OM protein, BamA initiates the β-structure formation and BamBCDE help to stabilize the interaction and help with the dissociation once the protein is folded.

The Bam complex, as seen in Figure 21, is composed of BamA, the main and essential protein of the complex, followed by four other proteins, BamB, BamC, BamD and BamE, forming the complex that folds the unfolded OM protein then inserts the now folded OM protein into the OM. The Bam complex does not support the transport of the unfolded protein from the plasma membrane (or inner membrane, hence IM) to the OM. Instead, it relies on the Sec pathway, which is common to the OM proteins and the IM proteins[188,189] , to transport the proteins from the IM to the periplasm, and relies on the SurA, Skp or DegP chaperone proteins to protect the unfolded protein during its travel through the periplasm[189].

4.3.5.2 Lol pathway

The Localization of lipoproteins (Lol) pathway is in charge for exporting lipoproteins from the IM to the insertion of the lipoproteins into the inner layer of the OM[190,191].

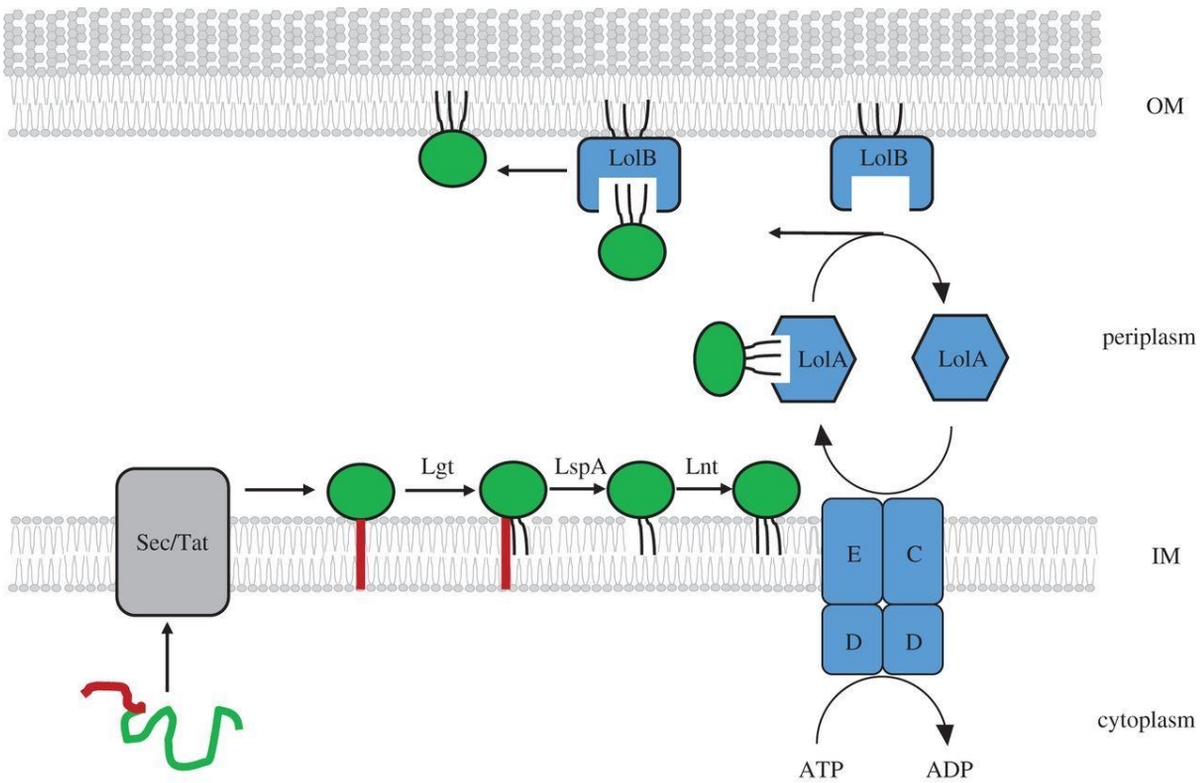


Figure 22: Lol pathway from Konovalova & Silhavy 2015[191]. The lipoprotein, in green, is flipped and inserted in the IM by its N-terminal signal sequence through the Sec/Tat pathway. The LolCDDE complex then extracts the lipoprotein from the IM to transport it to LolB on the OM by using the LolA chaperone.

As seen in Figure 22, the Lol pathway is dependent, like the Bam complex, from the Sec pathway for the passage of the lipoprotein from the IM to the periplasm. The first part of the complex, LolC, LolD (x2) and LolE, releases the lipoprotein from the IM and makes it available to LolA. The role of LolA is to chaperone the lipoprotein in the periplasm as it travels in the periplasm until it reaches LolB, located at the OM. LolB retrieves the lipoprotein from LolA and inserts it into the inner leaflet of the OM[191].

4.3.5.3 Lpt system

The role of the Lpt system is the transport of LPS from the IM to the outer leaflet of the OM[192]. Seven components form this system, LptABCDEFG.

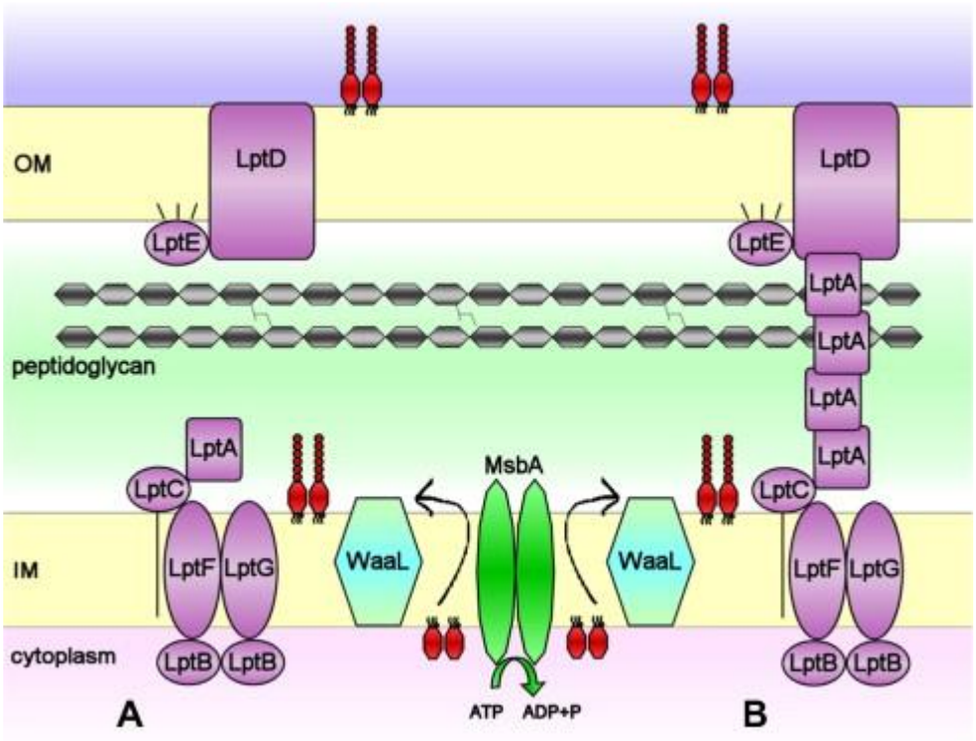


Figure 23: Lpt system, the two theories from Bowyer et al. 2011[192]. (A) a monomer or oligomer of LptA delivers the LPS through the periplasm; (B) a bridge of LptA proteins between the IM and OM delivers the LPS through the periplasm.

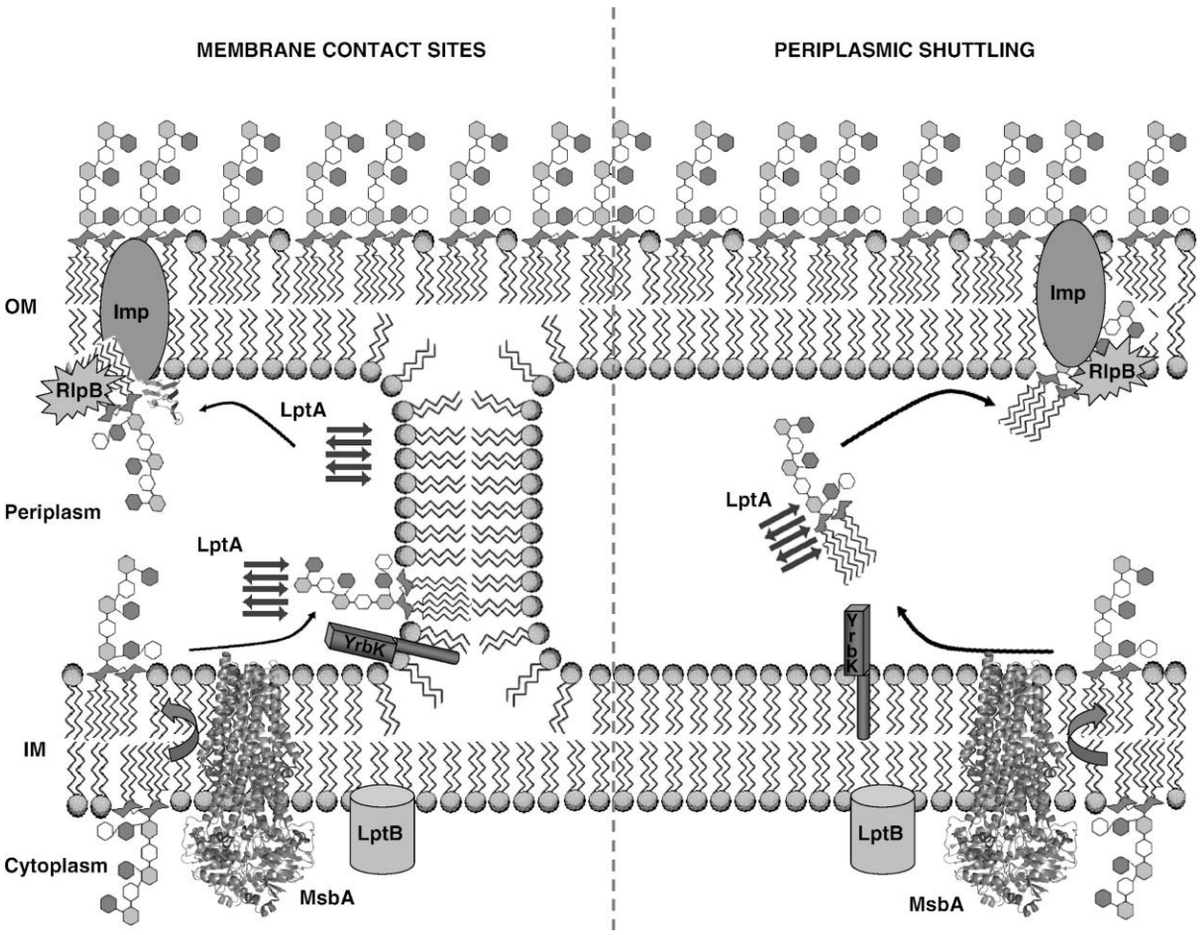


Figure 24: Lpt system, the two theories from Suits et al. 2008[193]. Left, LptA, linked to the LPS, guides the LPS through a Bayer bridge; Right, LptA, linked to the LPS, travels through the periplasm to deliver the LPS.

In Figure 23, two possible mechanisms explaining the transport of LPS are shown; a third possibility is also considered and shown in Figure 24 (left). The difference lies in the way it transports the LPS through the periplasm. In all cases, the complex LptB(x2)FG is supposed to release the LPS from the IM and give it to LptC, which then transfers it to LptA. The difference between the three mechanisms is how the LptA protein transports the LPS to the LptDE complex located on the inner leaflet of the OM. The first way is a monomer or oligomer of LptA making the travel through the periplasm to deliver the LPS. The second way is LptA forming a bridge between the IM and the OM, serving as a sort of conveyor belt of LPS. The last way is the formation of a Bayer junction[194], a contact point between the two membranes. Once the LPS has reached the OM, it is taken by the LptDE complex to transport and anchor it to the outer leaflet of the OM[192].

4.3.5.4 Tol-Pal system

The Tol-Pal system is a seven-component system consisting of TolA, TolB, TolQ and TolR, YbgC and YbgF, and finally Pal. Its function is still currently unknown but appears to be essential for maintaining the integrity of the OM. Indeed, mutants lacking these genes liberate periplasmic proteins or are more sensitive to drugs [195]. It is also used by the colicin (a toxin produced by *E. coli* and relatives) and the bacteriophages DNA to penetrate the cell[196].

YbgF is a cytoplasmic protein whereas TolA, TolQ and TolR are transmembrane proteins of the IM. Pal is an OM protein involved with the PG. The rest of the Tol-Pal system is periplasmic. It

has been suggested that Tol-Pal could have an important role during the cell division (at least in *E. coli*) since, during the division, it accumulates at the constriction sites[197]. Recent results show that the daughter cells of a *tol-pal* mutant (whole cluster) of *E. coli* remain attached by their PG layer, thereby hinting at a broader set of functions for the Tol-Pal system[187].

4.4 Objectives

The aim of this thesis was to study the evolution of the cell-wall architecture in the bacterial domain, and more specifically to build a scenario based on phylogenomic and phenotypic data to account for cell-wall evolution from the LBCA to exant bacterial lineages. To this end we needed a diverse selection of genomes, the creation of orthologous groups (OGs) for these genomes, a phylogenomic tree, the status of the cell wall for the selection of Bacteria used for the tree, genes of interest involved with the cell wall, tools for cluster reconstruction and ancestral trait reconstruction. Two main chapters describe our work, the first has been published in *PeerJ* and the second is currently in preparation for a first submission in *Frontiers in Genetics*.

The study of the evolution of the cell-wall architecture is of interest due to its complex situation in prokaryotes. Indeed, in Figures 1 and 5 from the Introduction of this work, we can see complex cell-wall architectures in both Archaea and Bacteria. They are however not neatly distributed within these groups, preventing these morphological characteristics to be used for the classification of the prokaryotes like morphological features can often be informative in the Animal kingdom. The cell wall remains nonetheless an important part of the prokaryotic cell, and the study of its complicated evolution represents an exciting endeavor. As it would be an almighty task if taken in its entirety, we limited ourselves to the study of the bacterial cell-wall architecture instead of the prokaryotic cell-wall architecture, hence excluding Archaea.

The diverse selection of genomes mentioned above is mandatory due to the sheer number of available genomes and their redundancy (Figure 15). Indeed, we need a more manageable number of genomes while maintaining the diversity (around a thousand genomes instead of more than 200,000 as of January 2021). Moreover, due to the (ever growing) size of the dataset that needs to be dereplicated, this first essential step has to be automated and easily scalable. This is why we created ToRQuEMaDA (Tool for Retrieving Queried Eubacteria, Metadata and Dereplicating Assemblies; TQMD)[132] to perform this task, through full genome comparison based on *k*-mers and a divide-and-conquer approach to produce a powerful and scalable tool, competitive with other existing tools. This part of the thesis has been published in *PeerJ* and is directly made available as the next chapter (or at the following address: https://peerj.com/articles/11348/).

Based on the result of an early version of TQMD, we produced OGs groups to identify the most conserved genes to use for our phylogenomic tree for our specific selection of genomes and also to study the synteny of the genes of interest involved with the cell wall (and maybe identify other genes of interest). Using the phylogenomic tree as a base, we could reconstruct the status of the cluster in the LBCA for the genes of interest belonging to a cluster. We could also reconstruct the LBCA cell wall once the cell wall of the organisms used in the phylogenomic tree were known through bibliographic searches. From the harvested information, possible scenarios for the evolution of the LBCA cell wall architecture could be devised. These steps are further expanded in the dedicated part of this work and are currently being prepared for a first submission.

4.5 References

1. Gram, C. Ueber die isolirte Farbung der Schizomyceten in Schnitt-und Trockenpraparaten. *Fortschritte Med.* **2**, 185–189 (1884).

2. Gupta, R. S. Life's Third Domain (Archaea): An Established Fact or an Endangered Paradigm?: A New Proposal for Classification of Organisms Based on Protein Sequences and Cell Structure. *Theor. Popul. Biol.* **54**, 91–104 (1998).

3. Coico, R. Gram staining. *Curr. Protoc. Microbiol.* A-3C (2006).

4. Venter, S., Lotter, L., de Haas, D. & MacDonald, L. The use of the analytical profile index in the identification of activated sludge bacteria: problems and solutions. 4 (1989).

5. Cain, T. C., Lubman, D. M. & Weber, W. J. Differentiation of bacteria using protein profiles from matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **8**, 1026–1030 (1994).

6. Stanier, R. Y. & van Niel, C. B. The Concept of a Bacterium. **35**, 17–35 (1962).

7. Chatton, E. *Titres et Travaux Scientifiques de Edouard Chatton 1906–1937*. (Sottano: Sette, 1938).

8. Sapp, J. The Prokaryote-Eukaryote Dichotomy : Meanings and Mythology. **69**, 292–305 (2005).

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

10. Hartmann, E. & König, H. Comparison of the biosynthesis of the methanobacterial pseudomurein and the eubacterial murein. *Naturwissenschaften* **77**, 472–475 (1990).

11. Woese, C. R., Kandlert, O. & Wheelis, M. L. Towards a natural system of organisms : Proposal for the domains Archaea, Bacteria and Eucarya. **87**, 4576–4579 (1990).

12. Albers, S.-V. & Meyer, B. H. The archaeal cell envelope. *Nat. Rev. Microbiol.* **9**, 414–426 (2011).

13. Klingl, A., Pickl, C. & Flechsler, J. Archaeal Cell Walls. in *Bacterial Cell Walls and Membranes* 471–493 (Springer, 2019).

14. Zerbib, D. Bacterial Cell Envelopes: Composition, Architecture, and Origin. in *Handbook*

of Electroporation (ed. Miklavčič, D.) 417–436 (2017). doi:10.1007/978-3-319-32886-7.

15. Rodrigues-oliveira, T. *et al.* Archaeal S-Layers : Overview and Current State of the Art. *Front. Microbiol.* **8**, 1–17 (2017).

16. Kates, M. Archaebacterial lipids: structure, biosynthesis and function. in *Biochemical Society Symposium* vol. 58 51–72 (1992).

17. Revell, L. J., Harmon, L. J. & Collar, D. C. Phylogenetic signal, evolutionary process, and rate. *Syst. Biol.* **57**, 591–601 (2008).

18. Philippe, H. & Moreira, D. Molecular phylogeny : pitfalls and progress. 9–16 (2000).

19. Ludwig, W. & Klenk, H.-P. Overview: A Phylogenetic Backbone and Taxonomic Framework for Procaryotic Systematics. in *Bergey's manual of Systematics of Archaea and Bacteria* (John Wiley & Sons, Inc., 2001).

20. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).

21. Woese, C. R., Olsen, G. J., Ibba, M. & Söll, D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* **64**, 202–236 (2000).

22. Baldauf, S. L., Palmert, J. D. & Doolittle, W. F. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. **93**, 7749–7754 (1996).

23. Rokas, A. & Holland, P. W. H. Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* **15**, 454–459 (2000).

24. Gupta, R. S. 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 Int. J. Gen. Mol. Microbiol.* **100**, 171–182 (2011).

25. Gupta, R. S. The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiol. Rev.* **24**, 367–402 (2000).

26. Griffiths, E. & Gupta, R. S. Signature sequences in diverse proteins provide evidence for the late divergence of the Order Aquificales. 41–52 (2004).

27. Lake, J. A., Herbold, C. W., Rivera, M. C., Servin, J. A. & Skophammer, R. G. Rooting the Tree of Life Using Nonubiquitous Genes. **24**, 130–136 (2007).

28. Singh, B. & Gupta, R. S. Conserved inserts in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins are essential for cellular growth. **60**, 361–373 (2009).

29. Lovejoy, A. The great chain of being. *Camb. Mass Harv. Univ. Press* **19**, 36 (1936).

30. Granger, H. The scala naturae and the continuity of kinds. *Phronesis* **30**, 181–200 (1985).

31. Barham, J. On the Objectivity of the Scala Naturae. *Evol. Cogn.* **5**, (1999).

32. Cavalier-Smith, T. The neomuran origin of archaebacteria , the negibacterial root of the universal tree and bacterial megaclassification. 7–76 (2002).

33. Cavalier-smith, T. The Neomuran Revolution and Phagotrophic Origin of Eukaryotes and Cilia in the Light of Intracellular Coevolution and a Revised Tree of Life. (2014) doi:10.1101/cshperspect.a016006.

34. Gribaldo, S., Poole, A. M., Daubin, V. & Forterre, P. The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? **8**, 743–752 (2010).

35. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).

36. Baurain, D., Wilmotte, A. & Frère, J.-M. Gram-Negative Bacteria:" Inner" vs." Cytoplasmic" or" Plasma Membrane": A Question of Clarity rather than Vocabulary. *J. Microb. Biochem. Technol.* **8**, 325–326 (2016).

37. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* **39**, 309–338 (2005).

38. Bryant, D. & Hahn, M. W. The Concatenation Question. in *Phylogenetics in the Genomic Era* (eds. Scornavacca, C., Delsuc, F. & Galtier, N.) 3.4:1-3.4:23 (No commercial publisher, 2020).

39. Anisimova, M. *et al.* State-of the art methodologies dictate new standards for phylogenetic analysis. (2013) doi:10.1186/1471-2148-13-161.

40. Philippe, H. & Roure, B. Difficult phylogenetic questions : more data , maybe ; better methods , certainly. 2–5 (2011).

41. Philippe, H. *et al.* Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol.* **9**, (2011).

42. Philippe, H. *et al.* Pitfalls in supermatrix phylogenomics. (2017).

43. Simion, P., Delsuc, F. & Philippe, H. To What Extent Current Limits of Phylogenomics Can Be Overcome? in *Phylogenetics in the Genomic Era* (eds. Scornavacca, C., Delsuc, F. & Galtier, N.) 2.1:1-2.1:34 (No commercial publisher, 2020).

44. Kapli, P., Yang, Z. & Telford, M. J. Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* (2020) doi:10.1038/s41576-020-0233-0.

45. Gribaldo, S. & Brochier, C. Phylogeny of prokaryotes: does it exist and why should we care? *Res. Microbiol.* **160**, 513–521 (2009).

46. Fernández, R., Gabaldón, T. & Dessimoz, C. Orthology: definitions, inference, and impact on species phylogeny inference. in *Phylogenetics in the Genomic Era* (eds. Scornavacca, C., Delsuc, F. & Galtier, N.) 2.4:1-2.4:14 (No commercial publisher, 2020).

47. Jensen, R. A. Orthologs and paralogs - we need to get it right. 1–3 (2001).

48. Fitch, W. M. Homology: a personal view on some of the problems. *Trends Genet.* **16**, 227–231 (2000).

49. Koonin, E. V. An apology for orthologs-or brave new memes. *Genome Biol.* **2**, 1–2 (2001).

50. Battistuzzi, F. U. & Hedges, S. B. A major clade of prokaryotes with ancient adaptations to life on land. *Mol. Biol. Evol.* **26**, 335–343 (2009).

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

52. Lartillot, N. The Bayesian Approach to Molecular Phylogeny. in *Phylogenetics in the Genomic Era* (eds. Scornavacca, C., Delsuc, F. & Galtier, N.) 1.4:1-1.4:17 (No commercial publisher, 2020).

53. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965–968 (2006).

54. Philippe, H. *et al.* Acoelomorph flatworms are deuterostomes related to Xenoturbella. (2011) doi:10.1038/nature09676.

55. Simion, P. *et al.* A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr. Biol.* **27**, 958–967 (2017).

56. Irisarri, I. *et al.* Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat. Ecol. Evol.* **1**, 1370–1378 (2017).

57. Wickett, N. J. *et al.* Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E4859-4868 (2014).

58. Leebens-Mack, J. H. *et al.* One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).

59. Li, Y. *et al.* A genome-scale phylogeny of the kingdom Fungi. *Curr. Biol. CB* **31**, 1653-1665.e5 (2021).

60. Hampl, V. *et al.* Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic “supergroups”. *Proc. Natl. Acad. Sci.* **106**, 3859–3864 (2009).

61. Burki, F. *et al.* Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc. Biol. Sci.* **283**, 20152802 (2016).

62. Boussau, B., Guéguen, L. & Gouy, M. Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. **18**, 1–18 (2008).

63. Yutin, N., Puigbo, P., Koonin, E. V. & Wolf, Y. I. Phylogenomics of Prokaryotic Ribosomal Proteins. *Curr. Sci.* **101**, 1435–1439 (2012).

64. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

65. Lasek-nesselquist, E. & Gogarten, J. P. The effects of model choice and mitigating bias on the ribosomal tree of life. *Mol. Phylogenet. Evol.* **69**, 17–38 (2013).

66. Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci.* 201420858 (2015) doi:10.1073/pnas.1420858112.

67. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

68. Castelle, C. J. & Banfield, J. F. Major new microbial groups expand diversity and alter our

understanding of the tree of life. *Cell* **172**, 1181–1197 (2018).

69. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* (2019) doi:10.1038/s41467-019-13443-4.

70. Cavalier-Smith, T., Ema, E. & Chao, Y. Multidomain ribosomal protein trees and the planctobacterial origin of neomura (eukaryotes, archaebacteria). *Protoplasma* 1–133 (2020).

71. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8**, 275–282 (1992).

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

73. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Syst. Biol.* **62**, 611–615 (2013).

74. Le, S. Q., Dang, C. C. & Gascuel, O. Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. *Mol. Biol. Evol.* **29**, 2921–2936 (2012).

75. Zoller, S. & Schneider, A. Improving Phylogenetic Inference with a Semiempirical Amino Acid Substitution Model. *Mol. Biol. Evol.* **30**, 469–479 (2013).

76. Le, S. Q., Lartillot, N. & Gascuel, O. Phylogenetic mixture models for proteins. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 3965–3976 (2008).

77. Mirarab, S. *et al.* ASTRAL : genome-scale coalescent-based species tree estimation. **30**, 541–548 (2014).

78. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, (2017).

79. Felsenstein, J. *Inferring phylogenies*. vol. 2 (Sinauer associates Sunderland, MA, 2004).

80. Liò, P. & Goldman, N. Models of molecular evolution and phylogeny. *Genome Res.* **8**, 1233–1244 (1998).

81. Whelan, S., Liò, P. & Goldman, N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *TRENDS Genet.* **17**, 262–272 (2001).

82. Uzzell, T. & Corbin, K. W. Fitting discrete probability distributions to evolutionary events. *Science* **172**, 1089–1096 (1971).

83. Lopez, P., Casane, D. & Philippe, H. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19**, 1–7 (2002).

84. Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. & Delsuc, F. Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol.* **5**, 1–8 (2005).

85. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).

86. Sullivan, J. & Joyce, P. Model Selection in Phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* **36**, 445–466 (2005).

87. Pupko, T. & Mayrose, I. A gentle introduction to probabilistic evolutionary models. in *Phylogenetics in the Genomic Era* (No commercial publisher| Authors open access book, 2020).

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

89. Bernard, G. *et al.* Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Brief. Bioinform.* **20**, 426–435 (2019).

90. Cornet, L. *et al.* Consensus assessment of the contamination level of publicly available cyanobacterial genomes. 1–26 (2018).

91. Longo, M. S., Neill, M. J. O. & Neill, R. J. O. Abundant Human DNA Contamination Identified in Non- Primate Genome Databases. **6**, 1–4 (2011).

92. Furuya, E. Y. & Lowy, F. D. Antimicrobial-resistant bacteria in the community setting. *Nat. Rev. Microbiol.* **4**, 36–45 (2006).

93. Barlow, M. What antimicrobial resistance has taught us about horizontal gene transfer. in *Horizontal Gene Transfer* 397–411 (Springer, 2009).

94. Lerminiaux, N. A. & Cameron, A. D. S. Horizontal transfer of antibiotic resistance genes in clinical environments. **44**, 34–44 (2019).

95. Gyles, C. & Boerlin, P. Horizontally transferred genetic elements and their role in

pathogenesis of bacterial disease. *Vet. Pathol.* **51**, 328–40 (2014).

96. Romiguier, J. & Roux, C. Analytical Biases Associated with GC-Content in Molecular Evolution. *Front. Genet.* **8**, 16 (2017).

97. Philippe, H. & Douady, C. J. Horizontal gene transfer and phylogenetics. 498–505 (2003) doi:10.1016/j.mib.2003.09.008.

98. Creevey, C. J., Doerks, T., Fitzpatrick, D. a., Raes, J. & Bork, P. Universally distributed single-copy genes indicate a constant rate of horizontal transfer. *PLoS ONE* **6**, (2011).

99. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).

100. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

101. Kozlov, A. M., Aberer, A. J. & Stamatakis, A. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* **31**, 2577–2579 (2015).

102. Nesbø, C. L. *et al.* The genome of Thermosipho africanus TCF52B: lateral genetic connections to the Firmicutes and Archaea. *J. Bacteriol.* **191**, 1974–1978 (2009).

103. Gupta, R. S. & Bhandari, V. Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups. *Antonie Van Leeuwenhoek* **100**, 1 (2011).

104. Bhandari, V., Naushad, H. S. & Gupta, R. S. Protein based molecular markers provide reliable means to understand prokaryotic phylogeny and support Darwinian mode of evolution. *Front. Cell. Infect. Microbiol.* **2**, 1–14 (2012).

105. Zhaxybayeva, O. *et al.* On the chimeric nature , thermophilic origin , and phylogenetic placement of the Thermotogales. **106**, 5865–5870 (2009).

106. Eveleigh, R. J. M., Meehan, C. J., Archibald, J. M. & Beiko, R. G. Being Aquifex aeolicus: Untangling a hyperthermophile's checkered past. *Genome Biol. Evol.* **5**, 2478–2497 (2013).

107. Jumas-Bilak, E., Roudiere, L. & Marchandin, H. Description of 'Synergistetes' phyl. nov. and emended description of the phylum 'Deferribacteres' and of the family Syntrophomonadaceae, phylum 'Firmicutes'. *Int. J. Syst. Evol. Microbiol.* **59**, 1028–1035

(2009).

108. Baurain, D., Brinkmann, H. & Philippe, H. Lack of Resolution in the Animal Phylogeny: Closely Spaced Cladogeneses or Undetected Systematic Errors? *Mol. Biol. Evol.* **24**, 6–9 (2007).

109. Bininda-Edmonds, O. R. P., Gittleman, J. L. & Steel, M. A. THE (SUPER) TREE OF LIFE : Procedures, Problems, and Prospects. (2002) doi:10.1146/annurev.ecolsys.33.010802.150511.

110. Ragan, M. A. Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.* **1**, 53–58 (1992).

111. Creevey, C. J. & McInerney, J. O. *Trees from Trees Trees from Trees : Construction of Phylogenetic Supertrees using Clann .* (2009). doi:10.1007/978-1-59745-251-9.

112. Tourasse, N. J. & Kolstø, A.-B. SuperCAT: a supertree database for combined and integrative multilocus sequence typing analysis of the Bacillus cereus group of bacteria (including B. cereus, B. anthracis and B. thuringiensis). *Nucleic Acids Res.* **36**, D461–D468 (2007).

113. Liu, L., Anderson, C., Pearl, D. & Edwards, S. V. Modern Phylogenomics: Building Phylogenetic Trees Using the Multispecies Coalescent Model. in *Evolutionary Genomics: Statistical and Computational Methods* (ed. Anisimova, M.) 211–239 (2019).

114. Rannala, B., Edwards, S. V., Leaché, A. D. & Yang, Z. The Multispecies Coalescent Model and Species Tree Inference. in *Phylogenetics in the Genomic Era* (eds. Scornavacca, C., Delsuc, F. & Galtier, N.) 3.3:1-3.3:21 (No commercial publisher, 2020).

115. Edwards, S. V. *et al.* Implementing and testing the multispecies coalescent model : A valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* **94**, 447–462 (2016).

116. Springer, M. S. & Gatesy, J. The gene tree delusion. *Mol. Phylogenet. Evol.* **94**, 1–33 (2016).

117. Mcvay, J. D. & Carstens, B. C. Phylogenetic Model Choice : Justifying a Species Tree or Concatenation Analysis. **1**, 1–8 (2013).

118. Maddison, W. P. Gene Trees in Species Trees. *Syst. Biol.* **46**, 523–536 (1997).

119. Coleman, G. A. *et al.* A rooted phylogeny resolves early bacterial evolution. *Science* **372**, (2021).

120. Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient Exploration of the Space of Reconciled Gene Trees. *Syst. Biol.* **62**, 901–912 (2013).

121. Zielezinski, A., Vinga, S., Almeida, J. & Karlowski, W. M. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.* **18**, 186 (2017).

122. Vinga, S. *Alignment-free methods in computational biology*. (Oxford University Press, 2014).

123. Miller, J. B., Mckinnon, L. M., Whiting, M. F. & Ridge, P. G. CAM : an alignment-free method to recover phylogenies using codon aversion motifs. *PeerJ* 1–24 (2019) doi:10.7717/peerj.6984.

124. Bernard, G., Chan, C. X. & Ragan, M. A. Alignment-free microbial phylogenomics under scenarios of sequence divergence , genome rearrangement and lateral genetic transfer. *Nat. Publ. Group* 1–12 (2016) doi:10.1038/srep28970.

125. Ren, F., Tanaka, H. & Yang, Z. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Syst. Biol.* **54**, 808–818 (2005).

126. Gil, M., Zanetti, M. S., Zoller, S. & Anisimova, M. CodonPhyML: fast maximum likelihood phylogeny estimation under codon substitution models. *Mol. Biol. Evol.* **30**, 1270–1280 (2013).

127. Kolmogorov, A. N. Three approaches to the quantitative definition of information. *Probl. Inf. Transm.* **1**, 1–7 (1965).

128. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

129. Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).

130. Tribus, M. & McIrvine, E. C. Energy and information. *Sci. Am.* **225**, 179–190 (1971).

131. Batista, M. V. A., Ferreira, T. A. E., Freitas, A. C. & Balbino, V. Q. An entropy-based approach for the identification of phylogenetically informative genomic regions of Papillomavirus. *Infect. Genet. Evol.* **11**, 2026–2033 (2011).

132. Léonard, R. R. *et al.* ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies. *PeerJ* **9**, e11348 (2021).

133. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

134. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).

135. Prescott, L. M., Hardy, M. P. & Klein, J. P. *Microbiology 8TH*. (McGraw Hill, 2006).

136. Wolf, M., Müller, T., Dandekar, T. & Pollack, J. D. Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. *Int. J. Syst. Evol. Microbiol.* **54**, 871–875 (2004).

137. Murray, R. G. E. & Schleifer, K. H. Taxonomic Notes : A Proposal for Recording the Properties of Putative Taxa of Procaryotes. 174–176 (1994).

138. Staley, J. T. & Konopka, A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. (1985).

139. Harwani, D. The Great Plate Count Anomaly and the Unculturable Bacteria. (2013).

140. Castelle, C. J. *et al.* Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nat. Commun.* **4**, 1–10 (2013).

141. Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).

142. Antunes, L. CS. *et al.* Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the Firmicutes. *eLife* **5**, 1–21 (2016).

143. Sutcliffe, I. C. A phylum level perspective on bacterial cell envelope architecture. *Trends Microbiol.* **18**, 464–470 (2010).

144. Cavalier-Smith, T. Rooting the tree of life by transition analyses. *Biol. Direct* **1**, 19 (2006).

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

146. Errington, J. L-form bacteria, cell walls and the origins of life. *Open Biol.* **3**, 120143 (2013).

147. White, D. C., Suttont, S. D. & Ringelberg, D. B. The genus Sphingomonas : physiology

and ecology. 301–306.

148. van den Berg van Saparoea, H. B. *et al.* Fine-mapping the Contact Sites of the Escherichia coli Cell Division Proteins FtsB and FtsL on the FtsQ Protein*,. *J. Biol. Chem.* **288**, 24340–24350 (2013).

149. Blaauwen, T. D., Hamoen, L. W. & Levin, P. A. The divisome at 25 : the road ahead. *Curr. Opin. Microbiol.* **36**, 85–94 (2017).

150. Wang, J. D. & Levin, P. A. Metabolism, cell growth and the bacterial cell cycle. *Nat. Rev. Microbiol.* **7**, 822–827 (2009).

151. Hirota, Y., Ryter, A. & Jacob, F. Thermosensitive mutants of E. coli affected in the processes of DNA synthesis and cellular division. in *Cold Spring Harbor symposia on quantitative biology* vol. 33 677–693 (Cold Spring Harbor Laboratory Press, 1968).

152. Mohammadi, T. *et al.* Identification of FtsW as a transporter of lipid-linked cell wall precursors across the membrane. *EMBO J.* **30**, 1425–1432 (2011).

153. Sham, L.-T. *et al.* MurJ is the flippase of lipid-linked precursors for peptidoglycan biogenesis. *Science* **345**, 220–222 (2014).

154. Kuk, A. C., Mashalidis, E. H. & Lee, S.-Y. Crystal structure of the MOP flippase MurJ in an inward-facing conformation. *Nat. Struct. Mol. Biol.* **24**, 171–176 (2017).

155. Bolla, J. R. *et al.* Direct observation of the influence of cardiolipin and antibiotics on lipid II binding to MurJ. *Nat. Chem.* **10**, 363 (2018).

156. Typas, A., Banzhaf, M., Gross, C. a. & Vollmer, W. From the regulation of peptidoglycan synthesis to bacterial growth and morphology. *Nat. Rev. Microbiol.* **10**, 123–136 (2011).

157. Ghuysen, J. M. & Hakenbeck, R. *Bacterial Cell Wall*. *Chemistry and Physics of Lipids* (1994). doi:10.1016/0009-3084(86)90009-5.

158. Lutkenhaus, J. Assembly dynamics of the bacterial MinCDE system and spatial regulation of the Z ring. *Annu Rev Biochem* **76**, 539–562 (2007).

159. Wu, L. J. & Errington, J. Nucleoid occlusion and bacterial cell division. *Nat. Rev. Microbiol.* **10**, 8–12 (2012).

160. Erickson, H. P. & Osawa, M. Cell division without FtsZ–a variety of redundant

mechanisms. *Mol. Microbiol.* **78**, 267 (2010).

161. Bernander, R. & Ettema, T. J. FtsZ-less cell division in archaea and bacteria. *Curr. Opin. Microbiol.* **13**, 747–752 (2010).

162. Pilhofer, M. *et al.* Discovery of chlamydial peptidoglycan reveals bacteria with murein sacculi but without FtsZ. *Nat. Commun.* **4**, 1–7 (2013).

163. Frandi, A., Jacquier, N., Théraulaz, L., Greub, G. & Viollier, P. H. FtsZ-independent septal recruitment and function of cell wall remodelling enzymes in chlamydial pathogens. (2014) doi:10.1038/ncomms5200.

164. Vicente, M., Gomez, M. J. & Ayala, J. a. Regulation of transcription of cell division genes in the Escherichia coli dcw cluster. **54**, 317–324 (1998).

165. Dubarry, N., Possoz, C. & Barre, F. Multiple regions along the Escherichia coli FtsK protein are implicated in cell division. **78**, 1088–1100 (2010).

166. Di Lallo, G., Fagioli, M., Barionovi, D., Ghelardini, P. & Paolozzi, L. Use of a two-hybrid assay to study the assembly of a complex multicomponent protein machinery: bacterial septosome differentiation. *Microbiology* **149**, 3353–3359 (2003).

167. Karimova, G., Dautin, N. & Ladant, D. Interaction Network among Escherichia coli Membrane Proteins Involved in Cell Division as Revealed by Bacterial Two-Hybrid Analysis. **187**, 2233–2243 (2005).

168. Grenga, L., Luzi, G., Paolozzi, L. & Ghelardini, P. The Escherichia coli FtsK functional domains involved in its interaction with its divisome protein partners. (2008) doi:10.1111/j.1574-6968.2008.01317.x.

169. Natale, P. & Vicente, M. Bacterial Cell Division. in *eLS* 1–9 (American Cancer Society, 2020). doi:10.1002/9780470015902.a0000294.pub3.

170. Margolin, W. FTSZ AND THE DIVISION OF PROKARYOTIC CELLS AND ORGANELLES. *Nat. Rev. Mol. Cell Biol.* **6**, 862–871 (2005).

171. Ent, F. V. D. *et al.* Structural and mutational analysis of the cell division protein FtsQ. *Mol. Microbiol.* **68**, 110–123 (2008).

172. Ikeda, M. *et al.* Structural similarity among Escherichia coli FtsW and RodA proteins and

Bacillus subtilis SpoVE protein, which function in cell division, cell elongation, and spore formation, respectively. *J. Bacteriol.* **171**, 6375–6378 (1989).

173. Barreteau, H. & Kovac, A. Cytoplasmic steps of peptidoglycan biosynthesis. **32**, 168–207 (2008).

174. Vollmer, W., Blanot, D. & De Pedro, M. A. Peptidoglycan structure and architecture. *FEMS Microbiol. Rev.* **32**, 149–167 (2008).

175. Höltje, J. V. Growth of the stress-bearing and shape-maintaining murein sacculus of Escherichia coli. *Microbiol. Mol. Biol. Rev. MMBR* **62**, 181–203 (1998).

176. Ghuysen, J. M. Use of bacteriolytic enzymes in determination of wall structure and their role in cell metabolism. *Bacteriol. Rev.* **32**, 425–464 (1968).

177. Egan, A. J. F., Errington, J. & Vollmer, W. Regulation of peptidoglycan synthesis and remodelling. *Nat. Rev. Microbiol.* **18**, 446–460 (2020).

178. Meeske, A. J. *et al.* MurJ and a novel lipid II flippase are required for cell wall biogenesis in Bacillus subtilis. *Proc. Natl. Acad. Sci.* **112**, 6437–6442 (2015).

179. Mingorance, J. & Tamames, J. The bacterial dcw gene cluster: an island in the genome? *Mol. Time Space* 249–271 (2004).

180. Razin, S. Peculiar properties of mycoplasmas: the smallest self-replicating prokaryotes. *FEMS Microbiol. Lett.* **100**, 423–431 (1992).

181. Morowitz, H. J. & Wallace, D. C. Genome size and life cycle of the Mycoplasma. 62–73 (1973).

182. Woese, C. R. Bacterial Evolution. **51**, 221–271 (1987).

183. Woese, C. R. *Prokaryote systematics: The evolution of a science. in '"The Prokaryotes"'(A. Balows, HG Tr. uper, M. Dworkin, W. Harder, and KH Schleifer, Eds.)*. (Springer-Verlag, New York, 1992).

184. Hara, H., Yasuda, S., Horiuchi, K. & Park, J. T. A promoter for the first nine genes of the Escherichia coli mra cluster of cell division and cell envelope biosynthesis genes, including ftsI and ftsW. *J. Bacteriol.* **179**, 5802–5811 (1997).

185. Eraso, J. M. *et al.* The highly conserved MraZ protein is a transcriptional regulator in

Escherichia coli. *J. Bacteriol.* **196**, 2053–2066 (2014).

186. Kimura, S. & Suzuki, T. Fine-tuning of the ribosomal decoding center by conserved methyl-modifications in the Escherichia coli 16S rRNA. *Nucleic Acids Res.* **38**, 1341–1352 (2010).

187. Yakhnina, A. A. & Bernhardt, T. G. The Tol-Pal system is required for peptidoglycan-cleaving enzymes to complete bacterial cell division. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 6777–6783 (2020).

188. Hagan, C. L., Silhavy, T. J. & Kahne, D. β -Barrel Membrane Protein Assembly by the Bam Complex. (2011) doi:10.1146/annurev-biochem-061408-144611.

189. Silhavy, T. J., Kahne, D. & Walker, S. The Bacterial Cell Envelope. **2**, 16 (2010).

190. Collin, S., Guilvout, I., Nickerson, N. N. & Pugsley, A. P. Sorting of an integral outer membrane protein via the lipoprotein-specific Lol pathway and a dedicated lipoprotein pilotin. **80**, 655–665 (2011).

191. Konovalova, A., Silhavy, T. J. & Silhavy, T. J. Outer membrane lipoprotein biogenesis : Lol is not the end. (2015).

192. Bowyer, A., Baardsnes, J., Ajamian, E., Zhang, L. & Cygler, M. Characterization of interactions between LPS transport proteins of the Lpt system. *Biochem. Biophys. Res. Commun.* **404**, 1093–1098 (2011).

193. Suits, M. D. L., Sperandeo, P., Dehò, G., Polissi, A. & Jia, Z. Novel Structure of the Conserved Gram-Negative Lipopolysaccharide Transport Protein A and Mutagenesis Analysis. *J. Mol. Biol.* **380**, 476–488 (2008).

194. Bayer, M. E. Zones of membrane adhesion in the cryofixed envelope of Escherichia coli. *J. Struct. Biol.* **107**, 268–280 (1991).

195. Sturgis, J. N. Organisation and evolution of the tol-pal gene cluster. *J. Mol. Microbiol. Biotechnol.* **3**, 113–122 (2001).

196. Walburger, A., Lazdunski, C. & Corda, Y. The Tol / Pal system function requires an interaction between the C-terminal domain of TolA and the N- terminal domain of TolB. **44**, 695–708 (2002).

197. Gerding, M. A., Ogata, Y., Pecora, N. D., Niki, H. & De Boer, P. A. The trans-envelope

Tol–Pal complex is part of the cell division machinery and required for proper outer-membrane invagination during cell constriction in E. coli. *Mol. Microbiol.* **63**, 1008–1025 (2007).

5 Results

ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies

Raphaël R. Léonard[1,2], Marie Leleu[2,3], Mick Van Vlierberghe[2], Luc Cornet[2,4], Frédéric Kerff[1] and Denis Baurain[2]

[1] InBioS – Centre d'Ingénierie des Protéines, Université de Liège, Liège, Belgium
[2] InBioS –PhytoSYSTEMS, Eukaryotic Phylogenomics, Université de Liège, Liège, Belgium
[3] UGSF –Unité de Glycobiologie Structurale et Fonctionnelle, Université de Lille/CNRS, Lille, France
[4] Mycology and Aerobiology, Sciensano, Service Public Fédéral, Bruxelles, Belgium

ABSTRACT

TQMD is a tool for high-performance computing clusters which downloads, stores and produces lists of dereplicated prokaryotic genomes. It has been developed to counter the ever-growing number of prokaryotic genomes and their uneven taxonomic distribution. It is based on word-based alignment-free methods (*k*-mers), an iterative single-linkage approach and a divide-and-conquer strategy to remain both efficient and scalable. We studied the performance of TQMD by verifying the influence of its parameters and heuristics on the clustering outcome. We further compared TQMD to two other dereplication tools (dRep and Assembly-Dereplicator). Our results showed that TQMD is primarily optimized to dereplicate at higher taxonomic levels (phylum/class), as opposed to the other dereplication tools, but also works at lower taxonomic levels (species/strain) like the other dereplication tools. TQMD is available from source and as a Singularity container at [https://bitbucket.org/phylogeno/tqmd].

Subjects Bioinformatics, Genomics, Microbiology, Taxonomy
Keywords Dereplication, Prokaryotes, Genome quality, Genome selection, Alignment-free methods, Phylogenomics, NCBI RefSeq, Singularity, Metagenomics

Submitted 24 November 2020
Accepted 4 April 2021
Published 5 May 2021

Corresponding author
Denis Baurain,
denis.baurain@uliege.be

Academic editor
Alexander Bolshoy

Additional Information and Declarations can be found on page 24

DOI **10.7717/peerj.11348**

INTRODUCTION

The fast-growing number of available prokaryotic genomes, along with their uneven taxonomic distribution, is a problem when trying to assemble high-quality yet broadly sampled genome sets for phylogenomics and comparative genomics. Indeed, most of the new genomes belong to the same subset of hyper-sampled phyla, such as Proteobacteria and Firmicutes, or even to single species, such as *Escherichia coli* (e.g., 105,081 out of 939,798 genomes in GenBank as of January 2021), while the continuous flow of newly discovered phyla prompts for regular updates of in-house databases. This situation makes it difficult to maintain sets of representative genomes combining lesser known phyla, for which only few species are available, and sound subsets of highly abundant phyla. An automated straightforward method is required but would be far too slow if based on regular alignment algorithms.

How to cite this article Léonard RR, Leleu M, Van Vlierberghe M, Cornet L, Kerff F, Baurain D.. 2021. ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies. *PeerJ* **9**:e11348 http://doi.org/10.7717/peerj.11348

Alignment-free methods are quantifiable ways of comparing the similarity of sequences without using an alignment (*Zielezinski et al., 2017*). They have several advantages over alignment-based methods: they are computationally less expensive, they are resistant to gene shuffling and recombination events, and they do not depend on assumptions about sequence changes. In the review of (*Zielezinski et al., 2017*), two main categories of methods are described: the information theory-based methods and the word-based methods. The rationale behind word-based methods is that similar sequences share a similar set of words. Sequence words are called k-mers and can be defined as all the words, of a given size k, that one can enumerate for a given alphabet. The idea is to compare the "dictionaries" of the words observed in two different genomes. The more similar two genomes are, the more words their respective "dictionaries" will have in common. In contrast, information theory-based methods compute the amount of information shared between two analyzed (genomic) sequences. Several different ways to assess this quantity do exist (e.g., through data compression) but they are not the subject of this paper (see *Shannon, 1948*; *Kullback & Leibler, 1951*; *Kolmogorov, 1965*; *Tribus & McIrvine, 1971*; *Batista et al., 2011*; *Zielezinski et al., 2017* for details).

Based on the review on the alignment-free sequence comparison methods of (*Zielezinski et al., 2017*), two main categories of software packages were theoretically suitable for dereplicating prokaryotic genomes: the species identification/taxonomic profiling programs (Table 1 in *Zielezinski et al., 2017*) and the whole-genome phylogeny programs (Table 2 in *Zielezinski et al., 2017*). First, we did not investigate software solutions made available as web services because of their intrinsic limitation with respect to the amount of genomic data that one regular user can process through these interfaces. Second, all the programs belonging to the taxonomic profiling category required a reference database to compare the genomes to, which would have led us to a circular conundrum, in which a (possibly handmade) database of reference genomes is required to (automatically) build a database of representative genomes. Third, all those presented in the whole-genome phylogeny category were either not suited for large-scale dereplication or did not provide small enough running time estimates for their test cases. For example, jD2Stat (*Chan et al., 2014*) gives results for 5000 sequences of 1500 nucleotides in 14 min, which would clearly make computationally intractable the dereplication of hundreds of thousands of whole prokaryotic genomes. As of January 2021, we only found two programs that were made to dereplicate genomes, dRep (*Olm et al., 2017*) and Assembly-Dereplicator (*Wick & Holt, 2019*) . These two programs are presented below.

Considering the limitations of the existing tools for assembling representative sets of prokaryotic genomes, the present article describes our own program called "ToRQuEMaDA" (abbreviated TQMD in the following for convenience) for Tool for Retrieving Queried Eubacteria, Metadata and Dereplicating Assemblies. TQMD is a word-based alignment-free dereplicating tool for both public and private prokaryotic genomes designed for both high-performance computing (HPC) clusters and powerful single-node computing servers. TQMD is available on Bitbucket and can be installed on any HPC with SGE/OGE (Sun/Open Grid Engine) installed as a scheduler. Few modifications are needed to adapt the scripts to most local setups. A Singularity (*Kurtzer, Sochat & Bauer, 2017*)

container is also available for single-node computers without a scheduler. TQMD works both in parallel and iteratively. Using default parameter values, each elemental job takes two to three hours to complete (see Materials and Methods for test hardware specifications), and if enough CPUs are available to run all jobs of a given round at the same time, such a round should only take two to three hours. Usually, four to five rounds are sufficient to achieve the dereplication. Therefore, a single run of TQMD against ~60,000 Bacteria in NCBI RefSeq takes 8 to 15 h to complete.

MATERIALS AND METHODS

Hardware

Almost all the computational work was performed on a grid computer IBM/Lenovo Flex System composed of one big computing node (x440) and nine smaller computing nodes (x240), featuring a total of 196 physical cores, 2.5 TB of RAM and 160 TB of shared mass storage, and operating under CentOS 6.6. Beyond "bignode" (running the scheduler and the MySQL database; see below), only four of the smaller computing nodes were used when testing TQMD; their specifications are as follow: 2 CPUs Intel Xeon E5-2670 (8 cores at 2.6 GHz with Hyper-Threading enabled), 128 GB of RAM. For the dRep test (see below), we had to use a desktop workstation (Ubuntu Linux 16.04) featuring 2 CPUs Intel Xeon E5-2620 v4 (8 cores at 2.1 GHz with Hyper-Threading enabled) and 64 GB of RAM. Based on the comparator found on the website http://cpubenchmark.net/, the CPUs in our cluster and in the workstation were roughly equivalent (from −0.5 to +5% difference).

It is important to mention that due to Hyper-Threading configuration of the grid computer and the fact that several teams shared the infrastructure, queueing time and disk usage could not be strictly controlled during the tests. Therefore, all running times provided in this article are informed estimates rather than exact measurements. These estimates are those we would communicate to a user inquiring about the waiting time for a specific analysis to complete. They are an approximation of the running time recorded when the grid computer usage is low (i.e., almost no other user).

Software architecture

TQMD is composed of a database and includes two main phases: (1) a periodic preparation phase in which newly available genome assemblies ("genomes" for short) are downloaded (or locally imported for private genomes) and individual genome metrics are computed, and (2) an "on-demand" dereplicating phase in which genomes (both new and old) are dereplicated on the fly to provide a list of high-quality representative genomes as a result (Fig. 1). The database stores the paths to the individual genome (FASTA) files, the individual genome metrics and the list of representative genomes produced by each TQMD run. Each piece of data is computed independently; if a dereplication request is issued during the computation of newly available genomes, TQMD only uses the genomes for which all the data is available in the database. Moreover, it is fully aware of the organisms (NCBI) taxonomy (*Federhen, 2012*), which means that taxonomic filters can be applied when downloading and/or when clustering genomes to spare time and/or focus on taxa of interest.

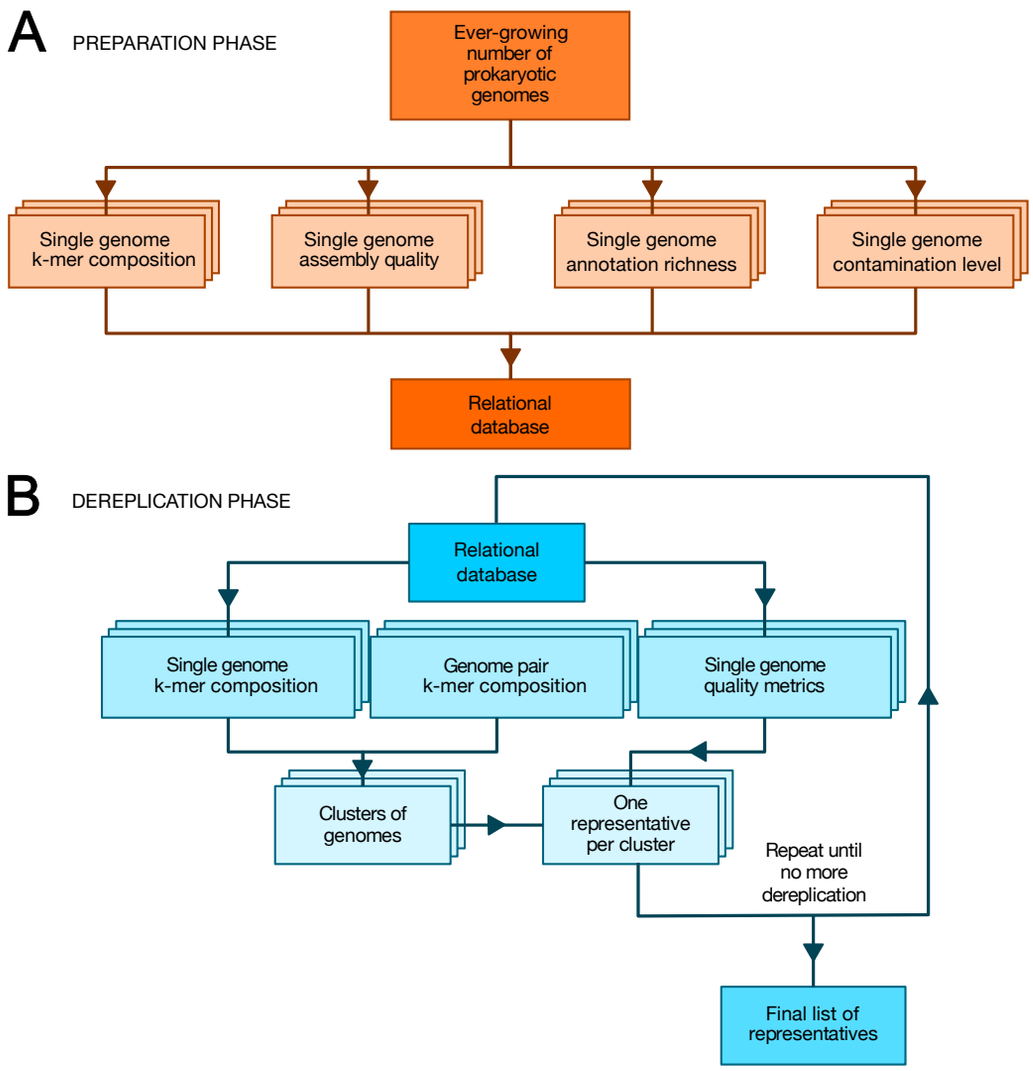


Figure 1 **Overview of TQMD phases and heuristics.** (A) The preparation phase consists in downloading newly released prokaryotic genomes from NCBI RefSeq and to pre-compute all per-genome information required to run the second phase: k-mer composition, assembly quality, annotation richness, contamination level (and completeness) level. Pre-computed information for single genomes is stored in a relational database associated with TQMD. (B) The dereplication phase then retrieves this information for all genomes to dereplicate from the database and clusters the genomes from pairwise distances computed on the fly. Cluster representatives (one per cluster created) are chosen for each cluster based on the single-genome metrics computed during the preparation phase. The dereplication is iterative and the process repeats until representative genomes cannot be dereplicated anymore, which produces the final list of representatives. Parallelized steps are shown as overlaid boxes.

Full-size DOI: 10.7717/peerj.11348/fig-1

During the preparation phase, we download the genomes and proteomes and pre-compute all the data required by the dereplication phase to store them in the database: indexes of nucleotide *k*-mers for single genomes, genome assembly quality metrics, genome annotation richness metrics, Small Subunit ribosomal RNA (SSU (16S) rRNA) predictions, contamination level and completeness level, whereas during the dereplication phase, we

cluster the genomes based on these *k*-mer indexes and select a representative for each cluster based on a user-modifiable ranking formula taking into account assembly quality, annotation richness, contamination and completeness level. These criteria were chosen, so as to select the best representative genomes (*Bowers et al., 2017*). By that, we mean that representative genomes (if available) are expected to be fully sequenced, correctly assembled, richly annotated and devoid of contaminating sequences. To satisfy this last requirement, TQMD can also use optional contamination statistics produced by Forty-Two (*Irisarri et al., 2017*; *Simion et al., 2017*) and/or CheckM (*Parks et al., 2015*) (see below).

Preparation phase

As shown in Fig. 1A, we first download the prokaryotic genomes from NCBI RefSeq (*O'Leary et al., 2016*) (or from GenBank (*Sayers et al., 2020*)). For the sake of data traceability, TQMD never gets rid of older genomes; newly released genomes are simply added to its internal database. The genomes from RefSeq and GenBank are kept physically separate. As TQMD was developed over five years, we have progressively accumulated several different versions of the RefSeq database, starting with release 79 (85,465 prokaryotic genomes, including 713 Archaea), then 79+92 (126,959 prokaryotic genomes, including 1,037 Archaea) and finally 79+92+203 (223,785 prokaryotic genomes, including 1,312 Archaea). Once RefSeq is up to date locally, we compute single-genome *k*-mer indexes and other metrics. For each of these computations, we use third-party programs and scripts (JELLYFISH, QUAST, RNAmmer, CD-HIT and Forty-Two or CheckM), except for the richness of the annotations, which we evaluate using an in-house script.

JELLYFISH (v1.1.12) (*Marçais & Kingsford, 2011*) is used to compute the *k*-mer indexes for single genomes (TQMD can also work with JELLYFISH v2.x and Mash (*Ondov et al., 2016*); see below). We tested several sizes for our *k*-mers. While JELLYFISH v1.x used to crash when using a size below 11 nucleotides, thus setting a hard lower bound on *k*-mer size, it is no longer an issue in JELLYFISH v2.x. On the other hand, while longer k-mers improve the specificity, they also require longer computing times (*Zielezinski et al., 2017*). With a size of 11, there are almost 4.2 millions (4^{11}) possible words. Consequently, a hypothetical genome featuring every possible *k*-mer without any repetition, could only be 4.2 Mbp long. Even if real genomes include repeats, genomes over 4 Mbp might still feature almost every *k*-mer, which would lead to useless *k*-mer indexes. To verify this idea, we examined the 85,465 genomes of RefSeq 79 and observed that about 15 genomes indeed almost exhaust the *k*-mer index (3 to 4 millions out of 4.2 millions), thus confirming that 11 is not a usable *k*-mer size. The next *k*-mer size, 12 nucleotides, offers over 16 millions (4^{12}) possibilities. The genomes with the largest index only reach 7.5 millions different *k*-mers, while the average index is below 2.7 millions *k*-mers. We could have used a *k*-mer size of 13 nucleotides, but our preliminary tests showed an important increase of the computing time. Whereas our tests with a *k*-mer size of 12 on all available Bacteria lasted between 8 and 15 h, depending on the distance threshold used (see below), our tests with a size of 13 required between 1 and 2 days to finish. Therefore, we chose to work with a *k*-mer size of 12 nucleotides. Above that, we would only have dereplicated closely related strains (i.e., belonging to the same species) due to a too high specificity (*Zielezinski et al., 2017*) and/or

the computing times would have become prohibitively long. Moreover, we did not use the "canonical" option for computing "strand-insensitive" *k*-mers with JELLYFISH (meant to be used on reads according to the manual) because we used RefSeq where the genomes are supposed to be fully assembled and thus gene orientation might be informative. If GenBank is used instead of RefSeq, it is highly recommended to enable the canonical option in TQMD due to the presence of genomes likely to be not assembled (still at the scaffold stage) or only very poorly assembled. Yet, one has to remember that canonical *k*-mers are twice less numerous for a given *k*-mer size than strand-specific *k*-mers, which might become an issue for distinguishing large genomes.

QUAST (v4.4) (*Gurevich et al., 2013*) is used to estimate the quality of genome assemblies (QUAST v5.x is also supported). We retrieve several quality metrics (13 in total) for each genome, these being the number of DNA sequences, the number of DNA sequences (or contigs) > 1 kbp, the size of the complete genome, the size of the complete genome composed of DNA sequences > 1 kbp, the number of contigs, the largest DNA sequence, the size of the complete genome composed of DNA sequences > 500 bp, the GC content, the N50, N75, L50 and L75 values, and the number of "N" per 100 kbp (N is the symbol used to scaffold contigs without matching ends). Given a minimal set of contigs ordered by descending length, the N50/N75 is defined as the length of the contig located at 50%/75% of the total genome length in the distribution, whereas the L50/L75 is defined as the rank of this specific contig. Among these metrics, we eventually decided to take into account (1) the relative length of the largest DNA sequence to the complete genome (> 1 kpb only) and (2) the fraction of "N" in the genome. In addition, we also use a size range (between 100 kbp and 15 Mbp) to remove the genomes too small to be complete and those too large to be considered uncontaminated (*Cornet et al., 2018b*).

For the richness of annotation, we compute what we call the "certainty" and the "completeness" of each genome. Importantly, this step necessitates (predicted) proteomes. While it is not an issue with RefSeq genomes, for which such predictions are always available, if TQMD is provided with an input genome set from a different source (GenBank or private genomes) with missing predicted proteomes, the related genomes will be automatically discarded (at least if the annotation metrics are used in the ranking formula). Our "certainty" metric corresponds to the proportion of sequences in a given proteome that we deem uncertain. To this end, we first count the number of sequence descriptions (in FASTA definition lines) with words indicating uncertainty, such as "probable", "hypothetical" or "unknown", then we compute a relative score as follows:

$$\text{Certainty} = 1 - \frac{\text{count of uncertain proteins}}{\text{total count of proteins}}$$

For "completeness", instead of counting the number of uncertain proteins, we count the number of proteins without any description:

$$\text{Completeness} = 1 - \frac{\text{count of unannotated proteins}}{\text{total count of proteins}}$$

Regarding genome contamination, RNAmmer (v1.2) (*Lagesen et al., 2007*) is used to predict the SSU (16S) rRNA of the genomes. By using cd-hit-est (v4.6) (*Li & Godzik, 2006*;

Fu et al., 2012) with an identity threshold of 97.5% (*Taton et al., 2003*), TQMD optionally creates a list of genomes featuring at least two SSU (16S) rRNA sequences belonging to different species (i.e., clustered in distinct CD-HIT clusters). This list of likely chimerical (or at least contaminated) genomes can be provided to filter out the genomes given as input to TQMD or produced in output by TQMD (see below). Another (more recent) possible threshold for species delineation based on SSU (16S) rRNA identity would be 99% (*Edgar, 2018*) and TQMD also supports such a setting.

Finally, another contamination metric is also available for the ranking: the genome contamination level estimated by the program Forty-Two (*Van Vlierberghe, 2021*) (v0.210570 or higher "42") based on the comparison of the genome ribosomal proteins to the reference sequences of the RiboDB database (*Jauffrit et al., 2016*). While this is the recommended approach for probing genome-wide contamination due to its speed, TQMD also supports CheckM (*Parks et al., 2015*) (v1.1.3) to predict "genome completeness" and "genome contamination". The contamination assessment of the latter is based on lineage-specific marker genes in addition to ribosomal proteins.

Once all these individual *k*-mer indexes and metrics are computed for all individual genomes, the genomes are ranked in a global ranking from the best to the worst genome (to be selected as a cluster representative), using an equal-weight sum-of-ranks approach available in the Perl module Statistics::Data::Rank. For each metric, a ranking is produced across all genomes and the final rank of a specific genome is computed as the sum of each of those individuals ranks without favoring one metric over another. For now, we do not consider all the metrics stored in the TQMD database, since all are optional and some are redundant. The five metrics (in TQMD syntax) used to compute the default ranking are: (1) assembly quality: quast.N.per.100.kbp; (2) assembly quality: quast.largest.contig.ratio (= quast.largest.contig / quast.total.len.1000.bp); (3) annotation richness: annot.certainty; (4) contamination level: 42.contam.perc; (5) contamination level: 42.added.ali. The first two metrics are obtained from QUAST, the third from our in-house script, and the fourth and fifth from "42". Finally, it is worth noting that TQMD allows the user to devise a custom ranking formula involving any combination of the 30 supported metrics (see details in TQMD manual).

Dereplication phase

Genome clustering can be carried out on the full set of genomes stored in the TQMD database or only on one or more taxonomic subsets of them. Moreover, both positive (inclusion and/or representativeness priority) and negative (exclusion) lists of GCA/GCF numbers can be provided to alter TQMD input and output genome sets. TQMD itself optionally produces such a negative list to exclude genomes featuring multiple SSU (16S) rRNA sequences (see above). Furthermore, both public (from RefSeq/GenBank) and private (i.e., custom) genomes can be dereplicated simultaneously. Moreover, the presence of at least one SSU (16S) rRNA predicted by RNAmmer can be used as a requirement for the genome to be selected, which would rule out some metagenome-assembled genomes (MAGs), for which rRNA genes are often missing (*Cornet et al., 2018a*). Consequently, this option is recommended when working with RefSeq but not GenBank, at least if the selection

of some lesser quality MAGs is important for the user. Regarding priority lists, they can be useful in comparative genomics, when one wants to include model organism genomes without sacrificing dereplication. As shown in Fig. 1B, the dereplication process is iterative and stops once it deems itself finished. Its decision is based on three different convergence criteria, for which we provide default threshold values but these can be modified by the user (see below). TQMD stops cycling as soon as one criterion is satisfied.

Two different distances can be used for clustering genomes with TQMD, each one derived from a distinct similarity metric, the Jaccard Index (JI; see (*Real & Vargas, 1996*)) or the Identical Genome Fraction (IGF; see (*Cornet et al., 2018b*)), both applied to shared *k*-mers at the nucleotide level. The effective distance used by TQMD is then obtained by subtracting the corresponding similarity metric from 1.

The JI is a measure of the similarity between two finite datasets. It is defined as the intersection over the union of the two datasets A and B:

$$\mathrm{JI}(\mathrm{A},\mathrm{B}) = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|}$$

The JI can be computed in two different manners: (1) exact computation using JELLYFISH (default option) and (2) approximate estimation using Mash (*Ondov et al., 2016*) (v1.1.1). If Mash is to be used, precomputation of single-genome *k*-mers is not required.

The IGF, for Identical Genome Fraction, replaces the union in the JI by the size of the smallest of the two datasets A and B:

$$\mathrm{IGF}(\mathrm{A},\mathrm{B}) = \frac{|\mathrm{A} \cap \mathrm{B}|}{\min(\mathrm{A},\mathrm{B})}$$

The TQMD algorithm works similarly for both distances and is inspired by the greedy clustering approach implemented in packages such as CD-HIT (*Jones, Pevzner & Pevzner, 2004*; *Li & Godzik, 2006*; *Fu et al., 2012*). The greedy clustering can work in two different modes, loose and strict. In both cases, we first sort the list of genomes based on the global ranking of the genomes (assembly quality and annotation richness metrics, indicators of genome contamination; see above for details) and the top-ranking genome is assigned to a first cluster. Then, in loose mode, every other genome is compared to every member of every cluster until it finds a suitable cluster of similar genomes; otherwise, such a genome becomes the first member of a new cluster. Hence, the second genome is compared to the single genome of the first cluster. If its distance to the latter genome is lower than specified threshold (let us say it is the case here), it is added to the cluster. Similarly, the third genome is compared to the first member of the first cluster; if its distance is higher than the threshold, it is compared to the second member of this first cluster. If it still is higher than the threshold, and since there is no other cluster, it creates a new (second) cluster. The fourth genome follows the same path, as will all remaining genomes do until every genome of the list is assigned to a cluster, whether singleton or part of a larger group. As genomes are processed from the best to the worst in terms of global ranking, representative genomes (which correspond to cluster founders) are automatically the best possible for each cluster. In strict mode, every other genome is only compared to the representative genome (here

too corresponding to the highest-ranking genome) of every cluster, which both speeds up the clustering process and mitigates the potential drawbacks of pure single-linkage.

To scale up the greedy clustering algorithm, we used a divide-and-conquer approach (*Bentley, 1980*; *Jones, Pevzner & Pevzner, 2004*) (Fig. 2). Indeed, when performing our own tests, we worked with about 112,000 genomes, a number making clearly impossible to compare all genomes at once. Therefore, we first partitioned the list of genomes into smaller batches (hereafter termed "packs") of 200 by default, either based on their advertised (NCBI) taxonomy (*Federhen, 2012*; *Sayers et al., 2020*) or completely at random. The clustering of each small pack yields a single representative, which we regroup into a new (shorter) list of genomes that is processed iteratively following the same algorithm. In the next round, only the selected representatives are compared between each other, thereby precluding the genomes that were not selected to be directly compared. While this heuristic results in an important speed-up, it may also prevent similar genomes to be mutually dereplicated because they were processed in distinct packs and replaced by representatives that are potentially less similar. The iterative algorithm stops based on any of the following three criteria (which can be specified by the user): (1) if it reaches a maximum number of rounds, (2) if it falls below an upper limit for the number of representatives (i.e., number of clusters) or (3) if the clustering ratio between two successive rounds falls below a minimum threshold. We define the clustering ratio as the percentage of genomes dereplicated at the end of a TQMD round compared to the number of genomes still in the game at the beginning of the round.

Phylogenomic analyses

We used TQMD runs as a source of representative bacterial genomes and obtained selections containing between 20 and 50 organisms for the six most populated phyla (the upper limit for the number of representatives was set to 50). We also generated two other selections to sample all Bacteria at once, one containing 49 organisms and the other 151. A last selection of Archaea was also produced and contained 86 organisms. For each TQMD run, we retrieved the proteomes of the selected representatives and used Forty-Two to retrieve their ribosomal proteins. Those proteins were taxonomically labelled by computing the last common ancestor of their closest relatives (best BLAST hits) in the corresponding alignments (excluding self-matches), provided they had a bit-score $\geq$80 and were within 99% of the bit-score of the first hit (MEGAN-like algorithm (*Cornet et al., 2018b*)). Thus, this strategy allowed us to simultaneously assess the completeness and the contamination level of each representative proteome while providing widely sampled ribosomal proteins for phylogenomic analyses (Table 1). For the bacterial dataset (B), the largest of the nine TQMD selections, this step took less than three hours to complete.

For each TQMD run, we assembled a supermatrix from the ribosomal proteins retrieved earlier (Table 1). Briefly, sequences were aligned with MAFFT v7.453 (*Katoh & Standley, 2013*), then the alignments were cleaned using ali2phylip.pl from the Bio::MUST::Core software package (D. Baurain, https://metacpan.org/release/Bio-MUST-Core), which implements the BMGE (*Criscuolo & Gribaldo, 2010*) filter (min=0.3, max=0.5, bmge-mask=loose). This step reduced the proportion of missing sites in the alignments. Next,

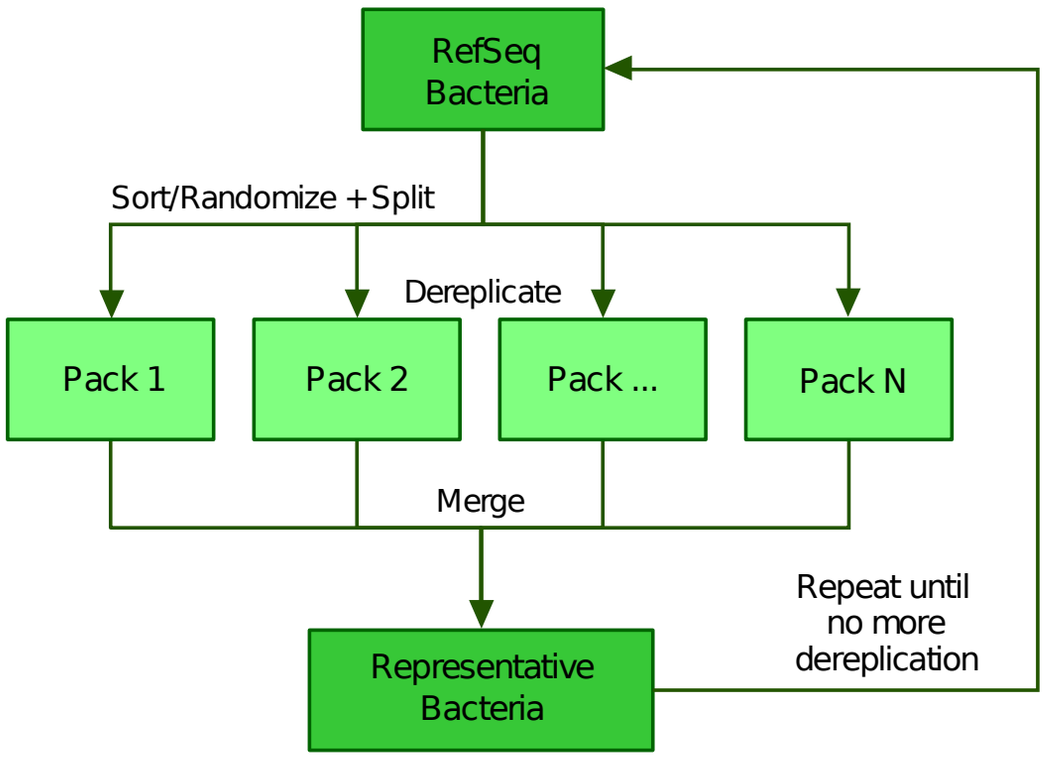


Figure 2 Illustration of the divide and conquer strategy of the dereplication phase. From a list of Bacteria downloaded from RefSeq (or GenBank), TQMD either sorts (based on the NCBI taxonomic lineage of each genome) or randomizes the list and splits it into packs of a given size. This allows each pack to be separately dereplicated, especially in parallel. Then all resulting lists of representative genomes are merged back and TQMD decides if it can stop or must refeed the merged list for another round.
Full-size DOI: 10.7717/peerj.11348/fig-2

we used Scafos v1.30k (*Roure, Rodriguez-Ezpeleta & Philippe, 2007*) to create the nine different supermatrices, using the Minimal evolutionary distance as a criterion for choosing sequences, the threshold set at 25%, the maximal percent of missing sites for a "complete sequence" set to 10 and the maximum number of missing OTUs set to 25, except for Firmicutes (22). Finally, IQ-TREE (*Nguyen et al., 2015*; *Hoang et al., 2018*) was used to infer the phylogenomic tree associated with each supermatrix, using the LG4X model with ultrafast bootstraps. Trees were automatically annotated and colored using format-tree.pl (also from Bio::MUST::Core) and then visualised with iTOL v4 (*Letunic & Bork, 2019*). The whole pipeline, from the launch of TQMD to the tree produced by IQ-TREE required approximately 3 working days for the larger bacterial selection (Table 1, line B).

RESULTS AND DISCUSSION

The TQMD workflow has two separate phases: a preparation phase (Fig. 1A) and a dereplication phase (Fig. 1B). The objective of the preparation phase is to compute the genome-specific data that will be needed during the dereplication phase. These operations are embarrassingly parallel and very easy to speed up. In contrast, the dereplication

Table 1 Details of TQMD runs and phylogenomic datasets built on eight different subsets of Bacteria. For each dataset, TQMD was launched with the Jaccard Index as a distance, a pack size of 200, the loose clustering mode, and was allocated a maximum of 50 CPUs. Other parameters (direct or indirect strategy and distance threshold) are provided in the table, along with the total running time in CPU hours (h.CPU), the initial number of genomes (# starting), the number of representatives obtained (# repr.), the number of ribosomal protein alignments used in the supermatrix (# prot.), and the number of unambiguously aligned amino acids in the supermatrix (# AA). Further details (taxonomy and download links, Krona taxonomic plots, Forty-Two reports, supermatrices and trees) are available at https://doi.org/10.6084/m9.figshare.13238936.

Label	Dataset	Strategy	Threshold	h.CPU	# starting	# repr.	# prot.	# AA
A	Bacteria (49)	indirect	0.900	656	63,863	49	53	6338
B	Bacteria (151)	indirect	0.880	656	63,863	151	53	6187
C	Actinobacteria	direct	0.900	96	8859	20	51	6562
D	Bacteroidetes	direct	0.850	16	1225	37	49	6605
E	Chlamydia	direct	0.800	6	360	32	44	6131
F	Cyanobacteria	direct	0.800	8	428	46	48	6314
G	Firmicutes	direct	0.900	242	21,544	22	52	6536
H	Proteobacteria	direct	0.885	310	30,690	36	53	6471
I	Archaea	direct	0.850	8	432	86	57	7810

phase considers all genomes at once, with the aim of clustering similar genomes based on pairwise distances and selecting the best representative for each cluster. To achieve this in the presence of many genomes, TQMD resorts to a greedy iterative heuristic in which each round is parallelized through a divide-and-conquer approach. The two phases are interconnected by the means of a relational database (see 'Materials and Methods' for details). Hereafter, we study the effects of TQMD parameters and heuristics on its dereplication behavior, then we compare its performance to those of two similar solutions, dRep and Assembly Dereplicator and, finally, we provide some application examples in the field of prokaryotic phylogenomics.

Analysis of TQMD behavior, parameters and heuristics

The dereplication phase is governed by a number of parameters and heuristics. One important issue is the inter-genome distance, which can either be based on the well-known Jaccard index (JI) or the identical genome fraction (IGF; see Materials and Methods for details). The latter was developed in an attempt to handle the comparison of genome pairs in which one is either partial or strongly reduced due to streamlining evolution or metagenomic source (*Cornet et al., 2018a*). Whatever the selected distance, genomes that are less distant than a user-specified threshold will end up in the same cluster. This distance threshold is thus the main "knob" for controlling the aggressivity of TQMD dereplication: the higher the threshold the tighter the clustering. Another point to consider are TQMD heuristics and their parameterization. Since TQMD is iterative, one can always decide to dereplicate genomes that are themselves representatives obtained in one or more previous runs. When trying to dereplicate very large and taxonomically broad genome sets, this raises the possibility to "guide" the dereplication by first clustering several phylum-wide subsets before merging the selected representatives in a single dataset to be dereplicated once more. This "indirect strategy" is to be contrasted with the "direct strategy", in which TQMD is left

dealing with the whole dataset from the very beginning. Regarding the divide-and-conquer algorithm operating during a single round, four parameters might be relevant: the pack size (e.g., 200 to 500), the clustering mode (loose or strict) and the dividing scheme (random or taxonomically-guided). Obviously, larger pack sizes require more time to be processed but are less likely to be affected by the impossibility to dereplicate two genomes that are in different packs. The clustering mode will also influence the number of pairwise comparisons required and thus the time necessary to cluster the genomes within a pack. Finally, in an attempt to balance such negative effects and the clustering speed, genome packs can either be composed at random (random sort) or by preferentially grouping taxonomically related organisms (taxonomic sort).

Performance criteria

Before studying the behavior of TQMD under different sets of parameters and heuristics, one has to keep in mind that its aim is to generate dereplicated lists of genomes that maintain the phylogenetic diversity of the input genomes, especially at the highest levels of the prokaryotic taxonomy. Therefore, we identified two metrics of interest when examining TQMD output: (1) the number of phyla with at least one representative genome ("diversity") and (2) the taxonomic mixing amongst the clusters ("mixity"). The diversity can be put in perspective with the number of representatives using what we call a redundancy index, i.e., the number of representatives divided by the number of phyla, with the lower the better. Regarding the concept of taxonomic mixing, we use it when the group of genomes behind a representative genome is not taxonomically homogeneous at some specific taxonomic level. Since our objective is mostly to dereplicate at the phylum level, we checked the taxonomic mixing at the phylum level. For example, if within a group of Proteobacteria, one (or several) Firmicutes is present, then the group is considered "mixed".

Iterative algorithm: dereplication kinetics

We first compared the results of the two distance metrics (JI or IGF) on the full set of RefSeq Bacteria passing our quality control (see Materials and Methods). To study the effect of the distance threshold used for dereplication, we selected two ranges of six values giving similar final numbers of representatives for the two metrics (JI: from 0.8 to 0.9; IGF: from 0.6 to 0.7). Figure 3 shows the dereplication kinetics observed when using a medium threshold (JI: 0.84; IGF: 0.66) and the direct strategy. The extreme efficacy (i.e., clustering ratio; see Materials and Methods) of the first round of dereplication is clear and subsequent rounds reach a plateau almost immediately. Whereas there is no notable difference between the two metrics in terms of kinetics, the height of the plateaus are not the same, with the IGF distance appearing greedier than the JI distance, especially when considering represented phyla rather than representative genomes.

Iterative algorithm: effect of parameters and heuristics

While TQMD was designed to be run without manual intervention (direct strategy), it is also possible to funnel the process by feeding it taxonomically homogeneous subsets of representative genomes (indirect strategy). To contrast the two strategies, we first separated

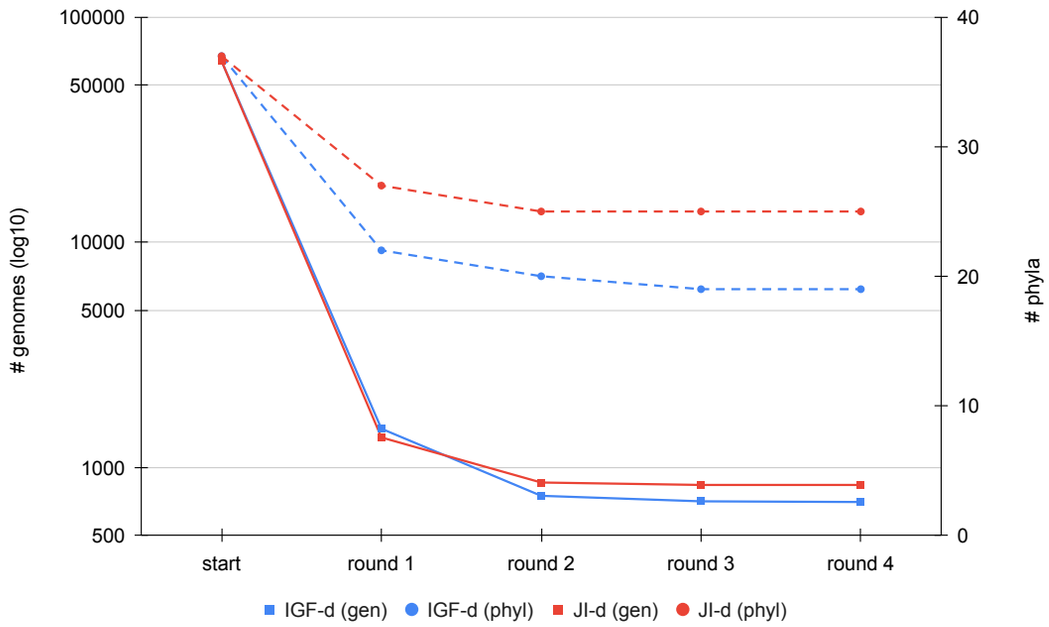


Figure 3 Comparison of the dereplication kinetics of TQMD when varying the distance metric. Two runs were launched on all RefSeq Bacteria (63,836 genomes; 37 phyla) using the direct strategy, a pack size of 200 and the loose clustering mode, one with the Jaccard Index (JI-d, distance threshold of 0.84, red curves) and one with the Identical Genome Fraction (IGF-d, distance threshold of 0.66, blue curves). The left *Y*-axis shows the log10 of the number of remaining genomes (square dots and solids lines), whereas the right *Y*-axis shows the number of phyla for which at least one representative is still present at a given round of dereplication (round dots and dashed lines).

Full-size DOI: 10.7717/peerj.11348/fig-3

Bacteria into five groups corresponding to the four largest phyla in terms of numbers of genomes available in RefSeq and a fifth group with the rest of Bacteria: Proteobacteria (39,011 genomes), Firmicutes (26,972 genomes), Actinobacteria (10,248 genomes), Bacteroidetes (1,639 genomes), other bacteria (2,682 genomes).Then we dereplicated the four phyla separately using the JI and a distance threshold of 0.8. Finally, we pooled the representatives obtained through the four TQMD runs with the remaining Bacteria and launched a final run on this reconstructed list. For this final run, we tried the two metrics and the full range of thresholds. The results of this multidimensional comparison are provided in Table 2 and Fig. 4.

Starting with an initial number of bacterial phyla equal to 37, it appears that the two JI strategies are better than any IGF strategy in terms of diversity, since the former retain a higher number of represented phyla for a given number of representative genomes. For example, when ending with about 500 representatives, the JI distance preserves 22–24 phyla, whereas the IGF distance only retains 15–19 phyla. These numbers translate to redundancy index (RI) values of 25–20 (JI) and 31–23 (IGF), respectively (Table 2). With the IGF distance, the indirect strategy appears better at all thresholds, with a number of represented phyla systematically higher for a number of representatives systematically lower. This translates to, e.g., RI = 50 (IGF-i) vs 65 (IGF-d) with about 1550 representatives and RI = 30 (IGF-i) vs 33 (IGF-d) for about 720 representatives. In contrast, this is less

Table 2 Comparison of the clustering properties when varying the distance metric, the distance threshold or the clustering strategy. Analyses were run on 63,863 RefSeq Bacteria using two different distance metrics, either based on the Jaccard Index (JI) or the Identical Genome Fraction (IGF), six different distance thresholds (from 0.8 to 0.9 and from 0.6 to 0.7, respectively), and two different clustering strategies, either direct (JI-D and IGF-D) or indirect (JI-i and IGF-i; see text for details). All pack sizes were 200 and the clustering mode was set to "loose". RI, Redundancy Index (# groups / # phyla).

Jaccard Index (JI)

Direct strategy (JI-d)							Indirect strategy (JI-i)						
threshold	0.80	0.82	0.84	0.86	0.88	0.90	threshold	0.80	0.82	0.84	0.86	0.88	0.90
RI	59	47	35	25	14	10	RI	54	46	35	20	12	4
# phyla	34	34	29	24	19	11	# phyla	34	31	25	22	13	11
# groups	2005	1589	1025	598	268	109	# groups	1845	1430	870	446	151	49
—pure groups	1992	1576	1009	587	261	106	– pure groups	1835	1416	853	434	149	45
– singletons	1201	904	557	325	143	56	– singletons	1727	818	488	242	88	24
— mixed groups	13	13	16	11	7	3	– mixed groups	10	14	17	12	2	4
– paraphyletic	0	0	0	0	0	0	– paraphyletic	0	1	0	0	0	0
– super-phyla	10	10	12	5	2	0	– super-phyla	10	13	9	7	0	1
– polyphyletic	3	3	4	6	5	3	– polyphyletic	0	0	8	5	2	3

Identical Genome Fraction (IGF)

Direct strategy (IGF-d)							Indirect strategy (IGF-i)						
threshold	0.60	0.62	0.64	0.66	0.68	0.70	threshold	0.60	0.62	0.64	0.66	0.68	0.70
RI	74	65	58	45	33	31	RI	50	55	44	30	23	11
# phyla	24	24	22	22	22	15	# phyla	31	25	24	24	19	16
# groups	1776	1548	1271	988	719	464	# groups	1536	1369	1061	715	440	176
—pure groups	1758	1530	1271	971	706	456	– pure groups	1514	1345	1042	701	426	167
– singletons	1094	939	755	587	419	260	– singletons	905	784	595	404	219	77
—mixed groups	18	18	19	17	13	8	– mixed groups	22	24	19	14	14	9
– paraphyletic	4	2	2	2	1	1	– paraphyletic	2	3	1	0	2	0
– super-phyla	11	11	13	10	4	1	– super-phyla	17	17	14	10	8	4
– polyphyletic	3	5	4	5	8	6	– polyphyletic	3	4	4	4	4	5

obvious with the JI distance, where the indirect strategy does not perform significantly better, the number of representatives also decreases but the number of represented phyla is also lower (or equal for the 0.9 threshold).

In the majority of the groups, the genome count per cluster is low with a significant proportion of singletons (i.e., only one representative genome, Table 2). However, in a few cases, large phyla (e.g., Proteobacteria, Firmicutes) gather into mixed groups that reach extreme genome counts and are visible as peaks in Fig. 4. Neither strategy changes this tendency but it is of notice that the JI distance with the indirect strategy is the combination leading to the lowest genome count per cluster and the lowest count of mixed groups (Table 2 and panel JI-i in Fig. 4), indicating a tendency to prevent the appearance of polyphyletic groups. When looking at the mixing (Table 2), it appears that unless at the highest thresholds, the mixity remains marginal in all strategies. To analyze the situation within the mixed groups, we separated them into three categories: (1) paraphyletic groups (only one case, Firmicutes and Tenericutes), (2) super-groups (e.g., FBC, PVC, Terrabacteria; see Fig. 5), and (3) polyphyletic groups. Since the TQMD objective is aggressive dereplication,

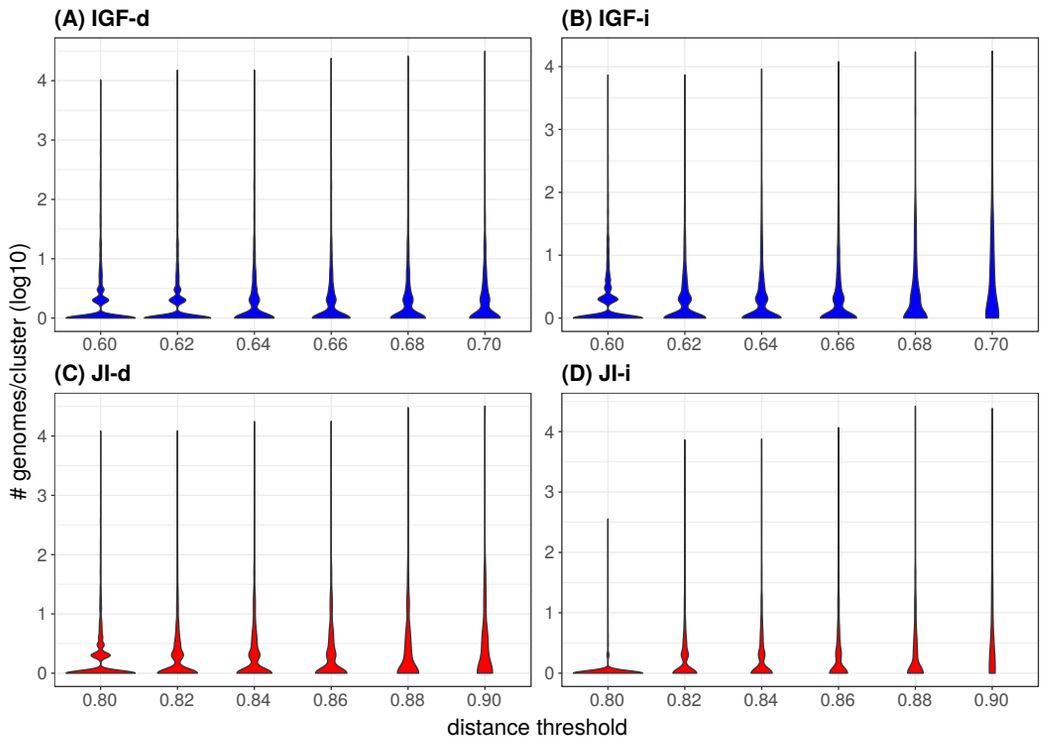


Figure 4 Distribution of the number of genomes per cluster when varying the distance metric, the distance threshold or the clustering strategy. (A) IGF-d, (B) IGF-i, (C) JI-d, (D) JI-i. These violin plots are a companion to Table 3 and abbreviations are as in the latter table. The *Y*-axes are in log10 units and the violin plot width is proportional to the number of clusters containing the given number of genomes.
Full-size DOI: 10.7717/peerj.11348/fig-4

the first two types of mixing are not problematic. Indeed they show that TQMD works as intended by first regrouping similar genomes together before regrouping the more dissimilar genomes. This also confirms that multiple scales of genuine phylogenetic signal lie in the nucleotide *k*-mers used in TQMD (*Wen et al., 2014*; *Allman, Rhodes & Sullivant, 2017*).

Amongst polyphyletic groups, the "early" groups, i.e., those that appear at lower thresholds (0.8 for JI and 0.6 for IGF), are (1) Firmicutes/Tenericutes clustered with Thermotogae and other thermophilic bacteria and (2) Terrabacteria clustered with Synergistetes. Thermotogae are likely mixed with Firmicutes due to their chimeric nature, Firmicutes being one of the main gene contributors (through lateral gene transfer, LGT) to Thermotogae (*Nesbø et al., 2009*; *Gupta & Bhandari, 2011*). At higher thresholds, Thermotogae attract the other thermophilic bacteria, leading to the formation of a polyphyletic group. This result is a consequence of our single-linkage approach, which reveals to be a weakness when it comes to chimeric organisms that can bridge unrelated bacterial genomes. It might be possible to alleviate this effect by using the strict clustering mode (see below). Regarding the clustering of Synergistetes with other Terrabacteria, when only a few genomes were available, Synergistetes were dispersed within two other phyla, Deferribacteres and Firmicutes (*Jumas-Bilak, Roudiere & Marchandin, 2009*). Nowadays,

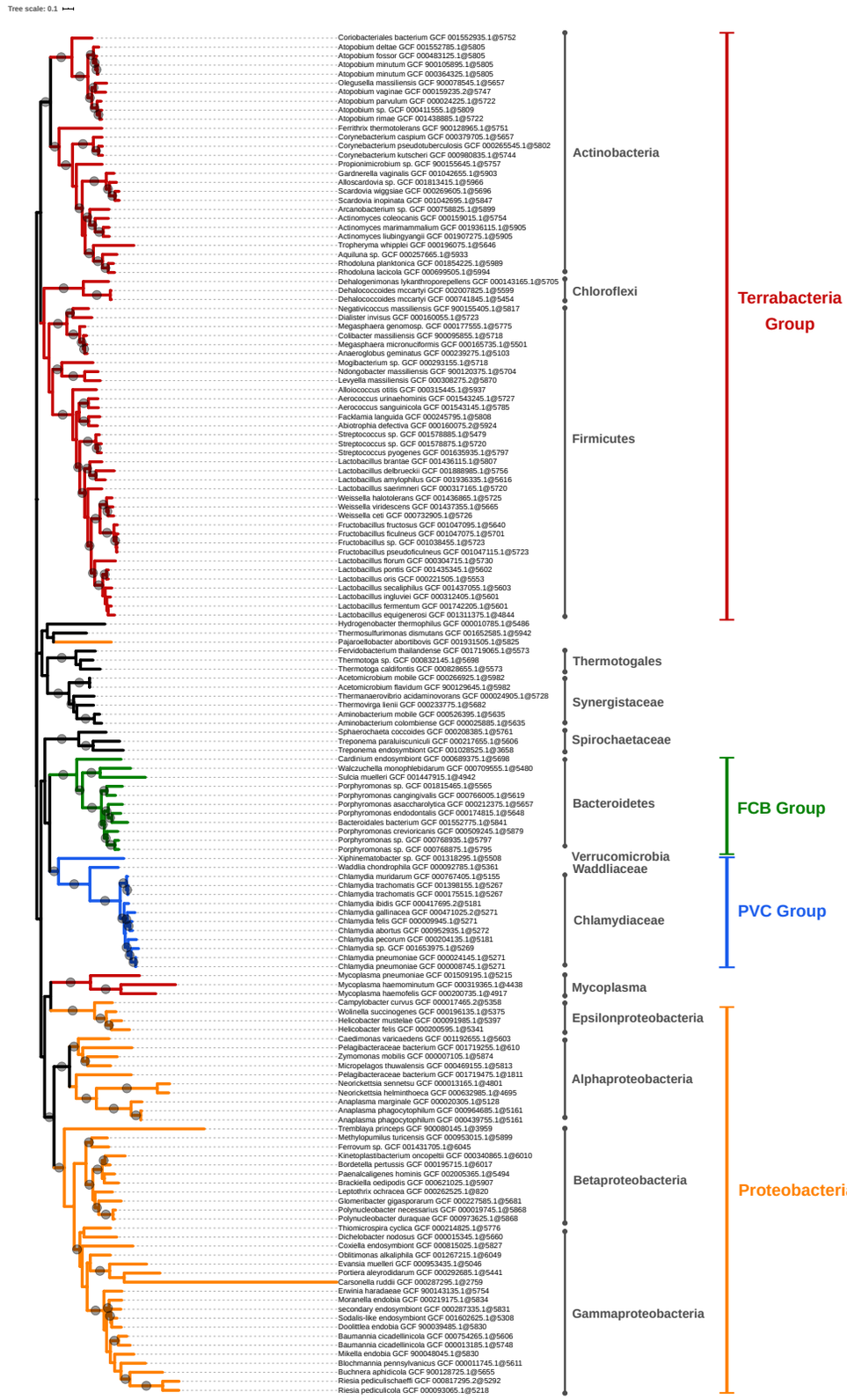


Figure 5 **Phylogenomic tree of the largest selection of Bacteria.** Tree inferred from a supermatrix of concatenated ribosomal proteins (Table 1B) under the LG4X model using IQ-TREE. Dots on branches indicate maximum bootstrap support values (100%).

Full-size DOI: 10.7717/peerj.11348/fig-5

Synergistetes form a monophyletic group that is sister to Deferribacteres (*Jumas-Bilak, Roudiere & Marchandin, 2009*). We hypothesize that conflicting (maybe artifactual) signals cause (at least some) Synergistetes to cluster with Firmicutes, and then to attract other Terrabacteria in a snow-ball effect due to single linkage. In other words, as the thresholds are increased, Thermotogae and Synergistetes serve as bridges between other bacterial phyla, creating or enlarging polyphyletic groups. This highlights that, just like alignment-based phylogeny, *k*-mer based approaches are also affected by chimeric organisms and LGT (*Daubin, Moran & Ochman, 2003*).

Divide-and-conquer algorithm: effect of parameters and heuristics

With respect to the parallelization of TQMD, the pack size has an influence on the results, since every time the size is diminished, the number of representatives returned at the end increases, whatever the distance metric (Table S1). This can be explained easily. In each pack, there is a list of genomes, to which each genome is compared in turn until it finds a cluster to join or creates a new cluster on its own. For each group, the selected representative is the best genome to work with in downstream applications, but not the "centroid" genome for the cluster. This means that a representative can be in the "outskirt" of its cluster in terms of sequence, which makes it less able to attract other genomes in subsequent groups. On the opposite, the single-linkage approach of the loose mode helps to alleviate the outskirt effect by enabling a genome to join a cluster as soon as any genome of that cluster is within the specified distance threshold. Another way to solve this issue is by increasing the pack size yet at the cost of speed. For example, 25 genomes require approximately 30 min to be processed, while 200 genomes take 2 h and 500 genomes take several days, which corresponds to a quadratic complexity.

The clustering mode (either loose and strict) also affects the clustering results. In Table 3, when compared to the corresponding (upper-left) part of Table 2, the effect of the strict mode on the number of representatives is obvious. As expected, they are more numerous than in loose mode since it becomes more difficult to cluster genomes together. Yet, if this effect is noticeable at the lower distance thresholds, it is barely noticeable at the higher thresholds. A second effect is that the polyphyletic groups of mixed genomes appear later (i.e., at higher thresholds) in strict mode than in loose mode.

Finally, TQMD tries to speed up the dereplication process by assembling packs following a taxonomic sort of the genomes to dereplicate. This heuristic should improve the clustering ratio of each iteration by directly comparing genomes that are more likely to be similar, thereby greatly reducing the required number of rounds of the whole process. As expected, five independent runs launched on all RefSeq Bacteria using JI-d (Table 4) with genomes sorted randomly returned selections of 904 representatives (on average) in 17 to 18 rounds whereas, the same run with genomes sorted according to taxonomy returned 836 representatives in only four rounds. Similarly, five runs using IGF-d with genomes sorted randomly yielded 456 representatives (on average) in 9 to 10 rounds, in contrast to 702 representatives in four rounds by enabling the taxonomic sort. However, when dereplicating subsets corresponding to Proteobacteria, the random dividing scheme returned less representatives (124, worst result) than the taxonomic dividing scheme (165),

Table 3 Effect of the strict clustering mode on the clustering properties when varying the distance threshold. Analyses were run on 63,863 RefSeq Bacteria using the Jaccard Index and the direct strategy (JI-d) with six different distance thresholds (from 0.8 to 0.9). All pack sizes were 200. RI, Redundancy Index (# groups / # phyla). This table has to be compared to the upper-left quarter of Table 2.

Thresholds	0.80	0.82	0.84	0.86	0.88	0.90
RI	66	49	37	24	15	8
# phyla	34	33	28	26	20	14
# groups	2231	1609	1035	614	300	112
- pure groups	2220	1592	1021	598	283	104
– singletons	1289	875	551	328	149	52
- mixed groups	11	17	14	16	17	8
– paraphyletic	1	0	0	0	2	0
– super phyla	10	16	10	10	11	4
– polyphyletic	0	1	4	6	4	4

in approximately the same number of rounds (3 to 5). Similar results were observed with Firmicutes: 224 representatives using the random scheme (worst result) vs 333 representatives using the taxonomic scheme. These results suggest that the random sort can be useful while working with a taxonomically homogeneous subset of bacteria. In other cases, it should be avoided because a higher number of rounds translates to a longer computing time.

A word about the genome source

In addition to RefSeq genomes, TQMD can also download and cluster GenBank genomes, along with (optional) custom genomes provided by the user. To test the effect of the source database, we studied the dereplication of RefSeq and GenBank Archaea (release 203), which have the advantage of combining a small number of genomes (941 and 4129 genomes, respectively) while featuring a lot of unclassified organisms, candidate phyla and metagenomic assemblies in GenBank (Table 5). Beyond the speed penalty due to sheer difference in the number of genomes, which influences the number of comparisons TQMD has to perform, switching to GenBank as the genome source also requires using canonical *k*-mers to account for the lesser assembly quality of many genomes (see Materials and Methods for details) and/or selecting Mash as the *k*-mer engine. Moreover, with GenBank, the diversity of representative genomes is expanded with candidate phyla, but at the cost of more unclassified genomes and also (meta)genomes of lesser assembly quality. Unclassified genomes are genomes without higher-level taxonomic taxa, which hinders the taxonomic sort heuristic and makes it harder for TQMD to dereplicate them (since they can start in packs distinct from those including the genomes they are the most similar to). Regarding genomes of lesser quality, some can act as a bridge between two clusters that should not be clustered together (as discussed above with the polyphyletic groups) if they are chimerical in any way (either genuinely or due to the mixing of different organisms). In the worst case, all genomes end up lumped together in a single large cluster (last row of Table 5). As our primary objective with TQMD was to provide high-quality representatives, we

Table 4 Comparison of the number of rounds and final representatives when modifying the distance metric and/or the dividing scheme for parallel processing. Five replicates of each combination were carried out for the random sort, whereas the taxonomic sort is deterministic. JI-based (direct) analyses were run using a distance threshold of 0.84, where IGF-based (direct) analyses used a threshold of 0.66. Pack size was 200 and the clustering mode was set to "loose".

Dataset	dist./appr.	sort	# rounds	# repr.
Bacteria	JI-d	taxonomic	4	836
Bacteria	JI-d	random	18	902
Bacteria	JI-d	random	17	903
Bacteria	JI-d	random	17	894
Bacteria	JI-d	random	18	915
Bacteria	JI-d	random	17	908
Bacteria	IGF-d	taxonomic	4	702
Bacteria	IGF-d	random	10	435
Bacteria	IGF-d	random	10	458
Bacteria	IGF-d	random	10	456
Bacteria	IGF-d	random	9	438
Bacteria	IGF-d	random	10	493
Proteobacteria	IGF-d	taxonomic	3	165
Proteobacteria	IGF-d	random	3	115
Proteobacteria	IGF-d	random	3	105
Proteobacteria	IGF-d	random	3	100
Proteobacteria	IGF-d	random	3	124
Proteobacteria	IGF-d	random	3	114
Firmicutes	IGF-d	taxonomic	4	333
Firmicutes	IGF-d	random	4	190
Firmicutes	IGF-d	random	5	212
Firmicutes	IGF-d	random	5	224
Firmicutes	IGF-d	random	4	194
Firmicutes	IGF-d	random	4	172

decided to focus this presentation on RefSeq, but Table 5 shows that TQMD also works with GenBank.

Comparison with dRep, assembly-dereplicator and mash

When we began our work on TQMD in 2015, there was no published program for genome dereplication. Now two different software packages are available, dRep (*Olm et al., 2017*) and Assembly-Dereplicator, both built on top of Mash (*Ondov et al., 2016*). Mash itself was created to estimate the Jaccard distance (derived from the JI) within sets of genomes and metagenome-assembled genomes (MAGs) based on nucleotide *k*-mer counts (*Ondov et al., 2016*). dRep was designed especially for the dereplication of MAGs, whereas Assembly-Dereplicator (A-D) was designed for groups of bacteria which are sufficiently close relatives. A comparison of the working principles and features of dRep, A-D and TQMD is available in Table 6.

To compare TQMD to dRep (v2.2.3), we chose two different datasets from RefSeq (release 79), the phylum Bacteroidetes (1127 genomes) and the order Streptomycetales

Table 5 Effect of the genome source (either RefSeq or GenBank) on clustering results using Archaea as a test case. The runs carried out on GenBank Archaea used canonical k-mers. The JI runs used a distance threshold of 0.90 and the IGF runs a threshold of 0.80. The super-phyla are the Asgard group, the TACK group and the DPANN group. Unclassified genomes are genomes without a phylum in the NCBI Taxonomy. JI: Jaccard Index; IGF: Identical Genome Fraction.

Source	# super-phyla	# phyla	# unclassified genomes	# genomes	Clustering mode
RefSeq	3	7	0	941	NA
GenBank	3	24	265	4129	NA
JI RefSeq	2	6	0	46	strict
JI RefSeq	2	6	0	29	loose
IGF RefSeq	2	6	0	38	strict
IGF RefSeq	1	3	0	16	loose
JI GenBank	3	17	38	313	strict
JI GenBank	3	15	18	145	loose
IGF GenBank	2	10	6	34	strict
IGF GenBank	1	1	0	1	loose

(648 genomes; phylum Actinobacteria). Because of technical difficulties with the installation of dRep, we had to use a workstation less powerful than the grid computer used to run TQMD (see 'Materials and Methods'). That is why we did not use all the available bacterial genomes in these tests. Regarding Bacteroidetes, dRep required five hours (using 10 CPUs and default parameters) to select 835 genomes. With TQMD, we used a threshold of 0.6 on the JI to obtain comparable results. TQMD run lasted 10 h (on at most 6 CPUs) and selected 789 representative genomes, of which 707 were in common with those of dRep. Since our main objective is to maintain as much as possible the diversity when dereplicating, we verified how many species were retained after the dereplication. Before dereplication, we had 528 different species of Bacteroidetes; dRep produced a list covering 516 of these species, whereas TQMD produced a list of 517 species, of which 511 were in common (see Table 7 for details). With Streptomycetales, dRep (again using default values), selected 430 genomes out of 648 in approximately 12h30min using 20 CPUs. To emulate such a result with TQMD, we had to use a threshold of 0.4 and obtained 486 representatives (392 in common, of which 175 species) in about 10 h using at most 4 CPUs in parallel (details given in Table 6).

dRep is a less aggressive program than TQMD, which is unsurprising as the former is meant to be used on sets of MAGs and to dereplicate at the species level, while the latter is meant to be used on every completely sequenced prokaryotic genome available and to dereplicate at the phyla/class level. Moreover, from the very start, TQMD was designed with scalability in mind, so as to accommodate the ever growing number of sequenced genomes. In principle, dRep could be used aggressively like TQMD, by fine-tuning two different thresholds (primary and secondary clusters), but this would need dRep to allow the user to choose a different Mash *k*-mer size, which does not appear to be possible (for the average user). On the other hand, TQMD can be used to dereplicate down to the species level more easily (only one threshold to specify) but it would take a longer time to

Table 6 **Feature comparison between dRep, Assembly-Dereplicator (A-D) and TQMD.**

Feature	dRep	A-D	TQMD
main engine(s)	Mash + ANIm (or gANI)	Mash	JELLYFISH or Mash
other dependencies	CheckM (optional)	none	QUAST (optional), RNAmmer (optional), CD-HIT-EST (optional), Forty-Two (optional), CheckM (optional)
relational database	N	N	Y
genome source	custom	custom	RefSeq, GenBank, custom
taxonomic filters	N	N	Y (when downloading and clustering)
automatic genome download	N	N	Y
distance metric(s)	Mash distance (estimated JI) then ANI	Mash distance (estimated JI)	1-JI (exact) or Mash distance (estimated JI) or 1-IGF (exact)
heuristic(s)	biphasic approach: Mash for fast and rough clustering followed by ANI for slow and accurate clustering	d-and-c strategy (serial)	iterative greedy algorithm (serial) + d-and-c strategy (parallel)
stop condition(s)	unspecified	first failure to dereplicate any serial batch	any of 3 possible cut-offs (number of rounds, number of representatives, clustering ratio)
d-and-c dividing scheme	unspecified	random	random or taxonomic
selection of representatives	formula based on genome size, assembly quality and contamination level (incl. strain heterogeneity)	assembly quality	formula based on genome size, assembly quality, annotation richness and contamination level (fully customisable with 30 possible metrics)
parameterization of representative selection	Y (parameter weights)	N	Y (simplified formula)
grid engine support	N	N	Y (SGE/OGE) (optional)
distribution	source (pip), conda, Galaxy	source	source (Bitbucket), Singularity container
CPU usage	fixed on launch	fixed on launch	specified as a maximum (decreases over time)

Notes.
JI, Jaccard Index; IGF, Identical Genome Fraction; ANI, average nucleotide identity; d-and-c, divide-and-conquer; SGE/OGE, Sun/Open Grid Engine; Y, present feature; N, absent feature.

Table 7 **Performance comparison between TQMD and dRep on two smaller datasets.** # gen, starting number of genomes; # repr, final/common number of representative genomes; # spec, starting/final/common number of species; h.CPU, upper bound on CPU use (i.e., product of wall-clock time and number of CPUs). With TQMD, a distance threshold of 0.6 was used for Bacteroidetes and a threshold of 0.4 for Streptomycetales. In both cases, the pack size was 200, the clustering mode was set to "loose" and the taxonomic sort was selected.

Dataset	Starting		TQMD - JELLYFISH k12			dRep			Intersection	
	# gen.	# spec.	# repr.	# spec.	h.CPU	# repr.	# spec.	h.CPU	# repr.	# spec.
Bacteroidetes	1,127	528	789	517	60	835	516	50	707	511
Streptomycetales	648	220	486	207	40	430	189	250	392	175

finish since it would require a longer JELLYFISH *k*-mer size (see Material and Methods). In conclusion, the dRep and TQMD can do each other's work but become less efficient when trying to do so, thereby rather making them complementary: dRep to dereplicate at the species level and TQMD at phylum/class level. For intermediate taxonomic levels, it is up to the user to decide which one s/he prefers. It is of note that, except for the centrality metric of dRep, the five other metrics used by dRep are available amongst the 30 metrics offered by TQMD and can be used through its customisable ranking formula (see Materials and Methods).

A-D is a program that is more recent but, as of April 2021, not yet published; its last update dates from November 2019. Its main advantage is ease of use, since it is a simple (no-installation) script that only needs Mash as a prerequisite. A–D takes as input the path to a folder containing the genomes to be dereplicated and rearranges them randomly and separated into smaller packs (500 genomes per pack by default). The next step is the clustering of each pack serially using Mash. A–D stops as soon as it cannot dereplicate at least one genome from the current pack. However, at least in our hands, A-D revealed to be unstable and/or to perform poorly on our test datasets (see Supplementary Materials for details).

TQMD allows the use of two different *k*-mer engines, JELLYFISH and Mash. With JELLYFISH, TQMD can compute a distance that is based on the exact JI (or the exact IGF), whereas with Mash, it relies on a distance based on the estimate of the JI. From the user perspective, this means that a given distance threshold will not produce exactly the same results depending on the active *k*-mer engine. We compared the results and run times of JELLYFISH and Mash using RefSeq Cyanobacteria (release 203) (Table 8). At an equivalent *k*-mer size (12), Mash is indeed faster than JELLYFISH (in both strict and loose clustering modes) and produces a similar number of clusters. The speed benefit provided by Mash approximation allows the use of larger *k*-mers, as illustrated by the results of a run based on a *k*-mer size of 16, whereas such a setup would be computationally intractable with JELLYFISH. Therefore, the integration of Mash as a *k*-mer engine makes TQMD competitive even while dereplicating on lower taxonomic levels. Finally, the relationship between the distance threshold and the Jaccard distance is not straightforward, notably depending on the size ratio between the two genomes under comparison. To help with the selection of an appropriate threshold when using JELLYFISH, we produced Fig. S9 as a guideline. For Mash, we refer the reader to *Ondov et al. (2016)*, who provide similar information in their Fig. S3 (and Eq. (4)).

Application example of TQMD

To check whether TQMD output was indeed useful in a practical context, we computed phylogenomic trees based on concatenations of ribosomal proteins sampled from selected representative genomes. We performed two runs on all RefSeq Bacteria (release 79; 63,863 genomes passing our prerequisites ; see Materials and Methods for details) using the indirect strategy and the JI, one at a distance threshold of 0.9 (Table 1, line A) and the other at 0.88 (Table 1, line B). The first run yielded a selection of 49 genomes while the second run retained 151 genomes. Seven additional runs using the direct strategy were

Table 8 Comparison of run time for JELLYFISH/Mash and strict/loose modes. All runs were carried out on RefSeq Cyanobacteria (918 genomes) using a distance threshold of 0.80 (JELLYFISH k12, 1-JI), 0.091 (Mash k12, Mash distance) and 0.069 (Mash k16, Mash distance). JI, Jaccard Index.

k-mer engine	Time		# representatives	
	Strict	Loose	Strict	Loose
Mash k12	0h56	1h44	73	49
Mash k16	11h19	13h15	550	529
JELLYFISH k12	3h14	7h30	73	52

carried out on the six largest bacterial phyla of RefSeq (in terms of numbers of organisms: Proteobacteria, Firmicutes, Actinobacteria, Bacteroidetes, Cyanobacteria and Chlamydia) and on Archaea. These phylum-wide selections contained about 20 to 50 genomes, each collectively representing the diversity of their respective phyla (Table 1, line C-H), whereas the archaeal selection contained 86 genomes (Table 1, line I). In this text, we only show and describe the larger phylogenomic tree of all Bacteria (Table 1, line B). The eight other trees are available as Figs. S1 to S8.

The larger bacterial tree (Fig. 5) results from an extremely aggressive selection (Table 1, B) but it still shows what we consider as the main groups of Bacteria (Proteobacteria, PVC, FBC, and "monoderm" phyla) and, after accounting for the idiosyncratic taxon names, most groups described by T. Cavalier-Smith (*Cavalier-smith & Chao, 2020*) are visible (with the exception of Eoglycobacteria and Hadobacteria, which were both absorbed in polyphyletic groups). Regarding the topology of the tree, all the organisms from the main super-phyla are generally regrouped in the same subtree, with some exceptions. These exceptions are the mycoplasma branch, which ends up within Proteobacteria, and *Pajaroellobacter abortibovis*, a proteobacterium that is separated from other Proteobacteria.

In Fig. 5, some genera and even species appear to be overrepresented in the selected genomes and form monophyletic subtrees within the tree. This is the case of *Lactobacillus*, for example, with 11 representatives (10 species). To investigate an eventual selection bias in TQMD, we launched two different TQMD runs using only the *Lactobacillus* genomes (841 which passed TQMD prerequisites). Both runs used the same values as the larger run for Bacteria (Table 1, B). The difference was the way of sorting the genomes before dividing them in packs, one used the taxonomic sort and the other the random sort. The run with the taxonomic sort yielded 19 *Lactobacillus* representatives (15 species), of which 10 in common with the larger run for Bacteria, whereas the random sort run yielded 21 representatives (16 species), of which 10 in common with the larger run for Bacteria and 16 with the taxonomic run. These results suggest that the taxonomic sort does not especially lead to a selection biased towards identically named genera or species, but that the representative genomes adequately sample the underlying phylogenetic diversity of the group. Along the same lines, dRep results for Bacteroidetes also show genomes of the same "species" not clustered together as in our Bacteroidetes tree (Table 6 and Fig. S3). This indicates that the genomes of such identically named organisms are actually quite different, thereby not reflecting a technical issue of TQMD or of dRep, but rather a genuine property of these genomes. Consequently, it is worth mentioning that a purely taxonomic (i.e., manual

based on NCBI Taxonomy) selection of representative genomes would have overlooked this genomic diversity, thereby reducing the relevance of the selection. In contrast, if the user is willing to accept a fixed number of representatives, a valuable alternative is to sample genomes from GTDB, since its taxonomy stems from ANI computations across RefSeq genomes, which is conceptually similar to what we dynamically do with TQMD. As for the current release (17/06/2020), GTDB features 111 "phyla" and 327 "classes" (*Parks et al., 2020*).

CONCLUSION

TQMD is an efficient dereplication tool initially designed for the assembly of phylum-level datasets of representative prokaryotic genomes. It manages to maintain the taxonomic diversity of input genomes while being fast, owing to its aggressive dereplication heuristics, which makes it able to scale with the ever growing number of genome assemblies in public repositories, such as NCBI RefSeq and GenBank. At lower taxonomic levels, TQMD becomes slower, probably because it has to compare more genomes before finding pairs close enough to be clustered and dereplicated. However, the use of the "strict" mode for the clustering can at least partially offset this effect. To dereplicate at the lowest taxonomic levels (species or strains), a longer *k*-mer would be better suited. While this is computationally intractable with the JELLYFISH engine, the support of the faster Mash engine makes it possible. The development of the first version of TQMD is now finished and highly benefited from the input of *PeerJ* reviewers. Yet, it could be further improved by adding new distance metrics beyond JI and IGF, and/or by including additional metrics for the selection of representative genomes. And now, with the Singularity container, TQMD can even be run on a single-node computer without a scheduler, making it easier to install and use.

ACKNOWLEDGEMENTS

The authors are grateful to Damien Sirjacobs for his support of the computing cluster and to Rosa Gago for her help with the design of the figures.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Raphaël R. Léonard and Mick Van Vlierberghe were supported by FRIA fellowships of the Belgian National Fund for Scientific Research (F.R.S.-FNRS). Marie Leleu is supported by the French Agence Nationale de la Recherche (ANR, project MATHTEST). Frédéric Kerff is a Research Associate employed by the F.R.S.-FNRS. Computational resources were provided through two grants to DB (University of Liège "Crédit de démarrage 2012" SFRD-12/04; F.R.S.-FNRS "Crédit de recherche 2014" CDR J.0080.15). This work (and Luc Cornet) was also supported by a research grant to DB (no. B2/191/P2/BCCM GEN-ERA) funded by the Belgian Science Policy Office (BELSPO). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Belgian National Fund for Scientific Research (F.R.S.-FNRS).
French Agence Nationale de la Recherche (ANR, project MATHTEST).
Belgian Science Policy Office (BELSPO): B2/191/P2/BCCM GEN-ERA.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Raphaël R. Léonard performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, developed the software ToRQuEMaDA, and approved the final draft.
- Marie Leleu performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Mick Van Vlierberghe performed the experiments, prepared figures and/or tables, and approved the final draft.
- Luc Cornet performed the experiments, authored or reviewed drafts of the paper, developed the Singularity container, and approved the final draft.
- Frédéric Kerff and Denis Baurain conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

TQMD software is available at Bitbucket: https://bitbucket.org/phylogeno/tqmd.

The datasets are available at figshare: Léonard, Raphaël R.; Leleu, Marie; Van Vlierberghe, Mick; Cornet, Luc; Kerff, Frédéric; BAURAIN, Denis (2020): Datasets for Léonard et al. ToRQuEMaDA: Tool for Retrieving Queried Eubacteria, Metadata and Dereplicating Assemblies. figshare. Dataset. https://doi.org/10.6084/m9.figshare.13238936.v2.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.11348#supplemental-information.

REFERENCES

Allman ES, Rhodes JA, Sullivant S. 2017. Statistically consistent k-mer methods for phylogenetic tree reconstruction. *Journal of Computational Biology* **24**:153–171 DOI 10.1089/cmb.2015.0216.

Batista MVA, Ferreira TAE, Freitas AC, Balbino VQ. 2011. An entropy-based approach for the identification of phylogenetically informative genomic regions of Papillomavirus. *Infection, Genetics and Evolution* **11**:2026–2033 DOI 10.1016/j.meegid.2011.09.013.

Bentley JL. 1980. Multidimensional divide-and-conquer. *Communications of the ACM* **23.4**:214–229 DOI 10.1145/358841.358850.

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**:725–731 DOI 10.1038/nbt.3893.

Cavalier-smith T, Chao EE. 2020. Multidomain ribosomal protein trees and the planctobacterial origin of neomura (eukaryotes, archaebacteria). *Protoplasma* **257**:621–753 DOI 10.1007/s00709-019-01442-7.

Chan CX, Bernard G, Poirion O, Hogan JM, Ragan MA. 2014. Inferring phylogenies of evolving sequences without multiple sequence alignment. *Scientific Reports* **4**6504.

Cornet L, Bertrand AR, Hanikenne M, Javaux EJ, Wilmotte A, Baurain D. 2018a. Metagenomic assembly of new (sub) polar Cyanobacteria and their associated microbiome from non-axenic cultures. *Microbial Genomics* **4**:e000212 DOI 10.1099/mgen.0.000212.

Cornet L, Meunier L, Van Vlierberghe M, Léonard RR, Durieu B, Lara Y, Misztak A, Sirjacobs D, Javaux EJ, Wilmotte A, Philippe H, Baurain D. 2018b. Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLOS ONE* **13.7**:e0200323 DOI 10.1371/journal.pone.0200323.

Criscuolo A, 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**:210 DOI 10.1186/1471-2148-10-210.

Daubin V, Moran NA, Ochman H. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science* **301**:829–832 DOI 10.1126/science.1086568.

Edgar RC. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**:2371–2375 DOI 10.1093/bioinformatics/bty113.

Federhen S. 2012. The NCBI taxonomy database. *Nucleic Acids Research* **40**:D136–D143 DOI 10.1093/nar/gkr1178.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**:3150–3152 DOI 10.1093/bioinformatics/bts565.

Gupta RS, Bhandari V. 2011. Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups. 1–34 DOI 10.1007/s10482-011-9576-z.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**:1072–1075 DOI 10.1093/bioinformatics/btt086.

Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35**:518–522 DOI 10.1093/molbev/msx281.

Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire J-Y, Kupfer A, Petersen J, Jarek M, Meyer A, Vences M. 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nature Ecology & Evolution* **1**:1370–1378 DOI 10.1038/s41559-017-0240-5.

Jauffrit F, Penel S, Delmotte S, Rey C, De Vienne DM, Gouy M, Charrier J-P, Flandrois J-P, Brochier-Armanet C. 2016. RiboDB database: a comprehensive resource

for prokaryotic systematics. *Molecular Biology and Evolution* **33**:2170–2172 DOI 10.1093/molbev/msw088.

Jones NC, Pevzner PA, Pevzner . 2004. *An introduction to bioinformatics algorithms*. Cambridge: MIT Press.

Jumas-Bilak E, Roudiere L, Marchandin H. 2009. Description of 'Synergistetes' phyl, nov. and emended description of the phylum 'Deferribacteres' and of the family Syntrophomonadaceae, phylum 'Firmicutes'. *International Journal of Systematic and Evolutionary Microbiology* **59**:1028–1035 DOI 10.1099/ijs.0.006718-0.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**:772–780 DOI 10.1093/molbev/mst010.

Kolmogorov AN. 1965. Three approaches to the quantitative definition of information. *Problems of Information Transmission* **1**:1–7.

Kullback S, Leibler RA. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* **22**:79–86 DOI 10.1214/aoms/1177729694.

Kurtzer GM, Sochat V, Bauer MW. 2017. Singularity: scientific containers for mobility of compute. *PLOS ONE* **12**:e0177459 DOI 10.1371/journal.pone.0177459.

Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**:3100–3108 DOI 10.1093/nar/gkm160.

Letunic I, Bork P. 2019. Interactive 'Tree of Life' (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**:W256–W259 DOI 10.1093/nar/gkz239.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**:1658–1659 DOI 10.1093/bioinformatics/btl158.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**:764–770 DOI 10.1093/bioinformatics/btr011.

Nesbø CL, Bapteste E, Curtis B, Dahle H, Lopez P, Macleod D, Dlutek M, Bowman S, Zhaxybayeva O, Birkeland N-K , et al. 2009. The genome of Thermosipho africanus TCF52B: lateral genetic connections to the Firmicutes and Archaea. *Journal of Bacteriology* **191**:1974–1978 DOI 10.1128/JB.01448-08.

Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**:268–274 DOI 10.1093/molbev/msu300.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**:D733–D745 DOI 10.1093/nar/gkv1189.

Olm MR, Brown CT, Brooks B, Banfield JF. 2017. *dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication*. London: Nature Publishing Group, 1–5 DOI 10.1038/ismej.2017.126.

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mashă: fast genome and metagenome distance estimation using MinHash. *Genome Biology* 1–14 DOI 10.1186/s13059-016-0997-x.

Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology* **38**:1079–1086 DOI 10.1038/s41587-020-0501-8.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**:1043–1055 DOI 10.1101/gr.186072.114.

Real R, Vargas JM. 1996. The probabilistic basis of Jaccard's index of similarity. *Systematic Biology* **45.3**:380–385 DOI 10.1093/sysbio/45.3.380.

Roure B, Rodriguez-Ezpeleta N, Philippe H. 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evolutionary Biology* **7(Suppl 1)**:S2 DOI 10.1186/1471-2148-7-S1-S2.

Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. 2020. GenBank. *Nucleic Acids Research* **48**:D84–D86 DOI 10.1093/nar/gkaa500.

Shannon CE. 1948. A mathematical theory of communication. *The Bell System Technical Journal* **27**:379–423 DOI 10.1002/j.1538-7305.1948.tb01338.x.

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Queinnec E, Ereskovsky A , et al. 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology* **27**:958–967 DOI 10.1016/j.cub.2017.02.031.

Taton A, Grubisic S, Brambilla E, Wit RD, Wilmotte A. 2003. Cyanobacterial diversity in natural and artificial microbial mats of Lake Fryxell (McMurdo Dry Valleys, Antarctica): a morphological and molecular approach. *Applied and Environmental Microbiology* **69.9**:5157–5169 DOI 10.1128/AEM.69.9.5157.

Tribus M, McIrvine EC. 1971. Energy and information. *Scientific American* **225**:179–190 DOI 10.1038/scientificamerican0971-179.

Van Vlierberghe M. 2021. Supplementary file 1. figshare. Dataset. London: Springer Nature. *Available at https:// doi.org/ 10.6084/ m9.figshare.14079866.v1*.

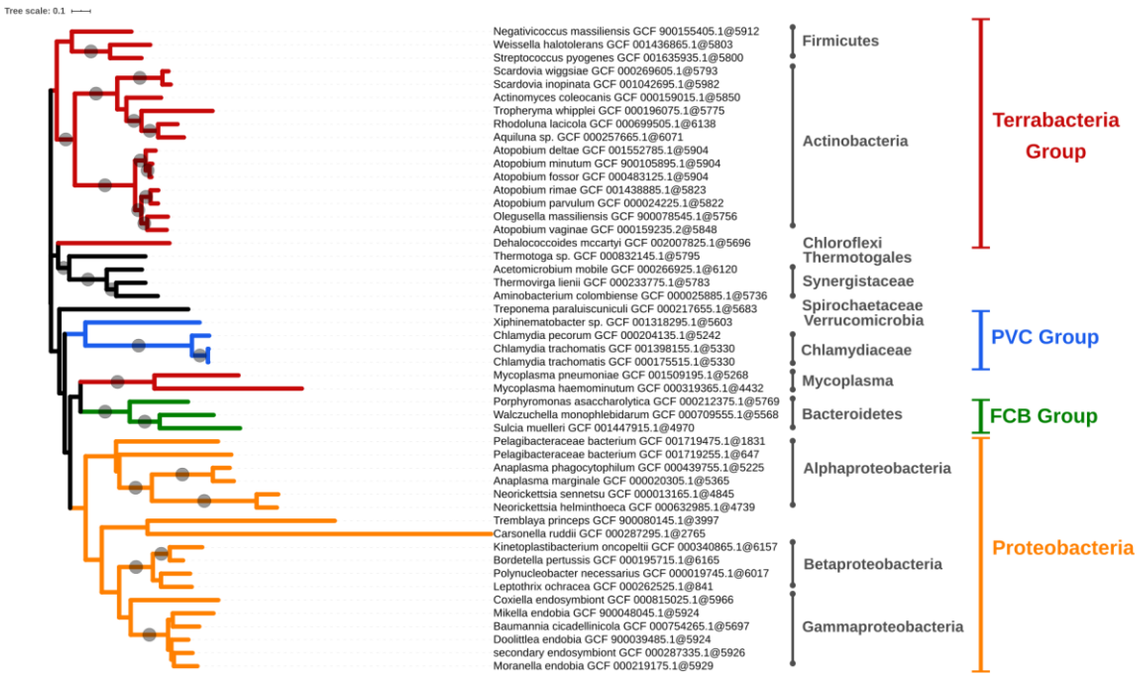
Wen J, Chan RH, Yau S-C, He RL, Yau SS. 2014. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene* **546**:25–34 DOI 10.1016/j.gene.2014.05.043.

Wick RR, Holt KE. 2019. rrwick/Assembly-Dereplicator: assembly dereplicator v0.1.0 (Version v0.1.0). Zenodo. DOI 10.5281/zenodo.3365572.

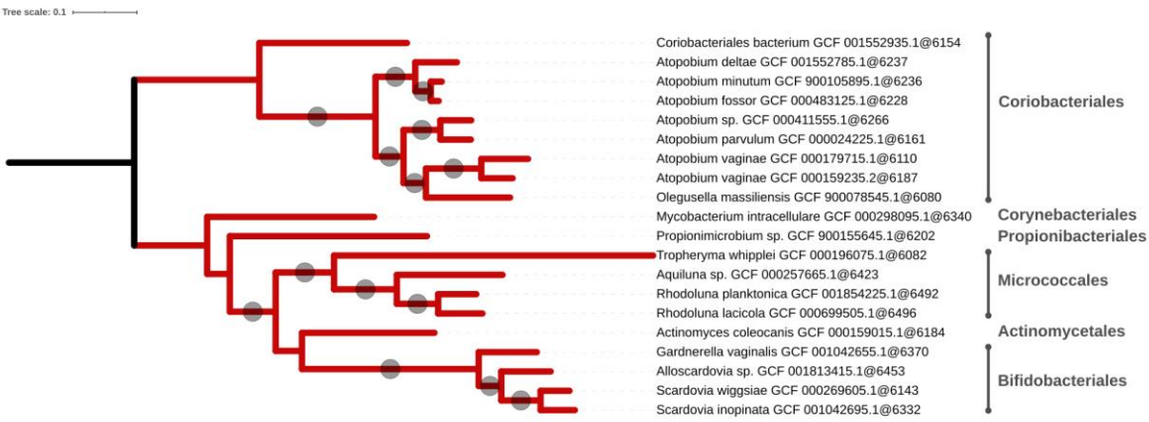
Zielezinski A, Vinga S, Almeida J, Karlowski WM. 2017. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology* **18**:186 DOI 10.1186/s13059-017-1319-7.

5.1.9 Supplementary materials

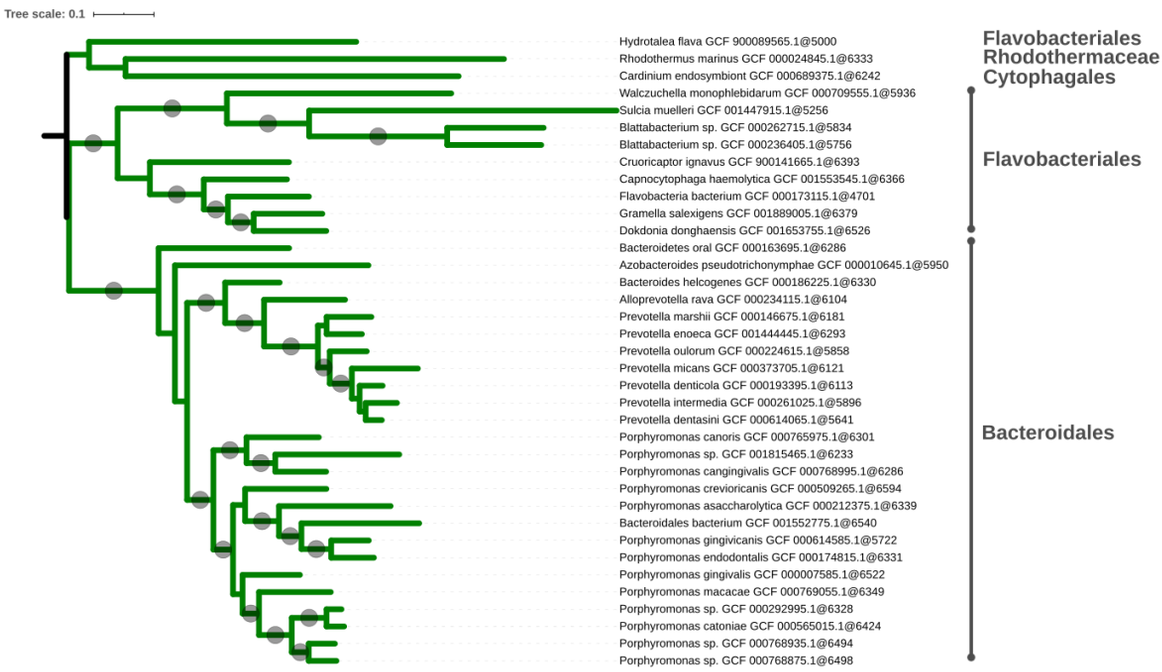
5.1.9.1 Supplementary figures



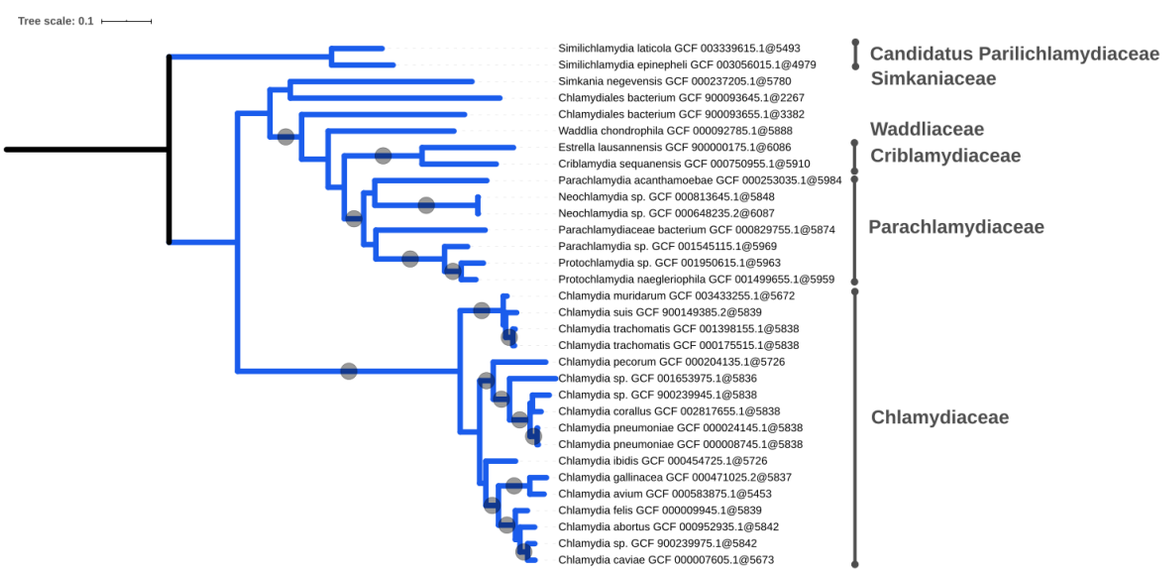
Supplementary Figure 1. Phylogenomic tree of the smallest selection of Bacteria. Tree inferred from a supermatrix of concatenated ribosomal proteins (Table 1, A) under the LG4X model using IQ-TREE. Dots on branches indicate maximum bootstrap support values (100%).



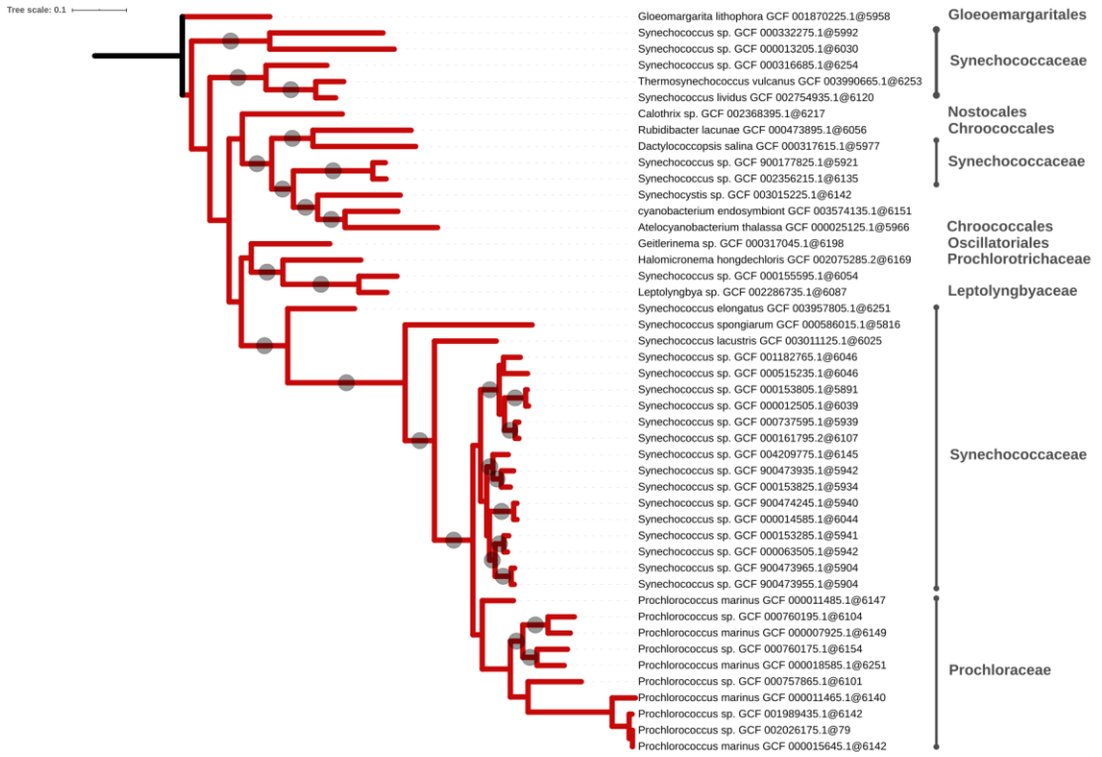
Supplementary Figure 2. Phylogenomic tree of the Actinobacteria. Tree inferred from a supermatrix of concatenated ribosomal proteins (Table 1, C) under the LG4X model using IQ-TREE. Dots on branches indicate maximum bootstrap support values (100%).



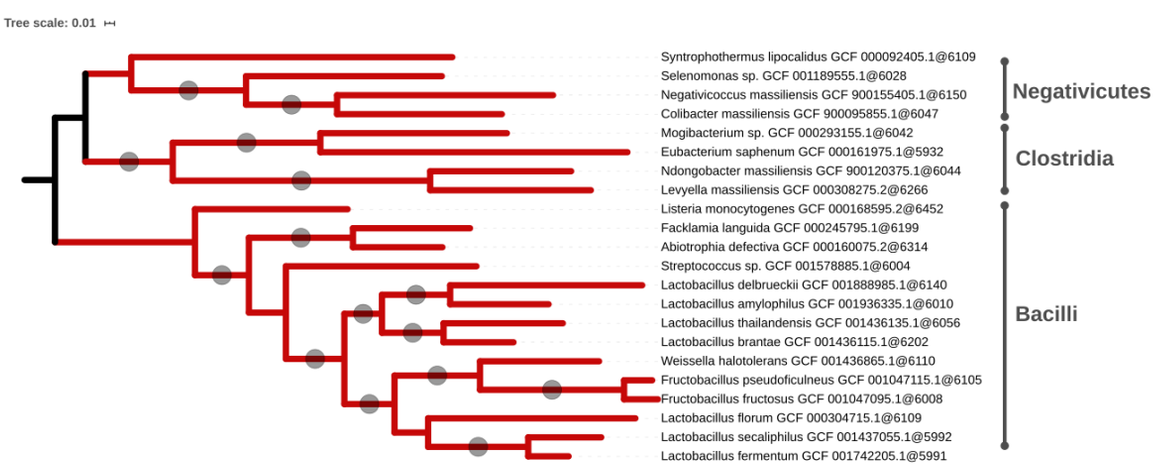
Supplementary Figure 3. Phylogenomic tree of the Bacteroidetes. Tree inferred from a supermatrix of concatenated ribosomal proteins (Table 1, D) under the LG4X model using IQ-TREE. Dots on branches indicate maximum bootstrap support values (100%).



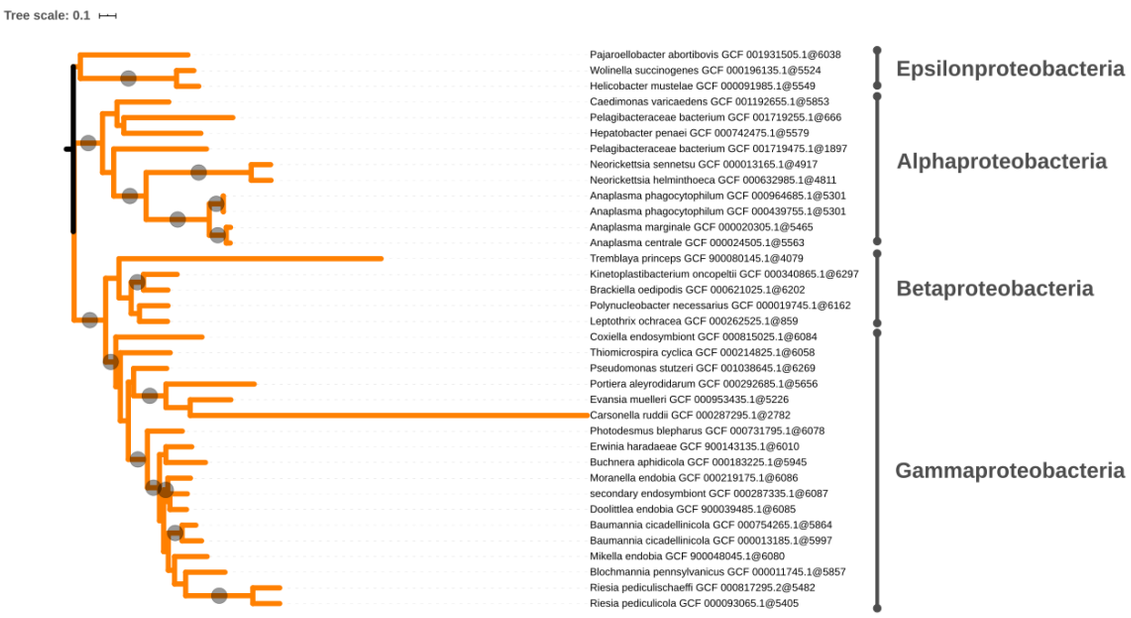
Supplementary Figure 4. Phylogenomic tree of the Chlamydia. Tree inferred from a supermatrix of concatenated ribosomal proteins (Table 1, E) under the LG4X model using IQ-TREE. Dots on branches indicate maximum bootstrap support values (100%).



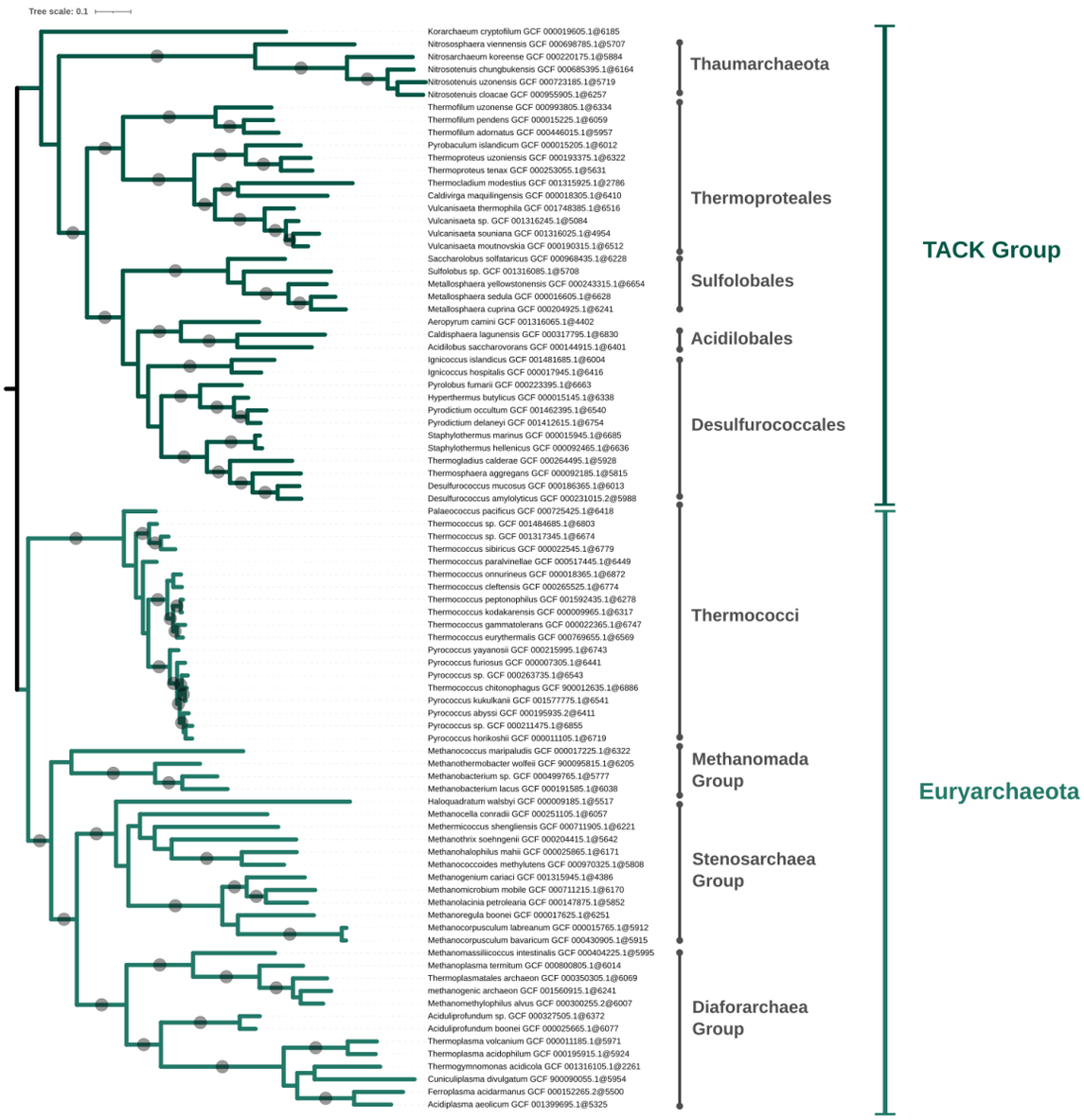
Supplementary Figure 5. Phylogenomic tree of the Cyanobacteria. Tree inferred from a supermatrix of concatenated ribosomal proteins (Table 1, F) under the LG4X model using IQ-TREE. Dots on branches indicate maximum bootstrap support values (100%).



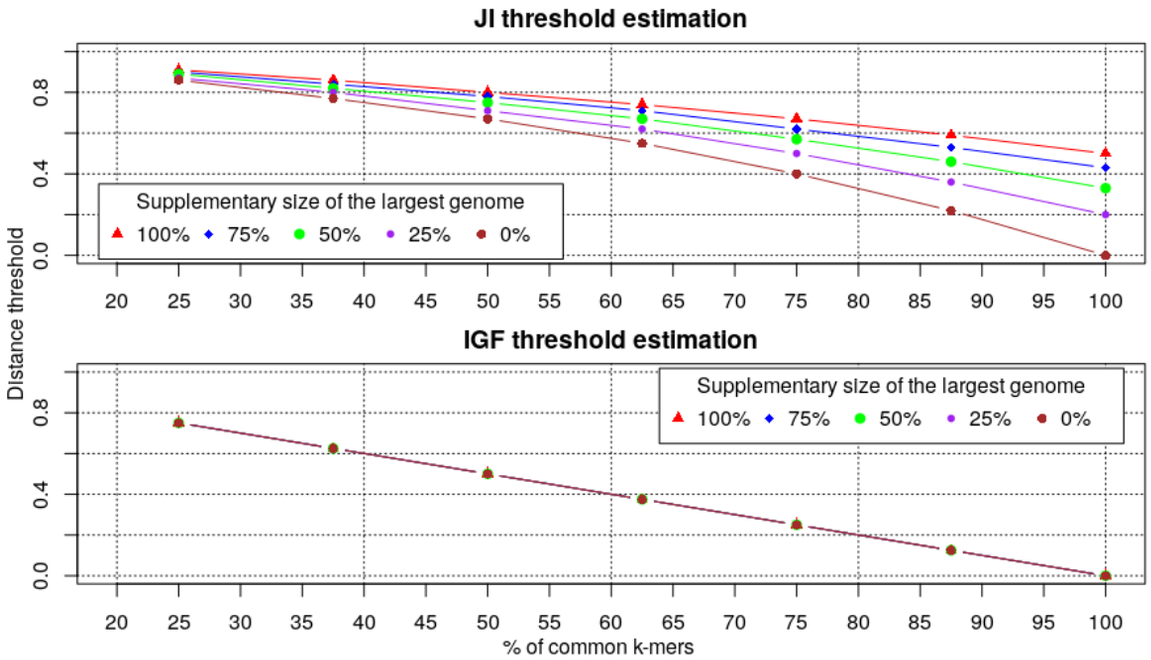
Supplementary Figure 6. Phylogenomic tree of the Firmicutes. Tree inferred from a supermatrix of concatenated ribosomal proteins (Table 1, G) under the LG4X model using IQ-TREE. Dots on branches indicate maximum bootstrap support values (100%).



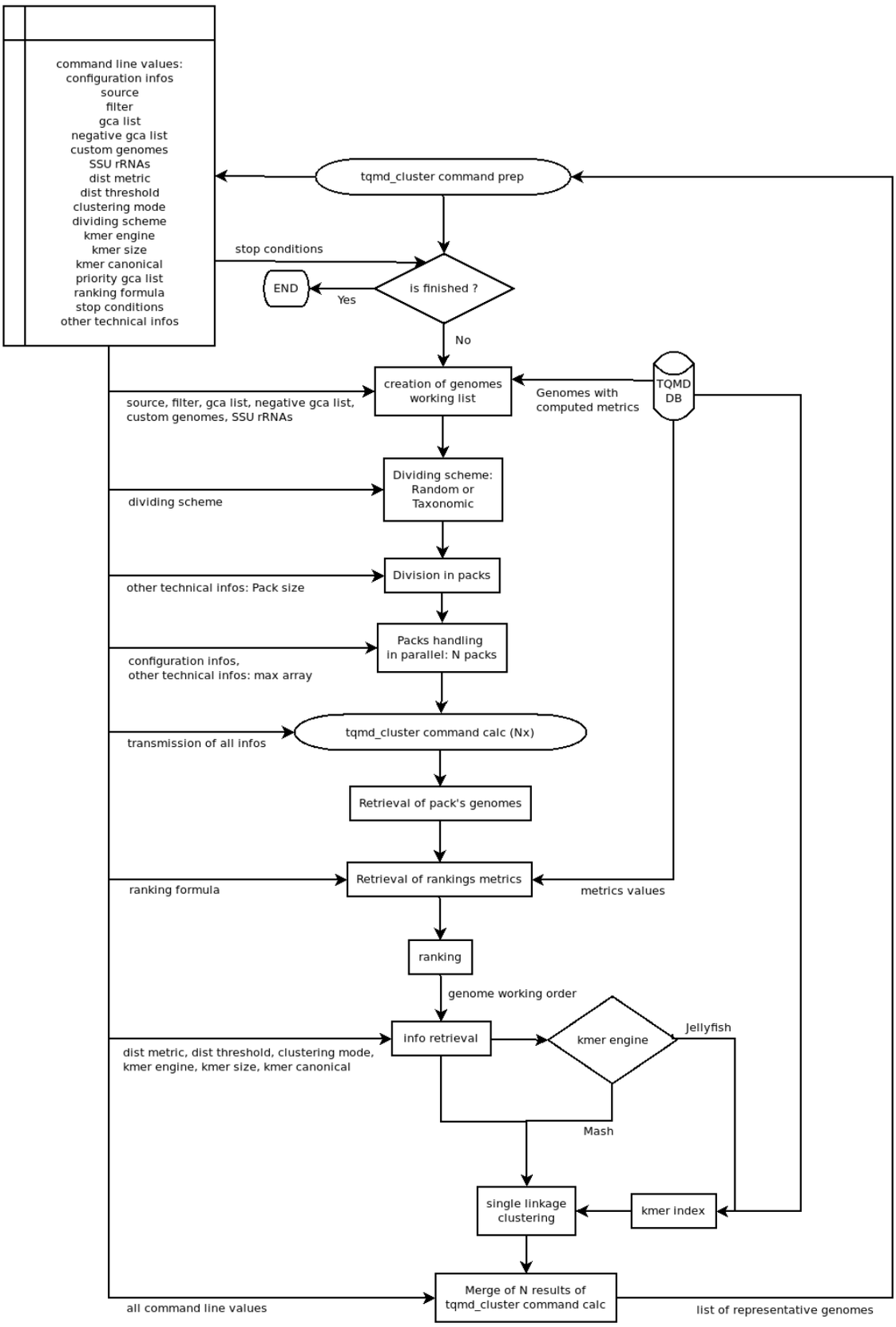
Supplementary Figure 7. Phylogenomic tree of the Proteobacteria. Tree inferred from a supermatrix of concatenated ribosomal proteins (Table 1, H) under the LG4X model using IQ-TREE. Dots on branches indicate maximum bootstrap support values (100%).



Supplementary Figure 8. Phylogenomic tree of the Archaea. Tree inferred from a supermatrix of concatenated ribosomal proteins (Table 1, I) under the LG4X model using IQ-TREE. Dots on branches indicate maximum bootstrap support values (100%).



Supplementary Figure 9. Evolution of the distance threshold (1-JI or 1-IGF) as a function of the proportion of common *k*-mers. The percentage of common *k*-mers is given from the smallest genome perspective, i.e., 25% of common *k*-mers means that 25% of all *k*-mers from the smallest genome are in common with the largest genome.



Supplementary Figure 10. Flowchart of tqmd_cluster.pl.

5.1.9.2 Supplementary tables

pack size	# representatives	
	JI-d	IGF-d
200	836	702
100	874	866
50	966	1197
25	1041	1387

Supplementary Table 1. Effect of the pack size on the final number of representative genomes. JI-based (direct) analyses were run using a distance threshold of 0.84, whereas IGF-based (direct) analyses used a threshold of 0.66. All analyses were run on 63,863 RefSeq Bacteria using the loose clustering mode

5.1.9.3 Comparison with Assembly-Dereplicator

A-D is a program that is more recent than dRep but, as of August 2020, not yet published; its last update dates from November 2019. Its main advantage is ease of use, since it is a simple (no-installation) script that only needs Mash as a prerequisite. A-D takes as input the path to a folder containing the genomes to be dereplicated and rearranges them randomly and separated into smaller packs (500 genomes per pack by default). The next step is the clustering of each pack serially using Mash. A-D stops as soon as it cannot dereplicate at least one genome from the current pack. Since it was compatible with our grid computer, we tried to test A-D (v0.1.0) with all prokaryotic RefSeq genomes, so as to mimic how TQMD is supposed to work in addition to the two datasets used with dRep.

For each of the two smaller datasets, A-D required only one CPU and took ten minutes when not partially crashing. Each dataset had to be relaunched several times due to A-D not finding the path to Mash for each pack. In the following, the TQMD results are the same as those reported for the dRep comparison. For Bacteroidetes, A-D selected 798 representatives (519 species), of which 704 were in common with TQMD, which represents 498 species in common. For the Streptomycetales, A-D selected 435 representatives (190 species), of which 408 were in common with TQMD, which represents 180 species (details given in Table S2).

In March 2019, all RefSeq prokaryotic genomes amounted to 112,254 genomes. We launched several A-D runs which all stopped after one hour and failed to dereplicate more than 5,000 genomes despite an increasingly lenient threshold (details given in Table S3). Investigation of the results revealed that this time the problem was not due to A-D not finding Mash but caused by the heuristics implemented in A-D.

dataset	starting		TQMD - JELLYFISH k12			Assembly-Dereplicator			intersection	
	# gen.	# spec.	# repr.	# spec.	h.CPU	# repr.	# spec.	h.CPU	# repr.	# spec.
Bacteroidetes	1127	528	789	517	60	798	519	0.1	704	498
Streptomycetales	648	220	486	207	40	435	190	0.1	408	180

Supplementary Table 2. Performance comparison between TQMD and Assembly-Dereplicator on two smaller datasets. The column titles and the TQMD results are taken from Table 7.

dist. threshold	# representatives	# derepl. packs
0.01	111,855	4
0.10	110,160	9
0.20	111,596	3
0.30	112,254	1
0.40	108,774	8
0.50	111,755	2

Supplementary Table 3. Attempts at dereplicating all RefSeq Bacteria (releases 79+92, 112,254 genomes) using Assembly-Dereplicator. Analyses were run using 6 different distance thresholds and the default pack size of 500 (225 packs).

Apparently, A-D does not work with very large and non-homogeneous groups of genomes. We did not investigate further the script with large non-homogenous datasets since our tests clearly showed that it cannot be compared, in its present state, with TQMD. A-D can dereplicate smaller sets of homogeneous genomes (such as the Cyanobacteria or the aforementioned Bacteroidetes and Streptomycetales) provided the bug with Mash not being recognized is solved. Yet, drawing on our own tests with TQMD (see main text), our intuition is that the A-D approach based on a random splitting of the genomes to dereplicate, if appropriate when working with homogeneous genomes, are likely to be inefficient when it comes to non-homogeneous datasets. Moreover, the stop conditions of the iterative heuristics obviously lead A-D to get stuck very easily.

5.2 Was the last bacterial common ancestor a monoderm after all?

Raphaël R. Léonard[a,b], Eric Sauvage[a], Valérian Lupo[a,b], Amandine Perrin[c,d], Damien Sirjacobs[b], Paulette Charlier[a], Frédéric Kerff[a,#], Denis Baurain[b,#]

[a] InBioS / Centre d'Ingénierie des Protéines, Université de Liège, Belgium
[b] InBioS / Phylogénomique des Eucaryotes / PhytoSYSTEMS, Université de Liège, Belgium
[c] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France.
[d] Hub de Bioinformatique et Biostatistique - Département Biologie Computationnelle, Institut Pasteur, Paris, France

Running Title: A monoderm ancestor for the bacterial domain

Address correspondence to:
Denis Baurain: denis.baurain@uliege.be
Frédéric Kerff: fkerff@uliege.be

5.2.1 Abstract

The very nature of the last bacterial common ancestor (LBCA), in particular the characteristics of its cell wall, is a critical issue to understand the evolution of life on earth. Although knowledge of the relationships between bacterial phyla has made progress with the advent of phylogenomics, many questions remain, including on the appearance or disappearance of the outer membrane (OM) of diderm bacteria (also called Gram-negative bacteria). The phylogenetic transition between monoderm (Gram-positive bacteria) and diderm bacteria, and the associated peptidoglycan expansion or reduction, requires clarification. Herein, using a phylogenomic tree as an evolutionary framework and a literature review of cell-wall characteristics, we used Bayesian ancestral state reconstruction to infer the cell-wall architecture of the LBCA. With the same phylogenomic tree, we further revisited the evolution of the division and cell-wall synthesis (*dcw*) gene cluster using homology- and model-based methods. Finally extensive similarity searches were carried out to determine the phylogenetic distribution of the genes involved with the biosynthesis of the OM in diderm bacteria. Quite surprisingly, our analyses suggest that all extant bacteria might have evolved from a common ancestor with a monoderm cell-wall architecture. If true, this indicates that the appearance of the OM was not a unique event and that selective forces have led to the repeated adoption of such an architecture.

5.2.2 Introduction

Cell-wall architecture has always been an important morphological character for bacterial classification (Schleifer and Kandler, 1972). Two main types of cell wall exist: the monoderm and the diderm architectures. While monoderm bacteria are generally surrounded by a thick peptidoglycan (and are positive to Gram coloration), in diderm bacteria, a thin peptidoglycan layer is sandwiched between the cytoplasmic membrane and the outer membrane (OM; and are negative to Gram coloration) (Coico, 2006; Silhavy et al., 2010). However, cell-wall features are insufficient to yield a classification that would correlate with phylogenetic trees based on molecular data (Woese, 1987). Hence, distantly related phyla may have apparently identical cell walls (e.g., Negativicutes and Proteobacteria), whereas closely related phyla or families may present variations in their peptidoglycan thickness or composition, and even in the number of surrounding membranes (e.g., Negativicutes and Halanaerobiales compared to other Firmicutes) (Megrian et al., 2020). Nonetheless, the evolution of the bacterial cell wall should be addressed in light of the phylogeny of the domain. The number of membranes (one or two) that surround a bacterial cell, their lipid composition and the thickness of the peptidoglycan layer are undoubtedly major characteristics of the bacterial cell wall, and these features frequently come into consideration when discussing the evolution of the bacterial domain. Hence, transition from one to two lipid membranes (or the opposite) has attracted much attention. Disappearance of the outer membrane leading from “diderm” to “monoderm” architecture has been proposed by Cavalier-Smith (Cavalier-Smith, 1987; Cavalier-Smith, 2010) but evolution from monoderm to diderm bacteria is usually favored by other evolutionary biologists (Sutcliffe, 2010; Gupta, 2011; Errington, 2013). It has been suggested that the endosymbiosis between an “actinobacterium” and a “clostridium” could be the starting point for the onset of double-membrane bacteria (Lake, 2009), but how exactly this symbiosis could have further evolved to form a diderm bacterium remains to be detailed. An attractive hypothesis accounting for the emergence of the OM is its evolution from a forespore of a spore-former “firmicute”. Based on 3D electron cryotomographic images of spore formation in the diderm firmicute *Acetonema longum*, Tocheva *et al.* showed that the inner membrane (IM) of the mother cell is inverted to become the OM of the forespore and ultimately of the germinating cell (Tocheva et al., 2011), leading to the assumption that the OM of diderm bacteria could have evolved from monoderms via sporulation (Tocheva et al., 2011, 2016; Vollmer, 2011; Errington, 2013). In contrast, some studies of the evolution of the cell-wall architecture in the phylum Firmicutes interpreted the double membrane found in Halanaerobiales and Negativicutes (two classes of Firmicutes) as a reminiscence of double membrane in the Firmicutes ancestor, and thus concluded that the OM was lost multiple times in this phylum (Antunes et al., 2016; Taib et al., 2020). This interpretation further opens the possibility that the last bacterial common ancestor (LBCA) was a *bona fide* diderm bacterium.

Cell division in bacteria involves a series of proteins that fulfil many functions as diverse as cytoplasmic membrane invagination, DNA transfer control, peptidoglycan synthesis and daughter cell separation. They assemble into a dynamical complex that overpasses the cytoplasmic membrane and has components in both the cytoplasm and the periplasm. A small number of these proteins are essential and conserved in the genome of almost all bacteria. Several of these proteins of cell division are generally clustered together with proteins involved in peptidoglycan synthesis in a single locus on the genome, the *dcw* (division and cell-wall synthesis) cluster (Mingorance and Tamames, 2004). This cluster is found in many bacteria and its composition and gene order are generally well conserved (Tamames, 2001; Real and Henriques, 2006). It has also been shown to be one of the most stable gene clusters, on par with the ribosomal clusters (Nikolaichik and Donachie, 2000; Barloy-hubler et al., 2001). The longest version of the *dcw* cluster includes 17 genes and encompasses genes coding for proteins responsible for

peptidoglycan precursors synthesis (DdlB, MurA, MurB, MurC, MurD, MurE, MurF, MurG, MraY), proteins integrated in the divisome (FtsA, FtsI, FtsL, FtsQ, FtsW, FtsZ), and proteins involved in regulation via DNA binding or RNA methylation (MraW, MraZ). The *E. coli dcw* cluster includes 15 genes, starting with *mraZ* and ending with *ftsZ*, but misses the *murA* and *murB* genes (Eraso et al., 2014). Many phyla, orders, classes, or families are apparently characterized by the lack of specific genes in the cluster, the absence of *ftsA* and *ftsZ* in Chlamydiae and Planctomycetes being a well-known example (Pilhofer et al., 2008). These observations suggest that the organization of the *dcw* cluster holds clues to bacterial evolution. Thus, its detailed study might complement sequence-based phylogenomic approaches, including in terms of rooting of the bacterial tree. For example, the integration of a gene in a specific position within the cluster probably happened only once in the history of the bacterial domain, whereas gene loss and genomic reorganization events, on the contrary, are expected to have been more frequent. Likewise, the phylogenetic distribution of the genes involved in the biosynthesis of the OM in diderm bacteria might provide useful information about their evolutionary status, ancestral or derived, with respect to the bacterial domain as a whole (Megrian et al., 2020; Taib et al., 2020; Coleman et al., 2021).

In this work, we built a Bayesian phylogenomic tree of the bacterial domain using a supermatrix of 117 single-copy orthologous genes sampled from 85 species representative of the bacterial diversity and for which a descriptive literature exists. We then researched the cell-wall architectures for these species and used the tree to reconstruct the evolution of two cell-wall traits, the number of membranes and the presence and thickness of the peptidoglycan layer, again with Bayesian inference. Moreover, we compared the composition and gene order of the *dcw* cluster in our 85 representative species, and used a new variant of a homology-based method to map the organization of the *dcw* cluster on the evolution of the bacterial domain. Contrary to our expectations based on recent literature and educated guesses, our Bayesian analyses inferred that the LBCA was a monoderm bacterium with a thick peptidoglycan. This reconstruction implies that the OM of diderm bacteria appeared more than once, an hypothesis that is indeed supported by differences in the genetic machinery involved in its biosynthesis across the various diderm lineages, as evidenced by our extensive similarity searches. Our results also show that the LBCA already possessed a complete *dcw* cluster and that its organization does not correlate with cell-wall architecture.

5.2.3 Results

5.2.3.1 A robust tree of the bacterial domain

To serve as the base for evolutionary analysis of the cell-wall architecture and reconstruction of the ancestral gene order in the *dcw* cluster, we needed a tree of Bacteria. With the growing availability of fully sequenced genomes, phylogenomics has developed as a discipline using the tools of phylogenetics but applied to tens to hundreds, or even thousands, sequences of broadly conserved genes (Delsuc et al., 2005). Phylogenomic trees can either be inferred from supermatrices of concatenated genes (Philippe et al., 2017) or through combination of single-gene trees into supertrees (Liu et al., 2019). Hence, the phylogenomic tree shown in Figure 1 was computed by Bayesian inference based on a dense (4.29% missing character states) supermatrix of 117 single-copy orthologous genes (see Materials and Methods) sampled from 85 representative bacterial genomes with PhyloBayes MPI under the site-heterogeneous CAT+GTR+Γ model of sequence evolution (Lartillot and Philippe, 2004, 2006; Lartillot et al., 2007, 2013). Congruence analyses were run on the 117 individual genes using Phylo-MCOA (De Vienne et al., 2012) and did not reveal incongruent genes or species, beyond 62 individual sequences, which might have experienced gene transfer and/or fast evolution. Once discarded, the overall

results did not change, as demonstrated by comparing two control trees (i.e., before and after outlier removal) inferred with RAxML under the LG+F+Γ model (see Figures S1 and S2). Regarding model selection, cross-validation analyses on four different models confirmed that CAT+GTR+Γ had the best fit to our dataset, followed by CAT+Γ, then GTR+Γ and finally LG+Γ (Table S1).

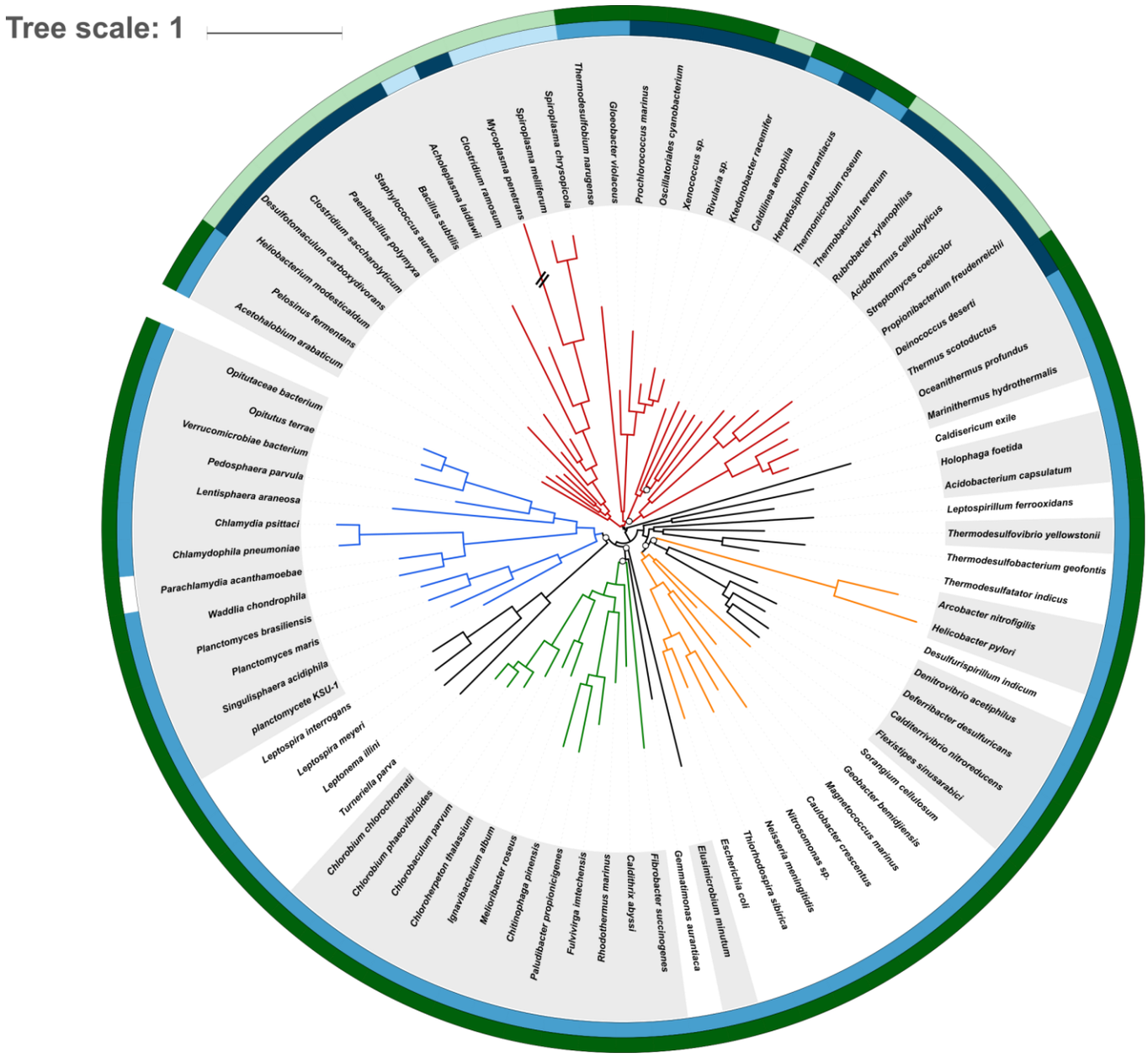


Figure 1: Phylogenomic tree of the bacterial domain based on a supermatrix concatenating 117 single-copy orthologous genes chosen for their broad conservation across Bacteria. The supermatrix contained 85 species and 19,959 unambiguously aligned amino-acid positions (< 5% missing character states). The tree was inferred from amino-acid sequences using PhyloBayes MPI and the CAT+GTR+Γ model of sequence evolution. Open symbols at the nodes are posterior probabilities (PP), and nodes without a symbol correspond to maximum statistical support for phylogenetic inference (PP of 1.0; averaged over two MCMC chains). The length of the branch marked with "//" has been reduced by 50% for the sake of clarity. Outer circles represent the status of the peptidoglycan (PG) and of the OM in the organisms, according to our literature survey. Dark blue = thick PG, blue = thin PG, light blue = no PG. Dark green = diderm, light green = monoderm. White = no information.

Our unrooted tree is in good agreement with most recent concatenating phylogenomic studies aimed at resolving bacterial evolution (Battistuzzi and Hedges, 2009; Wu et al., 2009; Yutin et al., 2012; Lasek-nesselquist and Gogarten, 2013; Rinke et al., 2013; Raymann et al., 2015; Hug et al., 2016; Castelle and Banfield, 2018; Parks et al., 2018; Zhu et al., 2019; Cavalier-Smith et al., 2020). In particular, we robustly recovered a bipartition of the bacterial lineages composing the Terrabacteria and the "Hydrobacteria" (= Gracilicutes sensu (Cavalier-Smith, 2006)). Within these

“megaphyla” first defined by Hedges and Battistuzzi (Battistuzzi and Hedges, 2009), resolution was weaker, as reflected in the lower posterior probabilities (PPs) observed at medium phylogenetic depth, whereas phyla and known superphyla (e.g., FBC, PVC) were always clearly resolved. In the first group, relationships between member lineages slightly varied from run to run (we ran a total of six independent chains, Figure S3), while in the second group, Epsilonproteobacteria were occasionally separated from other groups of Proteobacteria (Figures S4 and S5A to S5F). Some additional phyla initially present in our dataset (i.e., Synergistetes, Fusobacteria and Aquificae) were excluded from the tree shown in Figure 1 because they were difficult to robustly position (e.g., due to the chimerical nature of the Aquificae) without bringing more cell-wall architecture diversity (see also (Zhaxybayeva et al., 2009; Bhandari et al., 2012; Eveleigh et al., 2013)). Likewise, we further discarded the Thermotogae, which are also chimeras (Zhaxybayeva et al., 2009), even though their toga might be akin to a modified OM (Rachel et al., 1988, 1990) (see Figure S6 for a preliminary 101-species tree including all these lineages). Such uncertainties are not uncommon in bacterial phylogenomics and are the result of a combination of weak phylogenetic signal, widespread lateral gene transfer and systematic error (e.g., long-branch attraction artifacts) (Bapteste et al., 2004; Mira et al., 2004; Beiko et al., 2005; Koonin, 2005, 2016; Boussau et al., 2008; Philippe et al., 2011; Eveleigh et al., 2013; Gouy et al., 2015).

Rooting the different domains of Life is not an easy issue (Gouy et al., 2015). In Figure 1, we elected to set the root of Bacteria between Terrabacteria and Hydrobacteria/Gracilicutes, following studies having included Archaea as an outgroup (Lartillot et al., 2007; Coleman et al., 2021). Interestingly, this basal split mirrors cell-wall architecture differences. In the first group, Firmicutes, Tenericutes, Actinobacteria, and presumably Chloroflexi (see below), are mostly monoderm bacteria. Together with the atypical diderms (AD), i.e., Deinococcus-Thermus, Cyanobacteria, Synergistetes and Thermotogae, they compose Terrabacteria (Battistuzzi and Hedges, 2009). On the other hand, the remaining lineages are diderms mostly featuring lipopolysaccharides (LPS) and correspond to Hydrobacteria/Gracilicutes; these will be called “true diderms-LPS” (TDL) in this study. Over time, several positions for the bacterial root have been proposed (Table S2). In the following, since our Bayesian analyses required a rooted tree, we tested several of them, yet excluding roots lying within TDL, which are likely monophyletic (see below). Beyond the root of Figure 1, we thus explored the effect of setting the bacterial root within Terrabacteria on our inferences.

5.2.3.2 Evolution of the cell-wall architecture

In order to study the evolution of the cell-wall architecture, we carried out a thorough literature survey on all the bacteria retained in our tree (Tables S3 and S4). For each organism, we collected the number of membranes, the presence and thickness of the peptidoglycan layer and, if relevant, the type of spore, as there exists evidence of potential functional connection between sporulation and cell-wall remodeling processes (Tocheva et al., 2011, 2016). However, preliminary analyses indicated that the spore trait was difficult to encode reliably in terms of homologous states. Therefore, it was eventually discarded, whereas the two traits linked to the cell wall itself were analysed using BayesTraits under the MultiState model.

Based on this survey (Tables S3 and S4), most bacterial phyla have two membranes (diderm architecture) and a thin peptidoglycan layer. For example, Proteobacteria, Nitrospirae, Acidobacteria, Bacteroidetes and Chlorobi fall into this category and correspond to TDL lineages. For the organisms belonging to the PVC superphylum, this architecture might be slightly different (Rivas-Marín et al., 2016). Actinobacteria are essentially monoderms with a thick peptidoglycan, whereas Firmicutes and Chloroflexi both have monoderm and diderm representatives. Firmicutes

include Bacilli and Clostridia, two groups of endospore formers. Clostridia and Bacilli correspond to two well-defined classes, sharing many traits though being also very distinct. All Bacilli and most Clostridia are monoderms with a thick peptidoglycan, but some “clostridia” (Halanaerobiales and Negativicutes) have two membranes (some with LPS in the OM) and a relatively thin peptidoglycan layer (Mavromatis et al., 2009; Kivistö and Karp, 2011; Antunes et al., 2016). Regarding the status of the Chloroflexi cell-wall architecture, it is still controversial (Sutcliffe, 2011; Cavalier-Smith et al., 2020). Beside these canonical diderm and monoderm phyla, respectively corresponding to classical Gram- and Gram+ bacteria, there exist a series of organisms with atypical cell-wall architectures. Hence, Deinococcus-Thermus and Cyanobacteria are diderm bacteria with an OM, but their cell walls differ from those of the TDL by having a thick peptidoglycan instead of a thin layer (Table S4).

Consequently, the number of membranes observed in the extant organisms is either one (state 0) or two (i.e., there is an OM, state 1; Table S3). The evolutionary analysis of this trait suggests a LBCA surrounded by only one membrane. This inference is robust to five model variants (E, H1, H2, R1 and R2; see Materials and Methods) and six different positions for the bacterial root (P(0) = 94.2% to 98.2%; Figure S7). Due to the robustness of our results to alternative rootings, we will only present those obtained with a root located between Terrabacteria and TDL (as in Figure 1). In accordance with the inference of a monoderm LBCA, the posterior transition rates indicate that it is easier to gain (q01) an OM (range of the five model’s mean = 2.288-2.495, Table 1) than losing (q10) an existing one (range = 0.008-0.132). If we try to alter the H1/H2 model hyperpriors to promote the loss (q10 = 1-10) at the expense of the gain (q01 = 0-1), the LBCA remains inferred as a monoderm in 67.1% of the cases (mean P(0)), whereas it is inferred as a diderm in 32.9% of the cases (mean P(1)) (Table 1). Concerning the rates, the inferred loss rate remains weak (mean q10 = 0.000-0.187; Table 1), while the distribution of the gain rate (q01) becomes bimodal, with a mode at 0.2 and another at 1.8 [Figure S8A], and remain low for the loss rate (q10) [Figure S8B]. Consequently, under this extreme parameterization, we distinguish two main configurations for the pair of rates (Figure S8C) and the monoderm probability P(0) (Figure S8D).

Node	trait	statistic	E	H1	H2	R1	R2	H biased
LBCA	MBN	mean q01	2.495	2.352	2.477	2.288	2.411	1.431
LBCA	MBN	mean q10	0.132	0.113	0.121	0.012	0.008	0.210
LBCA	MBN	mean P(0)	94.951	94.204	95.375	97.134	98.161	67.092
LBCA	PG	mean P(0)	22.068	4.022	38.604	0.397	0.594	N/A
LBCA	PG	mean P(2)	76.497	94.622	60.147	99.535	99.358	N/A
LBCA	PG	mean q01	4.626	1.634	7.317	0.798	0.827	N/A
LBCA	PG	mean q02	6.935	2.020	20.967	0.953	1.041	N/A
LBCA	PG	mean q10	0.166	0.102	0.187	0.000	0.000	N/A
LBCA	PG	mean q12	0.128	0.109	0.118	0.001	0.000	N/A
LBCA	PG	mean q20	2.088	0.937	4.941	1.347	1.413	N/A
LBCA	PG	mean q21	1.890	2.165	1.600	1.398	1.419	N/A
Firmicutes	PG	mean P(0)	17.631	3.936	30.120	0.611	0.738	N/A
Firmicutes	PG	mean P(2)	81.891	95.648	69.435	99.378	99.237	N/A

Table 1: Overview of BayesTraits results. qij design posterior transition rates, whereas P(i) correspond to posterior ancestral state probabilities. For the membrane (MBN) trait, state 0 = one MBN and state 1 = two MBN, while for the peptidoglycan (PG) trait, state 0 = no PG, state 1 = thin PG and state 2 = thick PG. “H biased” is the model where the hyperprior has been purposely biased to favor a diderm LBCA (see Materials and Methods for details).

In the 85 extant organisms considered in our study, the peptidoglycan layer is either absent (state 0), present and thin (state 1) or present and thick (state 2; Table S3). The LBCA is inferred with a thick peptidoglycan. While this result is robust to alternative positions of the root, some models (E and H2) let the possibility open (22.0-38.6%, Table 1) for the LBCA having been devoid of peptidoglycan (Figure S9). Moreover, the posterior rates are highly heterogeneous, depending on the transition considered, and present a sensitivity to the model used (mean range = 0.000-20.967; Figure S10 and Table 1). Based on the values of the rates, the thin peptidoglycan state (state 1), once acquired, is unlikely to change towards another state, whereas the other two states (states 0 and 2) can exchange freely or change towards the thin peptidoglycan state (Figure S10 and Table 1).

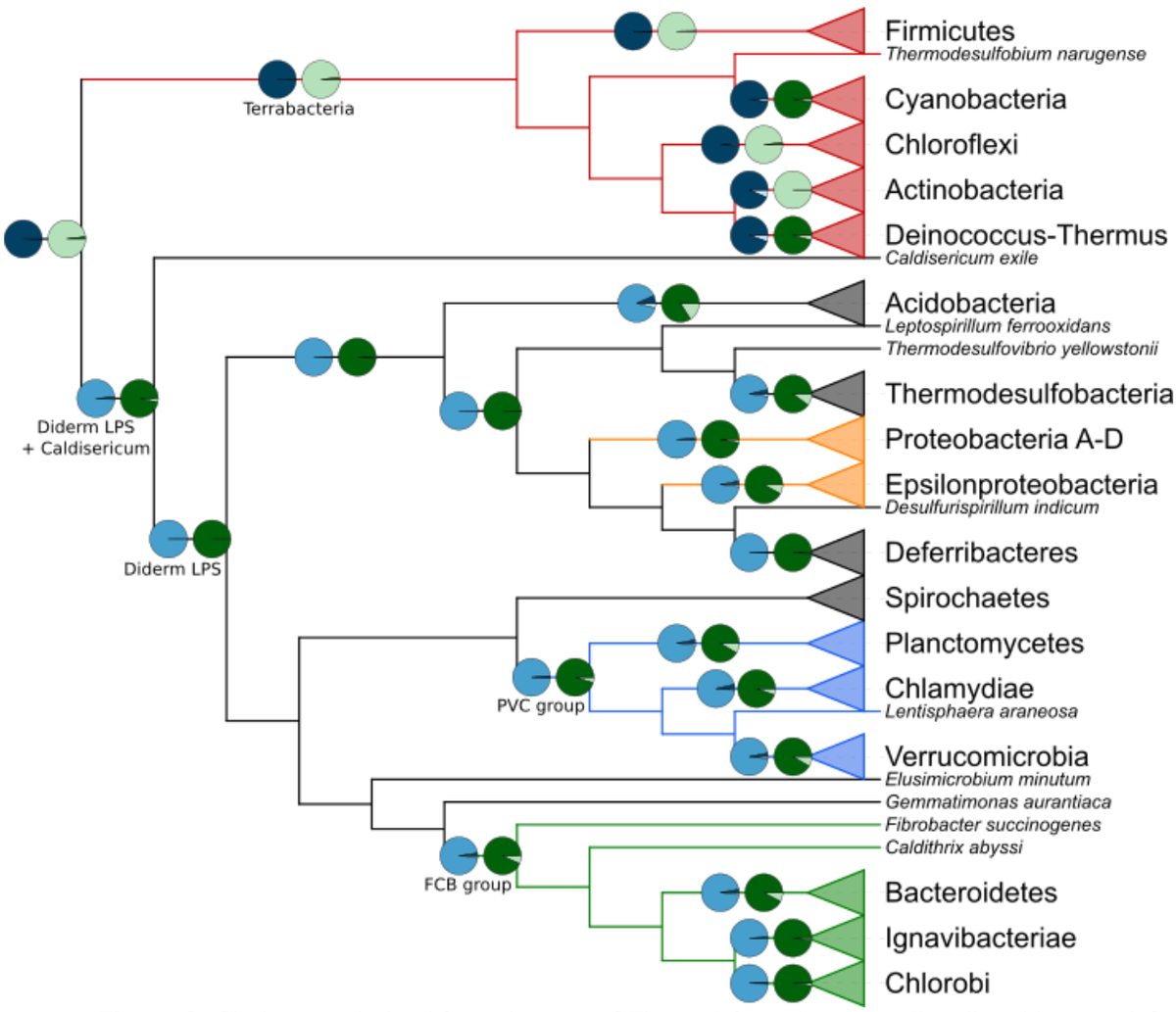


Figure 2: Cladogram derived from the tree of Figure 1 featuring the cell-wall architecture inferred for selected last common ancestors among Bacteria. The pie chart sectors correspond to the PP of the model reverse-jump hyperprior exponential 0 to 100 (R2). Dark blue = thick PG, blue = thin PG, light blue = no PG. Dark green = diderm, light green = monoderm.

In a second step, we used BayesTraits to reconstruct the state of the characters for the Last Common Ancestor (LCA) of every of the 15 bacterial phyla included in our study, as well as the LCA of several larger groups (e.g., PVC, Terrabacteria), still based on the Terrabacteria root (Figure 2). As expected, the LCA of the TDL bacteria is inferred as a diderm organism featuring a thin peptidoglycan layer, whereas the Terrabacteria LCA is reconstructed as a monoderm with thick peptidoglycan. The results obtained for the larger groups are homogeneous across the different models (Figure S11). For Firmicutes, which is the only phylum with some architectural diversity in our dataset, two of the five models (E and H2) do not completely settle on an LCA monoderm with a thick peptidoglycan, and instead do not dismiss an LCA without peptidoglycan (17.6% and 30.1%, respectively; Table 1). Finally, a comparison of the fit of the five models using Bayes Factors (Table 2) showed that model R1 was the best, followed by models R2, H1, E, and finally H2. Therefore, the two models that do not fully agree with the others about the peptidoglycan trait are also those that are deemed less fit by Bayes Factors (E and H2).

complex	simple	MBN	PG
R1	H2	7.41	22.86
	E	5.95	17.47
	H1	2.69	8.38
	R2	2.42	1.91
R2	H2	4.99	20.95
	E	3.53	15.56
	H1	0.27	6.47
H1	H2	4.71	14.47
	E	3.25	9.09
E	H2	1.46	5.39

Table 2: Pairwise comparisons of BayesTraits model fit using Bayes Factors (BF). BF > 2 are interpreted as positive evidence, 5 ≤ BF < 10 as strong evidence and BF > 20 as very strong evidence in favor of the more complex model (Gilks et al., 1995).

Hitherto, the two cell-wall traits were analysed separately, owing to the limitations of the MultiState model used. However, from a biological point of view, their evolution might be correlated. To account for this possibility, we conducted the BayesTraits procedure to estimate the correlation between two traits, which revealed that the peptidoglycan and the membrane characters are indeed linked. The actual strength of the correlation depended on the scheme used to recode the three-state peptidoglycan trait into a binary character, which was needed to estimate the correlation with the membrane trait (see Materials and Methods). When the coding scheme rewarded the mere presence of the peptidoglycan layer, whatever its thickness, the correlation was supported by strong evidence (log Bayes Factor for case A = 9.0), while it raised to very strong evidence when the scheme emphasized either a thick peptidoglycan (case B = 27.6) or a thin peptidoglycan (case C = 37.8). These differences in correlation can easily be explained. In case A, almost all organisms of our study without peptidoglycan are also deprived of the OM (see *Parachlamydia acanthamoebae* in Figure 1), whereas organisms with a peptidoglycan layer often have an OM. In case B, all organisms without peptidoglycan or with a thin peptidoglycan layer are put in the same category. In our study, all organisms with a thin peptidoglycan layer have an OM, and they are more numerous than the organisms without a peptidoglycan layer. In case C, the organisms with a thin peptidoglycan layer have their own category and, in our study, all these organisms also feature an OM.

5.2.3.3 Evolution of the gene order within the *dcw* cluster

Initially, we studied the organization of the *dcw* cluster in extant organisms based on the output of a custom visualisation software showing orthologous gene groups (OGs) in their syntenic context (see Materials and Methods for details and “synteny_85_dcw.pdf” available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder ProCARs, for the status of the *dcw* cluster in the 85 bacteria of our phylogenomic tree). This approach led us to identify the OGs for the 17 genes of (the most complete form of) the *dcw* cluster. In Cyanobacteria, the nearly total absence of the *dcw* cluster is noteworthy: *mraZ* and *ftsA* are missing from all cyanobacterial genomes examined, and all other genes of the cluster are generally present but completely dispersed on

almost as many loci as the number of genes, with some exceptions, the doublet *murC* and *murB* or the doublet *ftsQ* and *ftsZ* (see .xlsx file available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder ProCARs). The *murA* gene can be found in clusters or sub-clusters in several genomes. The complete form of the *dcw* cluster is only observed in a single order of Clostridia, the Halanaerobiales (more precisely, in *Acetohalobium arabaticum*). Halanaerobiales are robustly affiliated to Firmicutes, yet branching at the root of the phylum (Yutin and Galperin, 2013). However, *murA* is also present in sub-clusters in Cyanobacteria, Planctomycetes, Lentisphaerae and *Calditrix abyssi*. Otherwise, if present in the genome, *murA* is usually located outside of the *dcw* cluster. Beside this particular gene and particular phyla, several TDL phyla are characterized by the loss of specific genes from the cluster (*ftsW* in Thermodesulfobacteria, *murB* and *ddlB* in the FBC superphylum, *ftsA* and *ftsZ* in Chlamydiae and Planctomycetes) [.xlsx file available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder ProCARs].

Taking the rooted phylogenomic tree of Figure 1 as an evolutionary framework and the OGs identified just above as input extant data, we used a new variant of a homology-based reconstruction method (ProCARs)(Perrin et al., 2015) to retrace the evolution of the organization of the *dcw* cluster in our 85 representative organisms. Our reconstruction shows that both the LBCA and the LCA of the Terrabacteria group were organisms featuring a complete 17-gene *dcw* cluster. In contrast, the reconstructed cluster for the ancestor of the TDL group included 16 genes, with the *murA* gene located outside of the cluster (even if present in the genome). Detailed study revealed that the *murA* gene was also outside of the main cluster in every reconstructed ancestor among TDL [Figure 3A]. This gene is at best found on a small sub-cluster, and most of the time it exists as a singleton. An example of such a small sub-cluster reconstructed by ProCARs can be observed in the LCA of the FBC superphylum where *murA* and *murB* are located in tandem. Overall, the *dcw* cluster is conserved in almost all high-level ancestors down to the phyla (see Figure 3A for a summary and .xlsx file available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder ProCARs, for details). This conservation mostly takes the form of a single cluster (e.g., Proteobacteria LCA) or of a limited number of sub-clusters, with the synteny retained within individual sub-clusters (e.g., Chloroflexi LCA, Planctomycetes LCA). Thus, the *dcw* cluster appears as an ancient locus with mainly a history of gene loss or gene delocalization, but likely no gene gain since its establishment before the advent of the LBCA.

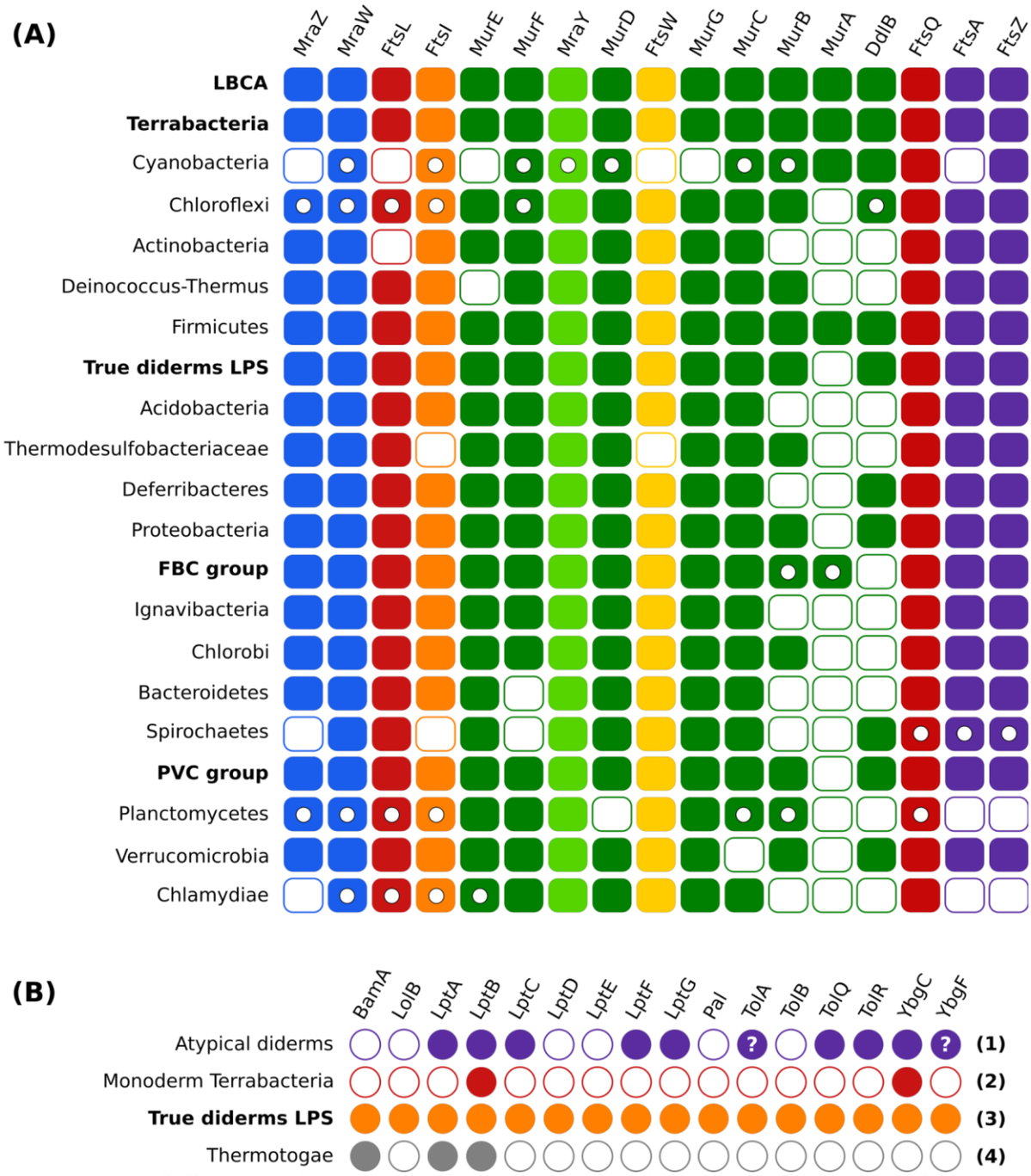


Figure 3: Overview of gene distribution and synteny analyses. (A) ProCARs results for *dcw* cluster organization in selected LCA among Bacteria. Full rectangle = gene present and located in the main cluster; empty circle in rectangle = gene present but located in a sub-cluster; empty rectangle = gene present but outside of any cluster. Note that the reconstruction procedure prevents the complete lack of a gene in an ancestral genome. (B) Recurring distribution patterns at the phylum level for the proteins involved with the OM. Full circle = gene present in the group; empty circle = gene absent in the group; "?" in a circle = potential presence of the gene in the group. Numbers in bold are the pattern numbers. Names written in bold are the names of groups regrouping several phyla.

Phylogenetic trees for the 17 genes of the *dcw* cluster were computed from protein sequences, but these trees are not well resolved ("DCW_17_SG.pdf" available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder Trees). Known phyla can be supported by low to high bootstrap proportions (BP: 9-100%) and PP (0.3-1.0), while the support is always too

low to resolve the relationships between phyla, even though general trends, such as the bipartition between Terrabacteria and TDL (Firmicutes – Chloroflexi – Actinobacteria – Deinococcus-Thermus vs. Proteobacteria – FBC – PVC), are observable in several single-gene trees. Moreover, trees inferred from genes frequently located outside of the *dcw* cluster (e.g., *murC*, *murB* and *ddlB*) are blurrier than those computed from genes retained in the cluster. Finally, the trees of the genes *ftsQ* and *ftsL*, for which the OGs had to be manually reconstructed (see Materials and Methods) are particularly chaotic. In contrast, the *mraY* tree (Figure S12) is better supported (BP: 39-100%; PP: 0.5-1.0) at the phylum level, and is the most congruent with the tree resulting from the 117-gene supermatrix (Figure 1). When concatenated, the *dcw* genes (all but *ftsQ* and ftsL) recover a similar tree (Figure S13), notably featuring the Terrabacteria group, the FBC group and the TDL, but with one exception: the PVC group is split in three, with the Planctomycetes and Verrucomicrobia on one side, the Chlamydia on the other side and the Lentisphaerae within the FBC group. This suggests that the *dcw* cluster mostly experienced a vertical evolution.

5.2.3.4 Evolution of the genes related to the outer membrane

According to our ancestral reconstruction of the cell wall, the LBCA had a single membrane around its cell, which implies that the AD lineages within Terrabacteria (Cyanobacteria, Deinococcus-Thermus and some Firmicutes, i.e., the Halanaerobiales and the Negativicutes) had to acquire their OM independently and in distinct events from the event at the origin of TDL. At face value, this inference might seem less parsimonious than hypothesizing a diderm LBCA and multiple independent OM losses over the evolution of the bacterial domain, as suggested repeatedly (Cavalier-Smith et al., 2020; Megrian et al., 2020; Coleman et al., 2021). To determine whether the OM could indeed have evolved several times independently, we studied the taxonomic distribution of 16 genes involved in OM synthesis and integrity: *bamA*, *lolB*, *lptA*, *lptB*, *lptC*, *lptD*, *lptE*, *lptF*, *lptG*, *pal*, *tolA*, *tolB*, *tolQ*, *tolR*, *ybgC*, *ybgF*. Briefly, BamA is the main protein of the Bam complex (to which the other Bam proteins attach to), which is responsible for the assembly of beta-barrel proteins in the OM (Hagan et al., 2011). LolB is the only OM-anchored protein of the Lol pathway, which delivers lipoproteins to the OM (Silhavy et al., 2010). The Lpt system (LptA to LptG) ensures the transport of the LPS from the cytoplasm to the OM (Bowyer et al., 2011). Finally, the Tol-Pal system (Pal, TolA, TolB, TolQ, TolR, YbgC, YbgF) is involved in the uptake of colicin, the uptake of filamentous bacteriophage DNA and the integrity of the OM (Walburger et al., 2002).

The distribution of these genes was examined across our initial selection of 903 bacterial genomes using curated (Hidden Markov model) HMM profiles built from OGs including *E. coli* reference sequences, and complemented by phylogenetic analyses when orthology was doubtful (see Materials and Methods for details). These results were then summarized at the phylum level to identify recurring patterns of gene distribution (Figure 3B & “OM_genes_presence-hmms.csv” available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder Outer_membrane, for details), while single-gene trees inferred from the corresponding protein sequences are available (“LBCA_OM_16_SG.pdf” available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder Trees). Altogether, our study of the genes encoding the proteins BamA, LolB, the Lpt system and the Tol-Pal system revealed four different patterns of presence/absence in bacterial phyla with diderm organisms. These four gene distribution patterns correspond to: (1) “atypical diderms” (AD) (see references in Table S4), i.e., Cyanobacteria, Deinococcus-Thermus and diderm Firmicutes; (2) “monoderm Terrabacteria” (MT), i.e., Chloroflexi, of which some may be monoderms but all are devoid of LPS (Sutcliffe, 2011; Cavalier-Smith et al., 2020), Actinobacteria,

and monoderm Firmicutes; (3) “true diderms with LPS” (TDL = typical Gram– bacteria); (4) Thermotogae, in which the OM has been replaced by a toga made of structural proteins and polysaccharide hydrolases (xylanases) (Rachel et al., 1988, 1990; Ranjit and Noll, 2016). Below, we briefly comment on these gene distributions from a functional perspective.

First, *bamA* is exclusive to TDL and Thermotogae, even though the latter lack nearly all other OM-related genes studied here. This result suggests a TDL origin for Thermotogae, which are now considered as chimeras partly derived from (or at least related to) Aquificales (Zhaxybayeva et al., 2009; Eveleigh et al., 2013; Bernard et al., 2016). This chimerical nature of Thermotoga is the reason why we did not include them in our phylogenomic tree (see above). Second, *lolB* is exclusive to Proteobacteria, a member of TDL, whereas *lptB* (Lpt system) and *ybgC* (Tol-Pal system) are found in all (or almost every) bacterial phylum of our selection of 903 genomes (including Chloroflexi), and are thus not informative about the origins of the OM. It is likely that these two genes have function(s) outside their respective system, functions that could be unrelated to the OM. This has already been proposed for *ybgF*, which might be part of a protein network involved in phospholipid biosynthesis (Gully and Bouveret, 2006). On the opposite, the LptB protein is known to assemble with LptF and LptG to form an ABC transporter for LPS (Narita and Tokuda, 2009; Bowyer et al., 2011), but the two corresponding genes are apparently lacking in Acidobacteria (TDL), Tenericutes and Chloroflexi. Perhaps surprisingly, this is also the case for Actinobacteria, these monoderm bacteria further sharing their whole gene distribution pattern with Chloroflexi.

Beyond *lptB* and *ybgC*, the Lpt and Tol-Pal systems are found in both AD and TDL but to a different extent. Indeed both systems are present in AD, albeit only in a largely reduced form, whereas in TDL, they range from a largely reduced form (e.g., Chlamydiae or Planctomycetes) to a (almost) complete form (e.g., Proteobacteria or Bacteroidetes), and this distribution is phylum-specific (Figure 3B). Hence, two genes from each system are only present in (most) TDL genomes, *lptD* and *lptE* on one side, *pal* and *tolB* on the other side, whereas all four genes are never found in AD genomes. Regarding *tolA* and *ybgF*, they may or may not be exclusive to TDL, depending on the biological reality of their scarce occurrence in some organisms belonging to AD (Firmicutes for *tolA* and Cyanobacteria for *ybgF*). Based on our trees of the corresponding proteins, the dubious sequences (denoted by “?” in Figure 3B and by stars in “OM_genes_presence-hmms.csv” available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder Outer_membrane) are sisters to Bacteroidetes (member of TDL) in both cases, plus one case with a sequence sister to *Moraxella* in *tolA* tree (Figures S14 and S15, see also “LBCA_OM_16_SG.pdf” available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder Trees). Therefore, provided they are not the product of genome contamination (Cornet et al., 2018), these genes are unlikely to have been vertically inherited.

From a functional point of view, the genes retained by AD for the Lpt system (*lptA*, *lptB*, *lptC*, *lptF* and *lptG*) are involved in the transport of the LPS from the cytoplasm to the OM and thus are not directly associated to the OM itself, contrarily to *lptD* and *lptE*, which form a complex at the OM that may serve as the recognition site for the LPS (Wang et al., 2014). Similarly, for the Tol-Pal system, AD genomes lack *pal* and *tolB*, two genes encoding proteins located in the periplasm and therefore directly associated to the OM (Rigal et al., 1997; Ray et al., 2000). Overall, the Lpt and Tol-Pal systems in AD are thus restricted to components that might have a function in the absence of an OM.

Interestingly, the genes of the Tol-Pal system are clustered in most genomes of Proteobacteria and Chlorobi, as well as in the lone genomes we studied within Fibrobacter and Gemmatimonadetes, and sporadically in those of Verrucomicrobia and Acidobacteria (available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder Outer_membrane sub-folder synteny_output). Since all these lineages belong to TDL, we cannot exclude that the conservation of the Tol-Pal cluster appears patchier than it really is, owing to uneven levels of genome assembly. In contrast, the genes of the Lpt system are not clustered in any of the genomes examined, except in Proteobacteria, where five of the seven genes are grouped on two loci (*lptFG* and *lptABC*) (available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder Outer_membrane sub-folder synteny_output).

5.2.4 Discussion

The nature of the LBCA is unknown, especially the architecture of its cell wall. The lack of reliably affiliated bacterial fossils outside Cyanobacteria (Demoulin et al., 2019) makes it elusive to determine the very nature of the LBCA. Nevertheless, phylogenomic inference leads to informative results, and our analysis of the cell-wall characteristics of extant bacteria, combined with ancestral state reconstruction and distribution of key genes, opens interesting possibilities: the LBCA might reasonably have been a monoderm bacterium featuring a complete 17-gene *dcw* cluster, two genes more than in the model *E. coli* cluster.

As diderm bacteria are not monophyletic, whatever the root retained for the bacterial domain, our reconstruction of a monoderm LBCA implies that the diderm character state has appeared several times, which goes against the principle of parsimony commonly invoked in such matters (Cavalier-Smith et al., 2020). Indeed, acquiring an OM is more than a simple mutation: it requires the acquisition of a whole new complex system. This makes the “monoderm-first” result counter-intuitive to the opposite of the alternative, widely held, educated guess, “diderm-first” hypothesis (Cavalier-Smith et al., 2020; Megrian et al., 2020; Taib et al., 2020; Coleman et al., 2021). Yet, our results are model-based, congruent across different roots and models and robust to a heavily biased hyperprior towards the diderm-first hypothesis. It contrasts with other recent studies, which do not rely on probabilistic models (Cavalier-Smith et al., 2020; Megrian et al., 2020) and conclude to a diderm LBCA, based on qualitative considerations. That being said, the diderm-first view has also been supported in the recent work of (Coleman et al., 2021). The latter study features a reconciliation tree and infers the diderm state of the LBCA based on the genes involved in the LPS synthesis and the flagellar subunits, notably PilQ, which is part of the Type IV pili. While the approach of Coleman and co-workers is also model-based, it differs from ours by first inferring the gene catalogue of the LBCA and then deducing its cell-wall architecture, whereas we directly infer the LBCA architecture and then study the underlying gene distribution patterns to corroborate our inference. It is of note that the Type IV pili is also present in monoderm bacteria (Melville and Craig, 2013), thus its presence does not automatically entail the inference of a diderm LBCA.

Hence, following a bibliographic search for proteins with functions exclusive to diderms (without distinguishing between diderms with and without LPS), we identified 16 candidates: BamA, which is a part of a complex assembling the proteins in the OM (Hagan et al., 2011), LolB, which is part of the proteins fixing the LPS to the OM (Silhavy et al., 2010), the Lpt proteins, which serve as a transport chain from the inner, i.e., cytoplasmic (Baurain et al., 2016), membrane (IM) to the OM (Bowyer et al., 2011), and the Tol-Pal system, the exact function of which is still unknown but important to the integrity of the OM (Walburger et al., 2002). Then, we studied the distribution of

the 16 corresponding genes in 903 broadly sampled bacterial genomes. Four recurring patterns of OM gene distribution were identified (Figure 3B): 1) AD (for atypical diderms: Deinococcus-Thermus and Cyanobacteria and diderm Firmicutes), 2) MT (for monoderm Terrabacteria: Actinobacteria, Chloroflexi and monoderm Firmicutes), 3) TDL (for true diderms-LPS), and 4) Thermotogae. Thermotogae have chimerical genomes (Zhaxybayeva et al., 2009) and are clearly derived with respect to other bacteria; thus, their cell-wall architecture is of secondary origin. This is why we do not elaborate further on their case. For similar reasons, the atypical cell-wall of the Corynebacteriales (an order of the Actinobacteria phylum) is not considered in this work. Indeed, Corynebacteriales are located deeply within Actinobacteria (Verma et al., 2013), which again implies a secondary origin for their peculiar cell-wall architecture.

From these patterns, it appears that even MT share some genes involved with the OM despite their lack of an OM. It implies that these genes provide at best circumstantial evidence concerning the presence or the absence of an OM. Thus, solely relying on their detection to infer the presence of an OM would be hazardous. In the study of (Coleman et al., 2021), the authors build upon two types of genes to justify their inference of a diderm LBCA: the genes involved with the LPS synthesis and the genes involved with the pili type IV. However, our results show that the mere presence of LPS genes is an unreliable feature to infer the presence of an OM, given that even monoderm bacteria can carry some of them. Similarly, the study of (Melville and Craig, 2013) shows that the type IV pili is not exclusive to the diderm bacteria. Therefore, the inference of a diderm LBCA by Coleman et al. is based on genes that only provide ambiguous evidence for the OM.

Pattern 2 shows that Chloroflexi shares the same gene distribution as MT, despite being mostly considered as diderms (3 out of 4) in our reconstruction of the cell wall. Currently, there is still debate on whether Chloroflexi are monoderm or diderm organisms (Sutcliffe, 2011). The fact that they share the same OM gene distribution pattern as MT is a clue in favor of Chloroflexi having only one membrane too. In this case, our reconstruction of the LBCA's cell wall would have had a small bias towards the diderm state and, in spite of that unwarranted handicap, we still recovered the LBCA as a monoderm bacterium. In our opinion, this result can be taken as additional evidence for a genuinely strong signal for a monoderm LBCA.

Patterns 1, 2 and 3 may be arranged following a gradual complexification, with pattern 2 being the simplest, pattern 1 the intermediate and pattern 3 the most complex. The study of the functions of the proteins characterizing the different patterns reveals that pattern 3 is the only one including proteins directly involved with the OM (i.e., linked to the OM), whereas pattern 1 only includes proteins indirectly involved with the OM (i.e., linked to the IM or interacting with the IM or located in the cytoplasm) and pattern 2 only includes proteins indirectly involved with the OM and located in the cytoplasm. Although we know (some of) the OM pathways functioning in TDL, regarding AD, we only identified the common parts between their pathways and TDL pathways. The rest of the TDL pathways should have an equivalent in the AD pathways but our approach by candidate genes did not allow us to identify them. This hints at the possibility of a different evolution from a common base, since some of the functions performed by the genes present in pattern 3 (TDL), but absent in pattern 1 (AD) should be carried out in one way or another (e.g., the maintenance of the OM or the OM invagination during cell division) (Yakhnina and Bernhardt, 2020). In this case, the common base would be the partial (primitive?) Lpt and Tol-Pal systems, and at least two different systems for handling the OM would have built upon it, respectively in TDL and (all or some) AD.

On the other hand, if the LBCA was a diderm, then extant monoderms would have been the result of several independent secondary simplifications. Consequently, the monoderms dispersed within the Terrabacteria group would share the same origin, a diderm ancestor, but would not necessarily end up with the same remaining genes after their respective simplification. Yet, they all display the same single pattern (pattern 1). Furthermore, based on single-gene trees, some OM genes found in AD genomes (e.g., LptF and LptG) might stem from horizontal transfer from some of the TDL genomes, rather than through vertical inheritance from a diderm LBCA ancestor. However, because most of these trees are poorly resolved (despite good multiple sequence alignments), the evidence is weak. Besides these patterns show that the TDL group is different from every other diderm, indicating that the relatively homogeneous TDL group is monophyletic, as suggested by phylogenomic trees. If correct, the bacterial root cannot lie within TDL and, as already mentioned, a root on (or within) Terrabacteria implies that the diderm cell-wall architecture appeared at least on two different occasions. The latter inference is necessary to account for diderms other than TDL in Firmicutes, Cyanobacteria, Chloroflexi and Deinococcus-Thermus, which then raises the issue of how the LPS is transported from the IM to the OM for these AD nested within Terrabacteria. Indeed, they do not share the same Lpt system as TDL since theirs is “reduced”, so they must have developed another system grafted (or not) onto the “reduced” Lpt system.

Another clue that might confirm our reconstruction is that the rare organisms amongst the CPR (Candidate Phylum Radiation, also known as Patescibacteria (Rinke et al., 2013; Parks et al., 2017)) to have been described feature a monoderm cell-wall architecture (Luef et al., 2015). In several trees including the CPR (with the Archaea used as the outgroup), these are the first to diverge from the other bacteria, while the remaining of those trees have the same structure as ours (Hug et al., 2016; Castelle and Banfield, 2018). However, in (Coleman et al., 2021), the CPR subtree is located within the Terrabacteria with strong support. Consequently, depending on the accepted topology, the CPR could either provide another (small) clue for a monoderm LBCA (CPR at the base of the bacterial tree) or only for a monoderm ancestor for the Terrabacteria group (CPR within the Terrabacteria group). Nonetheless, as most CPR genomes still lack detailed reliable information about the cell-wall architecture of the corresponding organisms, there was no point adding them to our study for now.

When it comes to the reconstruction of the *dcw* cluster, the LBCA is inferred as featuring a complete 17-gene cluster. This complete cluster has probably been vertically transmitted since then and often subject to parallel reduction, either by escape of one or several genes from the cluster or by disappearance of those genes from the genome. Since it is shared by both monoderm and diderm organisms, the *dcw* cluster does not provide a clue about the issue of the number of membranes of the LBCA. However, it confirms that the LBCA had a cell wall with a peptidoglycan layer, even if it does not inform on its original thickness.

In TDL and Terrabacteria, the *murA* gene is (almost) always absent from the main *dcw* cluster. In Firmicutes, which are at the base of Terrabacteria, this gene is nevertheless considered located within the cluster by our reconstruction, as this is the situation observed for five (out of nine) genomes from our selection of 85 representatives. The gene is also found in sub-clusters distributed relatively patchily across Cyanobacteria, Firmicutes, Epsilon-proteobacteria, Elusimicrobia, *Caldithrix abyssi*, *planctomycete* KSU1, and *Lentisphaera araneosa*. Both extant and reconstructed ancestors show that TDL have excised their *murA* from the main cluster after diverging from Terrabacteria, whereas Terrabacteria kept it longer in the main cluster. However, *murA* is found located on sub-clusters in both groups.

For the moment, there is no scenario to explain the appearance of the OM in the lineage leading to TDL, but such a scenario exists for the appearance of diderms in Firmicutes: it is the failed endospore origin (Dawes, 1981; Tocheva et al., 2011; Vollmer, 2011; Errington, 2013) According to this hypothesis, an ancestral monoderm endospore former would have experienced a failed sporulation, thereby locking the endospore within the cell while never finishing the spore. With time, it would have become a diderm bacteria. Indeed, during the course of sporulation, the prespore engulfed in the bacterial mother cell actually possesses two membranes. A thin layer of the mother peptidoglycan subsists between these membranes before the cortex is added around the prespore between this small layer and the OM. Although not yet a diderm-LPS architecture, a cortex-less spore could represent a starting point for the emergence of diderm bacteria in the specific case of Firmicutes. In 2016, Tocheva (Tocheva et al., 2016) amended the model by arguing that this founding event would have taken place in an ancestor not only to diderm Firmicutes but to all diderm bacteria. Regarding the origin of the OM in AD other than Firmicutes, we have already mentioned that Chloroflexi might actually be monoderms, based on their shared pattern (pattern 2) with MT. This leaves us with Cyanobacteria and Deinococcus-Thermus, along with the large TDL group. Since pattern 3 looks like a complexification of pattern 1, the origin of didermia in TDL might come from one of these AD phyla by horizontal gene transfer of OM genes, followed by complexification in an ancestor of TDL. Alternatively, TDL ancestors might have transferred OM genes to distinct ancestors of AD phyla, thus in the opposite direction. At this stage, this remains an open question because of the lack of resolution of the corresponding single-gene trees, which prevents any definitive answer.

5.2.5 Conclusion

Our results show that the LBCA was, against our intuition, a monoderm bacteria with a thick peptidoglycan layer. The reconstruction of the *dcw* cluster adds a strong hint towards an LBCA with a peptidoglycan layer but does not discriminate between a thick and a thin peptidoglycan layer. Concerning our study of the OM genes, their distribution suggests that indeed a monoderm ancestor is possible but the evidence is not decisive. Yet, further improving our results using the same methodology would require a more accurate description of the cell-wall architecture of the extant organisms, notably the presence or absence of the LPS, an information which is often lacking. Moreover, we observe that some OM genes involved with the precursors of the LPS synthesis are even present in genomes of bacteria that does not have LPS on their OM (or even an OM), thus relying solely on the presence of specific genes to determine the presence or absence of LPS is not adequate.

5.2.6 Materials and Methods

5.2.6.1 Dataset assembly

5.2.6.1.1 Data download

The initial dataset of prokaryotic genomes and proteomes was downloaded from Ensembl Bacteria release 20 (Kersey et al., 2014). This dataset contained 8848 Bacteria and 238 Archaea.

5.2.6.1.2 Genome dereplication and selection

We first reduced the number of genomes based on genomic signatures (Moreno-Hagelsieb et al., 2013) to regroup similar genomes into genome clusters with a prerelease version of our new

software ToRQuEMaDA (Léonard et al., 2021). Briefly, for five different *k*-mer sizes (from 2 to 6-nt), we computed the frequency of each word in each genome using the program compseq from the EMBOSS software package (Rice et al., 2000). The complete lineage of every genome was recovered from the NCBI Taxonomy database (Sayers et al., 2011) using the program fetch-tax.pl from the Bio::MUST::Core distribution (D. Baurain, https://metacpan.org/dist/Bio-MUST-Core). Each signature file was further analyzed in R (R Core Team, 2013) to cluster genomes into a predefined number of groups (300, 600, 900, 1200, 1500 and 2100) using various distance metrics (i.e., Euclidean, Pearson and Hamming) and clustering algorithms (i.e., *k*-means, ascending and descending hierarchical clusterings). In order to choose the best combination of methods and parameters, the available taxonomic information was used to evaluate the quality of the clustering. Briefly, we computed how many different taxa of each rank (phylum, class, order, family, genus, species) were found in each individual cluster of each set of clusters, and chose the combination that best separated the higher-level taxa (phylum, class, order, family) while merging the lower-level taxa (genus, species) (Léonard et al., 2021). This led us to settle on the following set of methods and parameters: 6-nt *k*-mer, 900 clusters, Pearson distance and ascending hierarchical clustering algorithm. Then, we selected a single representative for each cluster, based on the quality of genome annotations, as evaluated by the number of gene names devoid of uninformative words like “hypothetical”, “putative”, “unknown” etc (Léonard et al., 2021). After including a few additional well-characterized genomes (e.g., *Streptomyces coelicolor* A3(2), *Escherichia coli* O127:H6 str. E2348/69, *Staphylococcus aureus* subsp. *aureus* MRSA252), we ended up with a list of 903 genomes: 822 Bacteria and 81 Archaea.

5.2.6.1.3 Identification of orthologous groups

For every protein sequence of every one of these 903 genomes, we launched an all-versus-all BLAST-like similarity search using USEARCH v7.0.959 (Edgar, 2010) with the following parameters (evalue = 1e-5; accel = 1; threads = 64). Then, we used OrthoMCL v2.0.3 (Li et al., 2003) to cluster protein sequences into orthologous groups (OGs) based on USEARCH reports, using an e-value cut-off of 1e-5, a similarity cut-off of 50% and an inflation parameter of 1.5. The total number of proteins for the 903 genomes was 2,467,263, and these were partitioned into 124,422 OGs, whereas 326,269 sequences were considered as “singletons” by OrthoMCL (i.e., without homologues).

5.2.6.1.4 Database creation

Gene metadata (organism, genomic coordinates, strand, putative function) for every protein was extracted from the definition lines of the Ensembl FASTA files and stored into a custom designed MySQL (Oracle Corporation) relational database (see Figure S16), along with orthology relationships, based on our protein sequence clustering.

5.2.6.2 Evolution of the bacterial domain

5.2.6.2.1 Supermatrix assembly

To build a robust tree of the bacterial domain, we manually chose a subset of 85 genomes (out of the 903 genomes initially selected), trying to maximise the number of classes. Then, using classify-mcl-out.pl (Van Vlierberghe et al., 2021), we selected all OGs of proteins featuring at least one representative of eight major bacterial phyla (Firmicutes, Chloroflexi, Actinobacteria, Deinococcus-Thermus, Proteobacteria, Spirochaetes, Planctomycetes and Bacteroidetes) and in which at most 10% of the selected genomes contained more than one gene copy. This left us with a list of 176 broadly conserved and (mostly) single-copy genes. The final dataset was further

reduced to 117 OGs to ensure a maximum of 14 missing species in each individual OG (Table S5). The corresponding OGs were aligned with MAFFT v7.127b (Katoh and Standley, 2013) using default parameters. The protein sequence alignments were then filtered with Gblocks v0.91b (Castresana, 2000) using a set of “medium stringency” parameters (as predefined in Bio::MUST::Core) and concatenated with SCaFoS v1.30k (Roure et al., 2007). Finally, the resulting concatenation was further filtered for sites >50% missing character states, yielding a supermatrix of 85 species x 19,959 unambiguously aligned amino-acid (AA) positions (4.29% missing character states). A preliminary (more diverse) supermatrix was also created in the process, including 101 species and 19,959 unambiguously aligned AA positions (4.72% missing states).

5.2.6.2.2 Phylogenomic analyses

For Bayesian inference (BI), we used PhyloBayes MPI v1.5 (Lartillot et al., 2013) to produce six replicate MCMC chains of 50,000 cycles, with one tree sampled every 10 cycles, using the CAT+GTR+Γ model of sequence evolution (Lartillot and Philippe, 2004, 2006; Lartillot et al., 2007). Constant sites were deleted with the -dc option. Convergence was assessed using the program tracecomp from the PhyloBayes software package. Two consensus trees (along with their PP) were extracted after a burn-in of 10,000 cycles: one over the six chains (A to F) and another over the two most congruent chains (A and C; maxdiff = 0.130; meandiff = 0.001), both with the -c option of bpcomp set to 0.01. Cross-validation tests to determine the best-fit model (CAT+GTR+Γ) were carried out using PhyloBayes v3.3f (Lartillot et al., 2009), as suggested in PhyloBayes manual (page 38). For our preliminary tree, we ran two chains of 50,000 cycles, with one tree sampled every 10 cycles, under the simpler CAT+Γ model. The consensus tree was extracted after a burn-in of 5,000 cycles (maxdiff = 0.580; meandiff = 0.011). All trees (including those described below) were formatted semi-automatically using the scripts format-tree.pl, export-itol.pl and import-itol.pl (also from Bio::MUST::Core) and iTOL v6 (Letunic and Bork, 2021).

5.2.6.2.3 Congruence tests

Congruence tests were performed on the 85-species supermatrix genes with Phylo-MCOA v1.4 (De Vienne et al., 2012), then Maximum Likelihood (ML) reconstruction with RAxML v8.1.17 (Stamatakis, 2014) was used under the model PROTGAMMALGF (LG+F+Γ) to compare the topologies obtained with and without the “cell-by-cell outliers” (i.e., specific species in specific genes whose position is not concordant with their position in the other gene trees) identified by Phylo-MCOA.

5.2.6.3 Evolution of the cell-wall

5.2.6.3.1 Cell-wall architecture of extant organisms

For each one of the 85 bacterial species, a dedicated survey of the literature was conducted (Table S4). When no information about the cell-wall architecture was available at the species level, we searched at a higher taxonomic level, sometimes up to the phylum. Based on the collected data, we summarized the cell-wall architecture using two different traits: the number of membranes and the presence and thickness of the peptidoglycan layer (Table S3). For the membrane trait, we used the following binary coding: 0 for one membrane and 1 for two membranes, whereas for the peptidoglycan trait, we used three different states: 0 for no peptidoglycan, 1 for a thin peptidoglycan and 2 for a thick peptidoglycan. Cell-wall trait analyses were then performed using BayesTraits V3 (Pagel et al., 2004; Pagel and Meade, 2015; Meade and Pagel, 2017). For *Parachlamydia acanthamoebae*, no indication about peptidoglycan thickness was found, so this trait was coded as “12”, following the suggestion in BayesTraits manual (page 9).

5.2.6.3.2 Correlation between cell-wall traits

Correlation between cell-wall traits was tested by comparing the discrete independent and discrete dependent models using Bayes Factors (BF), as described in BayesTraits manual (page 13). We applied the stepping stone sampler, using 100 stones with 10,000 iterations per stone. As this procedure only allows for the comparison of two binary traits, and since our peptidoglycan trait had three possible states, we had to combine two different states into a single state. Three different combinations were tested to check the robustness of the correlation. For case A, the absence of peptidoglycan was coded as 0 and the presence of peptidoglycan (either thin or thick) as 1. For case B, both the absence of peptidoglycan and the thin peptidoglycan were coded as 0, while the thick peptidoglycan was coded as 1. For case C, both the absence of peptidoglycan and the thick peptidoglycan were coded as 0, while the thin peptidoglycan was coded as 1. Because *P. acanthamoebae* is a Chlamydiae, which belong to the diderm-LPS group, its undocumented peptidoglycan layer (see above) was considered as thin when recoding the peptidoglycan trait.

5.2.6.3.3 Ancestral state reconstruction of cell-wall traits

For ancestral state reconstruction, the two traits were considered separately. We used the Bayesian phylogenomic tree rooted on Terrabacteria as an input tree, and further checked the robustness of our inferences to five alternative roots, all within Terrabacteria. Branch lengths were scaled to have a mean of 0.1, as suggested in BayesTraits manual (page 10). Five different MultiState models were tested: prior exponential of 10 (model “E”), hyperprior exponential 0 to 10 (model “H1”), hyperprior exponential 0 to 100 (model “H2”), reverse-jump hyperprior exponential 0 to 10 (model “R1”), and reverse-jump hyperprior exponential 0 to 100 (model “R2”). Reversible-jump models had the opportunity to forbid some transitions (rate = 0) and/or to equate distinct rates. 10 MCMC chains were run for each combination of trait/root/model for 1,100,000 cycles, with one sample saved every 1000 cycles, and burnin set at 100,000 cycles. State probabilities and transition rates were summarized as means of the 10 x 10,000 samples. To investigate the sensitivity of the Bayesian inference of a monoderm LBCA to priors, one additional analysis (biased on purpose towards reversion from diderm to monoderm state) was re-run as 100 MCMC chains with q01 and q10 exponential hyperpriors set to 0 to 1 for and 1 to 10, respectively.

5.2.6.3.4 Comparison of the selected models

Building on the stepping stones sampler files produced by the BayesTraits ancestral state reconstruction, we compared the fit of our five models (in a systematic pairwise fashion) to both the membrane peptidoglycan data using Bayes Factors. We selected the stepping stones files from the runs with the tree rooted on the Terrabacteria. As above, the stepping stone sampler used 100 stones with 10,000 iterations per stone.

5.2.6.4 Evolution of the *dcw* cluster

5.2.6.4.1 Synteny analyses of extant genomes

To study the gene order of the *dcw* cluster across our 903 genomes, we developed a custom R script. This interactive interface allows us to select any subset of genomes and to focus on any region of the bacterial chromosome chosen as the reference genome for the comparison. To maximize the robustness of these analyses, the data (genomic coordinates, orthology relationships, functions) required for the visualization are fetched in real-time from the relational database. Examples of graphical outputs produced by this program (limited to the 85 final organisms) are shown in “synteny_85_dcw.pdf” available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder ProCARs. The OGs corresponding to the genes of the *dcw* cluster were identified by a combination of homology searches using reference protein sequences as queries and our R interface for visual confirmation of synteny conservation.

In most cases but the poorly conserved *ftsL* and *ftsQ*, a single OG was identified for each gene. For *ftsL* and, to a much lesser extent, *ftsQ*, several OGs had to be merged, based on the presence of an unidentified gene sequence at their respective expected location, i.e., between *mraW* and *ftsI* for *ftsL*, and immediately before *ftsA* for *ftsQ*. Moreover, HMM profiles (pHMM) (Eddy, 2011; Mistry et al., 2013) (see also below) were built from unambiguous reference sequences to ensure proper identification of *ftsL* and *ftsQ* genes in genomes with a fragmented *dcw* cluster. Overall, *ftsL* and *ftsQ* were spread over 36 and 24 OGs (many containing only 2-3 sequences), respectively, whereas *mraW*, *mraZ* and *ftsA* were spread over 2, 3 and 4 OGs, respectively.

5.2.6.4.2 Ancestral gene order reconstruction

To reconstruct the evolution of the *dcw* cluster, we used the program ProCARs (Perrin et al., 2015), modified to prevent gene inversions in the cluster (by enabling the -p option). ProCARs input files were built semi-automatically from the relational database, focusing on the 85 bacterial species retained in our phylogenomic analyses and informed by synteny analyses of extant genomes. Briefly, genes too far from other genes were encoded as lying on different "chromosomes" by introducing artificial telomeres. When several "orthologous" genes were available in a given genome for a specific gene, we first tried to select the gene copy lying on the artificial "chromosome" with the highest count of other *dcw* genes. If this failed due to ties, we turned to the gene copy located on the main DNA molecule (genuine chromosome or largest scaffold in the genome assembly); otherwise, as a last resort, we selected the gene copy in the same orientation as the *dcw* genes found on the genuine chromosome or largest scaffold. Finally, when two gene copies were in tandem, we considered them as a single (duplicated) gene for the purpose of the ancestral reconstruction.

5.2.6.4.3 Phylogenetic analyses

For the single-gene analyses of the *dcw* cluster in the 85 genomes of interest, we used the 17 identified OGs (possibly consolidated; see above) to produce trees according to two different approaches: (1) by ML using RAxML v8.1.17 under the PROTGAMMALGF (LG+F+Γ) model and (2) by BI using PhyloBayes v3.3f under the model GTR+C60+Γ, with two MCMC chains run for 10,000 cycles, with burnin of 5000 cycles and sampling every 10 cycles. Convergence was assessed as above (gene maxdiff's ranging between 0.208 and 1.000 and meandiff's between 0.013 and 0.062), with the -c option of bpcomp set to 0.25, which turned unresolved nodes to multifurcations. Then, a concatenation of 15 of the 17 genes of the *dcw* cluster was built using SCaFoS v1.30k, leaving out *ftsL* and *ftsQ* due to their poor conservation (see above). For these 15 genes, additional steps were carried out to ensure the orthology of the concatenated sequences. Briefly, we used our ProCARs input to select only the genes belonging to the *dcw* cluster (or sub-cluster) in each genome. Orthologues not supported by synteny evidence were removed from the alignments using prune-ali.pl (also from Bio::MUST::Core) before concatenation. We further filtered out sites with ≥50% missing character states, thereby yielding a sparser supermatrix of 85 species x 4571 AAs (8.47% missing character states). PhyloBayes MPI v1.4 was used to run two chains under the CAT+Γ model for 50,000 cycles. We chose a burnin of 10,000 cycles and kept only one sample every 10 cycles of the remaining 40,000 cycles. We selected both chains to compute the tree (maxdiff = 0.284; meandiff = 0.007), with the -c option of bpcomp set to 0.25. All trees were formatted as above.

5.2.6.5 Evolution of the genes related to the outer membrane

5.2.6.5.1 Homology searches in complete proteomes

For our broader study of the taxonomic distribution of 16 genes involved in synthesis of the OM across the 903 selected genomes, we did not rely on synteny as those were not part of a single

cluster in any organism. Instead, we searched for the OGs containing unambiguous reference sequences for these genes. For each set of OGs (from 1 to 9) potentially corresponding to a gene of interest, we computed an alignment over all sequences with MAFFT v7.453 (using the accurate LINSI strategy) and checked by eye if it was globally satisfactory or not, possibly after cleaning up a few divergent sequences. If the alignment was good enough, we built an HMM profile from it to search the complete proteomes of our 903 genomes using HMMER (Eddy, 2011; Mistry et al., 2013). Then, based on the E-value, length, pHMM profile coverage, copy number and taxonomy of the HMMER hits, we selected the probably orthologous proteins using the visual software Ompa-Pa (A.R. Bertrand and D. Baurain; available at https://metacpan.org/dist/Bio-MUST-Apps-OmpaPa). In contrast, when the alignment of all sequences was too poor, we focused on the original OG containing the *E. coli* sequence and tried to build a profile by adding up to 6 (for *lolB* and *lptC*) of the additional OGs using an iterative strategy as implemented in the software Two-Scalp (A.R. Bertrand and D. Baurain; available at https://metacpan.org/dist/Bio-MUST-Apps-TwoScalp). Then, we followed the same route as if the pHMM had been computed from a “good-enough” alignment.

5.2.6.5.2 Taxonomic and phylogenetic analyses

For each gene of the 16 genes, we retrieved the list of genomes having provided the (probably) orthologous proteins and tabulated the corresponding organisms at the phylum level. From these numbers, we tried to identify recurring patterns of gene distribution. For two genes, *tolA* and *ybgF*, the taxonomic distribution was discordant with respect to other genes (when present) in the AD group. In each case, only one of the expected phyla of the AD group had at least a copy, and this phylum was represented by a noticeably lower number of sequences compared to other genes present in the AD group (when they possessed copies of the gene). To determine if these discordances were due to genome contamination or very recent gene transfers, we aligned the sequences with MAFFT v7.453 (LINSI) and computed two phylogenetic trees using RAxML v8.1.17 under the PROTGAMMALGF (LG+F+Γ) model. Trees were also produced for the 14 other genes associated with the OM following the same method. All trees were formatted as above, with unresolved nodes (BP < 25%) turned to multifurcations.

5.2.7 Author contributions

Raphaël R. Léonard performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, developed the dereplication tool and the synteny tool, and approved the final draft. Eric Sauvage, Frédéric Kerff and Denis Baurain conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper and approved the final draft. Valérian Lupo prepared figures, reviewed drafts of the paper and approved the final draft. Amandine Perrin developed a specific version of ProCARs for this study, reviewed drafts of the paper and approved the final draft. Damien Sirjacobs provided technical support for the high-performance computing cluster, reviewed drafts of the paper and approved the final draft. Paulette Charlier reviewed drafts of the paper and approved the final draft.

5.2.8 Acknowledgments

This work was supported in part by the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister’s Office, Science Policy programming (IAP no. P6/19). It also benefited from computational resources made available on the Tier-1 supercomputer of the Fédération Wallonie-Bruxelles, infrastructure funded by the Walloon Region

under the grant agreement n°1117545, and on the "*durandal*" grid computer funded by three grants from the University of Liège, "Fonds spéciaux pour la recherche", "Crédit de démarrage 2012" (SFRD-12/03 and SFRD-12/04) and "Crédit classique 2014" (C-14/73). R.R.L. and V.L. are the recipients of FRIA (Fonds de la Recherche pour l'Industrie et l'Agriculture) fellowships (F.R.S.-FNRS, Brussels, Belgium). F.K. is a research associate of the F.R.S.-FNRS, Belgium. We thank D. de Vienne for his advice on the interpretation of Phylo-MCOA output and Nguyen The Anh for preliminary trees of the genes in the *dcw* cluster.

5.2.9 References

Antunes, L. CS., Poppleton, D., Klingl, A., Criscuolo, A., Dupuy, B., Brochier-Armanet, C., et al. (2016). Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the Firmicutes. *eLife* 5, 1–21. doi:10.7554/eLife.14589.

Bapteste, E., Boucher, Y., Leigh, J., and Doolittle, W. F. (2004). Phylogenetic reconstruction and lateral gene transfer. *Trends Microbiol.* 12, 406–411. doi:10.1016/j.tim.2004.07.002.

Barloy-hubler, F., Lelaure, V., Galibert, F., Génétique, L., Léon, P., and Cedex, R. (2001). Ribosomal protein gene cluster analysis in eubacterium genomics : homology between Sinorhizobium meliloti strain 1021 and Bacillus subtilis. 29, 2747–2756.

Battistuzzi, F. U., and Hedges, S. B. (2009). A major clade of prokaryotes with ancient adaptations to life on land. *Mol. Biol. Evol.* 26, 335–343. doi:10.1093/molbev/msn247.

Baurain, D., Wilmotte, A., and Frère, J.-M. (2016). Gram-Negative Bacteria:" Inner" vs." Cytoplasmic" or" Plasma Membrane": A Question of Clarity rather than Vocabulary. *J. Microb. Biochem. Technol.* 8, 325–326.

Beiko, R. G., Harlow, T. J., and Ragan, M. a (2005). Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 14332–14337. doi:10.1073/pnas.0504068102.

Bernard, G., Chan, C. X., and Ragan, M. A. (2016). Alignment-free microbial phylogenomics under scenarios of sequence divergence , genome rearrangement and lateral genetic transfer. *Nat. Publ. Group*, 1–12. doi:10.1038/srep28970.

Bhandari, V., Naushad, H. S., and Gupta, R. S. (2012). Protein based molecular markers provide reliable means to understand prokaryotic phylogeny and support Darwinian mode of evolution. *Front. Cell. Infect. Microbiol.* 2, 1–14. doi:10.3389/fcimb.2012.00098.

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. 18, 1–18. doi:10.1186/1471-2148-8-272.

Bowyer, A., Baardsnes, J., Ajamian, E., Zhang, L., and Cygler, M. (2011). Characterization of interactions between LPS transport proteins of the Lpt system. *Biochem. Biophys. Res. Commun.* 404, 1093–1098. doi:10.1016/j.bbrc.2010.12.121.

Castelle, C. J., and Banfield, J. F. (2018). Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* 172, 1181–1197.

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.* 17, 540–552. doi:10.1093/oxfordjournals.molbev.a026334.

Cavalier-Smith, T. (1987). The Origin of Eukaryote and Archaebacterial Cells. *Ann. N. Y. Acad. Sci.* 503, 17–54. doi:https://doi.org/10.1111/j.1749-6632.1987.tb40596.x.

Cavalier-Smith, T. (2006). Rooting the tree of life by transition analyses. *Biol. Direct* 1, 19. doi:10.1186/1745-6150-1-19.

Cavalier-Smith, T. (2010). Deep phylogeny, ancestral groups and the four ages of life.

Philos. Trans. R. Soc. B Biol. Sci. 365, 111–132.
Cavalier-Smith, T., Ema, E., and Chao, Y. (2020). Multidomain ribosomal protein trees and the planctobacterial origin of neomura (eukaryotes, archaebacteria). *Protoplasma*, 1–133.
Coico, R. (2006). Gram staining. *Curr. Protoc. Microbiol.*, A-3C.
Coleman, G. A., Davín, A. A., Mahendrarajah, T. A., Szánthó, L. L., Spang, A., Hugenholtz, P., et al. (2021). A rooted phylogeny resolves early bacterial evolution. *Science* 372. doi:10.1126/science.abe0511.
Cornet, L., Meunier, L., van Vlierberghe, M., Léonard, R. R., Durieu, B., Lara, Y., et al. (2018). Consensus assessment of the contamination level of publicly available cyanobacterial genomes. 1–26.
Dawes, I. W. (1981). Sporulation in evolution. *Mol. Cell. Asp. Microb. Evol. Camb. Univ. Press N. Y.*, 85–130.
De Vienne, D. M., Ollier, S., and Aguileta, G. (2012). Phylo-MCOA: A fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. *Mol. Biol. Evol.* 29, 1587–1598. doi:10.1093/molbev/msr317.
Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361–375. doi:10.1038/nrg1603.
Demoulin, C. F., Lara, Y. J., Cornet, L., François, C., Baurain, D., Wilmotte, A., et al. (2019). Cyanobacteria evolution: Insight from the fossil record. *Free Radic. Biol. Med.* 140, 206–223. doi:10.1016/j.freeradbiomed.2019.05.007.
Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput Biol* 7, e1002195.
Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi:10.1093/bioinformatics/btq461.
Eraso, J. M., Markillie, L. M., Mitchell, H. D., Taylor, R. C., Orr, G., and Margolin, W. (2014). The highly conserved MraZ protein is a transcriptional regulator in Escherichia coli. *J. Bacteriol.* 196, 2053–2066. doi:10.1128/JB.01370-13.
Errington, J. (2013). L-form bacteria, cell walls and the origins of life. *Open Biol.* 3, 120143. doi:10.1098/rsob.120143.
Eveleigh, R. J. M., Meehan, C. J., Archibald, J. M., and Beiko, R. G. (2013). Being Aquifex aeolicus: Untangling a hyperthermophile's checkered past. *Genome Biol. Evol.* 5, 2478–2497. doi:10.1093/gbe/evt195.
Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. CRC Press.
Gouy, R., Baurain, D., and Philippe, H. (2015). Rooting the tree of life: the phylogenetic jury is still out. *Philos. Trans. R. Soc. B Biol. Sci.* 370, 20140329. doi:10.1098/rstb.2014.0329.
Gully, D., and Bouveret, E. (2006). A protein network for phospholipid synthesis uncovered by a variant of the tandem affinity purification method in Escherichia coli. *Proteomics* 6, 282–293.
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 Int. J. Gen. Mol. Microbiol.* 100, 171–182. doi:10.1007/s10482-011-9616-8.
Hagan, C. L., Silhavy, T. J., and Kahne, D. (2011). β -Barrel Membrane Protein Assembly by the Bam Complex. doi:10.1146/annurev-biochem-061408-144611.
Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016). A new view of the tree of life. *Nat. Microbiol.* 1, 16048. doi:10.1038/nmicrobiol.2016.48.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
Kersey, P. J., Allen, J. E., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., et al. (2014). Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.* 42, D546–D552. doi:10.1093/nar/gkt979.
Kivistö, A. T., and Karp, M. T. (2011). Halophilic anaerobic fermentative bacteria. *J. Biotechnol.* 152, 114–124. doi:10.1016/j.jbiotec.2010.08.014.
Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39, 309–338. doi:10.1146/annurev.genet.39.073003.114725.
Koonin, E. V. (2016). Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Research* 5, F1000 Faculty Rev-1805. doi:10.12688/f1000research.8737.1.
Lake, J. A. (2009). Evidence for an early prokaryotic endosymbiosis. *Nature* 460, 967–971. doi:10.1038/nature08183.
Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* 7 Suppl 1, S4. doi:10.1186/1471-2148-7-S1-S4.
Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25, 2286–2288. doi:10.1093/bioinformatics/btp368.
Lartillot, N., and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* 21, 1095–1109. doi:10.1093/molbev/msh112.
Lartillot, N., and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Syst. Biol.* 55, 195–207. doi:10.1080/10635150500433722.
Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Syst. Biol.* 62, 611–615. doi:10.1093/sysbio/syt022.
Lasek-nesselquist, E., and Gogarten, J. P. (2013). The effects of model choice and mitigating bias on the ribosomal tree of life. *Mol. Phylogenet. Evol.* 69, 17–38. doi:10.1016/j.ympev.2013.05.006.
Léonard, R. R., Leleu, M., Vlierberghe, M. V., Cornet, L., Kerff, F., and Baurain, D. (2021). ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies. *PeerJ* 9, e11348. doi:10.7717/peerj.11348.
Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* doi:10.1093/nar/gkab301.
Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi:10.1101/gr.1224503.
Liu, L., Anderson, C., Pearl, D., and Edwards, S. V. (2019). “Modern Phylogenomics: Building Phylogenetic Trees Using the Multispecies Coalescent Model,” in *Evolutionary Genomics: Statistical and Computational Methods*, ed. M. Anisimova, 211–239.
Luef, B., Frischkorn, K. R., Wrighton, K. C., Holman, H. N., Birarda, G., Thomas, B. C., et al. (2015). Diverse uncultivated ultra-small bacterial cells in groundwater. doi:10.1038/ncomms7372.
Mavromatis, K., Ivanova, N., Anderson, I., Lykidis, A., Hooper, S. D., Sun, H., et al. (2009). Genome Analysis of the Anaerobic Thermohalophilic Bacterium Halothermothrix orenii. *PLOS ONE* 4, e4192. doi:10.1371/journal.pone.0004192.
Meade, A., and Pagel, M. (2017). *BayesTraits V3. 0.1*.
Megrian, D., Taib, N., Witwinowski, J., Beloin, C., and Gribaldo, S. (2020). One or two

membranes ? Diderm Firmicutes challenge the Gram-positive / Gram-negative divide. 659–671. doi:10.1111/mmi.14469.
Melville, S., and Craig, L. (2013). Type IV pili in Gram-positive bacteria. *Microbiol. Mol. Biol. Rev.* 77, 323–341.
Mingorance, J., and Tamames, J. (2004). The bacterial dcw gene cluster: an island in the genome? *Mol. Time Space*, 249–271.
Mira, A., Pushker, R., Legault, B. A., Moreira, D., and Rodríguez-Valera, F. (2004). Evolutionary relationships of Fusobacterium nucleatum based on phylogenetic analysis and comparative genomics. *BMC Evol. Biol.* 4, 50. doi:10.1186/1471-2148-4-50.
Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41, e121–e121.
Moreno-Hagelsieb, G., Wang, Z., Walsh, S., and ElSherbiny, A. (2013). Phylogenomic clustering for selecting non-redundant genomes for comparative genomics. *Bioinformatics* 29, 947–949. doi:10.1093/bioinformatics/btt064.
Narita, S., and Tokuda, H. (2009). Biochemical characterization of an ABC transporter LptBFGC complex required for the outer membrane sorting of lipopolysaccharides. *FEBS Lett.* 583, 2160–2164. doi:10.1016/j.febslet.2009.05.051.
Nikolaichik, Y. A., and Donachie, W. D. (2000). Conservation of Gene Order Amongst Cell Wall and Cell Division Genes In Eubacteria, and Ribosomal Genes in Eubacteria and Eukaryotic Organelles. 7.
Pagel, M., and Meade, A. (2015). Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. *Am. Nat.* doi:10.1086/503444.
Pagel, M., Meade, A., and Barker, D. (2004). Bayesian Estimation of Ancestral Character States on Phylogenies. *Syst. Biol.* 53, 673–684. doi:10.1080/10635150490522232.
Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36, 996–1004.
Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P., Woodcroft, B. J., Evans, P. N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 2. doi:10.1038/s41564-017-0012-7.
Perrin, A., Varré, J.-S., Blanquart, S., and Ouangraoua, A. (2015). ProCARs: Progressive Reconstruction of Ancestral Gene Orders. *BMC Genomics* 16 Suppl 5, S6. doi:10.1186/1471-2164-16-S5-S6.
Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., et al. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol.* 9. doi:10.1371/journal.pbio.1000602.
Philippe, H., Vienne, D. M. D., Ranwez, V., Roure, B., Baurain, D., and Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics.
Pilhofer, M., Rappl, K., Eckl, C., Bauer, A. P., Ludwig, W., Schleifer, K. H., et al. (2008). Characterization and evolution of cell division and cell wall synthesis genes in the bacterial phyla Verrucomicrobia, Lentisphaerae, Chlamydiae, and Planctomycetes and phylogenetic comparison with rRNA genes. *J. Bacteriol.* 190, 3192–3202. doi:10.1128/JB.01797-07.
R Core Team, Rf. (2013). *R: A language and environment for statistical computing*. R foundation for statistical computing Vienna, Austria.
Rachel, R., Engel, A. M., Huber, R., Stetter, K.-O., and Baumeister, W. (1990). A porin-

type protein is the main constituent of the cell envelope of the ancestral eubacterium Thermotoga maritima. *FEBS Lett.* 262, 64–68.
Rachel, R., Wildhaber, I., Stetter, K. O., and Baumeister, W. (1988). “The structure of the surface protein of Thermotoga maritima,” in *Crystalline Bacterial Cell Surface Layers* (Springer), 83–86.
Ranjit, C., and Noll, K. M. (2016). Distension of the toga of Thermotoga maritima involves continued growth of the outer envelope as cells enter the stationary phase. 1–7. doi:10.1093/femsle/fnw218.
Ray, M.-C., Germon, P., Vianney, A., Portalier, R., and Lazzaroni, J. C. (2000). Identification by genetic suppression ofEscherichia coli TolB residues important for TolB-Pal interaction. *J. Bacteriol.* 182, 821–824.
Raymann, K., Brochier-Armanet, C., and Gribaldo, S. (2015). The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci.*, 201420858. doi:10.1073/pnas.1420858112.
Real, G., and Henriques, A. O. (2006). Localization of the Bacillus subtilis murB Gene within the dcw Cluster Is Important for Growth and Sporulation. *J. Bacteriol.* 188, 1721–1732. doi:10.1128/JB.188.5.1721-1732.2006.
Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277.
Rigal, A., Bouveret, E., Lloubes, R., and Lazdunski, C. (1997). The TolB Protein Interacts with the Porins of Escherichia coli. 179, 7274–7279.
Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437. doi:10.1038/nature12352.
Rivas-Marín, E., Canosa, I., and Devos, D. P. (2016). Evolutionary Cell Biology of Division Mode in the Bacterial Planctomycetes-Verrucomicrobia- Chlamydiae Superphylum. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.01964.
Roure, B., Rodriguez-Ezpeleta, N., and Philippe, H. (2007). SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol. Biol.* 7 Suppl 1, S2. doi:10.1186/1471-2148-7-S1-S2.
Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., et al. (2011). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 39, D38–D51. doi:10.1093/nar/gkq1172.
Schleifer, K. H., and Kandler, O. (1972). Peptidoglycan types of bacterial cell walls and their taxonomic implications. *Bacteriol. Rev.* 36, 407–477.
Silhavy, T. J., Kahne, D., and Walker, S. (2010). The Bacterial Cell Envelope. 2, 16. doi:10.1101/cshperspect.a000414.
Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.
Sutcliffe, I. C. (2010). A phylum level perspective on bacterial cell envelope architecture. *Trends Microbiol.* 18, 464–470. doi:10.1016/j.tim.2010.06.005.
Sutcliffe, I. C. (2011). Cell envelope architecture in the Chloroflexi : a shifting frontline in a phylogenetic turf war. 13, 279–282. doi:10.1111/j.1462-2920.2010.02339.x.
Taib, N., Megrian, D., Witwinowski, J., Adam, P., Poppleton, D., Borrel, G., et al. (2020). Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat. Ecol. Evol.* 4, 1661–1672. doi:10.1038/s41559-020-01299-7.
Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. 1–11.
Tocheva, E. I., Matson, E. G., Morris, D. M., Moussavi, F., Leadbetter, J. R., and Jensen,

G. J. (2011). Peptidoglycan remodeling and conversion of an inner membrane into an outer membrane during sporulation. *Cell* 146, 799–812. doi:10.1016/j.cell.2011.07.029.
Tocheva, E. I., Ortega, D. R., and Jensen, G. J. (2016). Sporulation, bacterial cell envelopes, and the origin of life. *Nat. Publ. Group* 14, 535–542. doi:10.1038/nrmicro.2016.85.
Van Vlierberghe, M., Philippe, H., and Baurain, D. (2021). Broadly sampled orthologous groups of eukaryotic proteins for the phylogenetic study of plastid-bearing lineages. *BMC Res. Notes* 14, 143. doi:10.1186/s13104-021-05553-4.
Verma, M., Lal, D., Kaur, J., Saxena, A., Kaur, J., Anand, S., et al. (2013). Phylogenetic analyses of phylum Actinobacteria based on whole genome sequences. *Res. Microbiol.* 164, 718–728. doi:10.1016/j.resmic.2013.04.002.
Vollmer, W. (2011). Bacterial outer membrane evolution via sporulation? *Nat. Chem. Biol.* 8, 14–18. doi:10.1038/nchembio.748.
Walburger, A., Lazdunski, C., and Corda, Y. (2002). The Tol / Pal system function requires an interaction between the C-terminal domain of TolA and the N- terminal domain of TolB. 44, 695–708.
Wang, Z., Xiang, Q., Zhu, X., Dong, H., He, C., Wang, H., et al. (2014). Structural and functional studies of conserved nucleotide-binding protein LptB in lipopolysaccharide transport. *Biochem. Biophys. Res. Commun.* 452, 443–449.
Woese, C. R. (1987). Bacterial Evolution. 51, 221–271.
Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., et al. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462, 1056–1060. doi:10.1038/nature08656.
Yakhnina, A. A., and Bernhardt, T. G. (2020). The Tol-Pal system is required for peptidoglycan-cleaving enzymes to complete bacterial cell division. *Proc. Natl. Acad. Sci. U. S. A.* 117, 6777–6783. doi:10.1073/pnas.1919267117.
Yutin, N., and Galperin, M. Y. (2013). A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. *Environ. Microbiol.* 15, 2631–2641. doi:https://doi.org/10.1111/1462-2920.12173.
Yutin, N., Puigbo, P., Koonin, E. V., and Wolf, Y. I. (2012). Phylogenomics of Prokaryotic Ribosomal Proteins. *Curr. Sci.* 101, 1435–1439. doi:10.1371/Citation.
Zhaxybayeva, O., Swithers, K. S., Lapierre, P., Fournier, G. P., Bickhart, D. M., Deboy, R. T., et al. (2009). On the chimeric nature , thermophilic origin , and phylogenetic placement of the Thermotogales. 106, 5865–5870.
Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J. G., et al. (2019). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* doi:10.1038/s41467-019-13443-4.

5.2.10 Supplementary materials

mean score +/- stdev	CAT+GTR+Γ	CAT+Γ	GTR+Γ	LG+Γ
CAT+GTR+Γ	/	1279.9 +/- 116.634	8994.2 +/- 347.355	9317.5 +/- 358.479
CAT+Γ	-1279.9 +/- 116.634	/	7714.3 +/- 391.3	8037.6 +/- 407.53
GTR+Γ	-8994.2 +/- 347.355	-7714.3 +/- 391.3	/	323.3 +/- 58.9441
LG+Γ	-9317.5 +/- 358.479	-8037.6 +/- 407.53	-323.3 +/- 58.9441	/

Table S1: Results of the cross-validation procedure comparing four different models of sequence evolution available in PhyloBayes MPI.

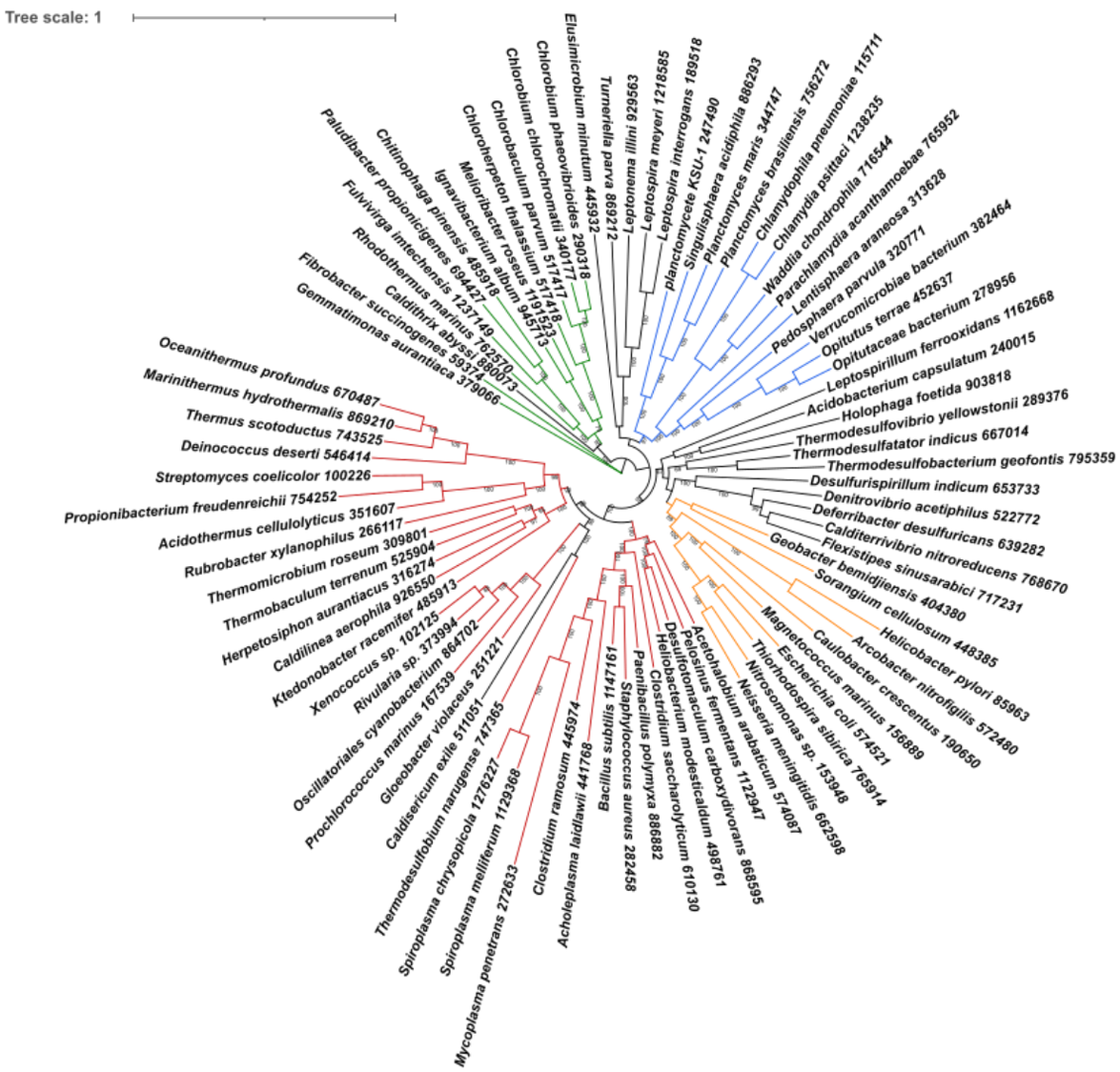


Figure S1: Phylogenomic tree of the bacterial domain based on a supermatrix concatenating 117 single-copy orthologous genes chosen for their broad conservation across Bacteria. The supermatrix contained 85 species and 19,959 unambiguously aligned amino-acid positions (< 5% missing character states). The tree was inferred from amino-acid sequences using RAxML v8.1.17 with the LG+F+Γ model of sequence evolution.

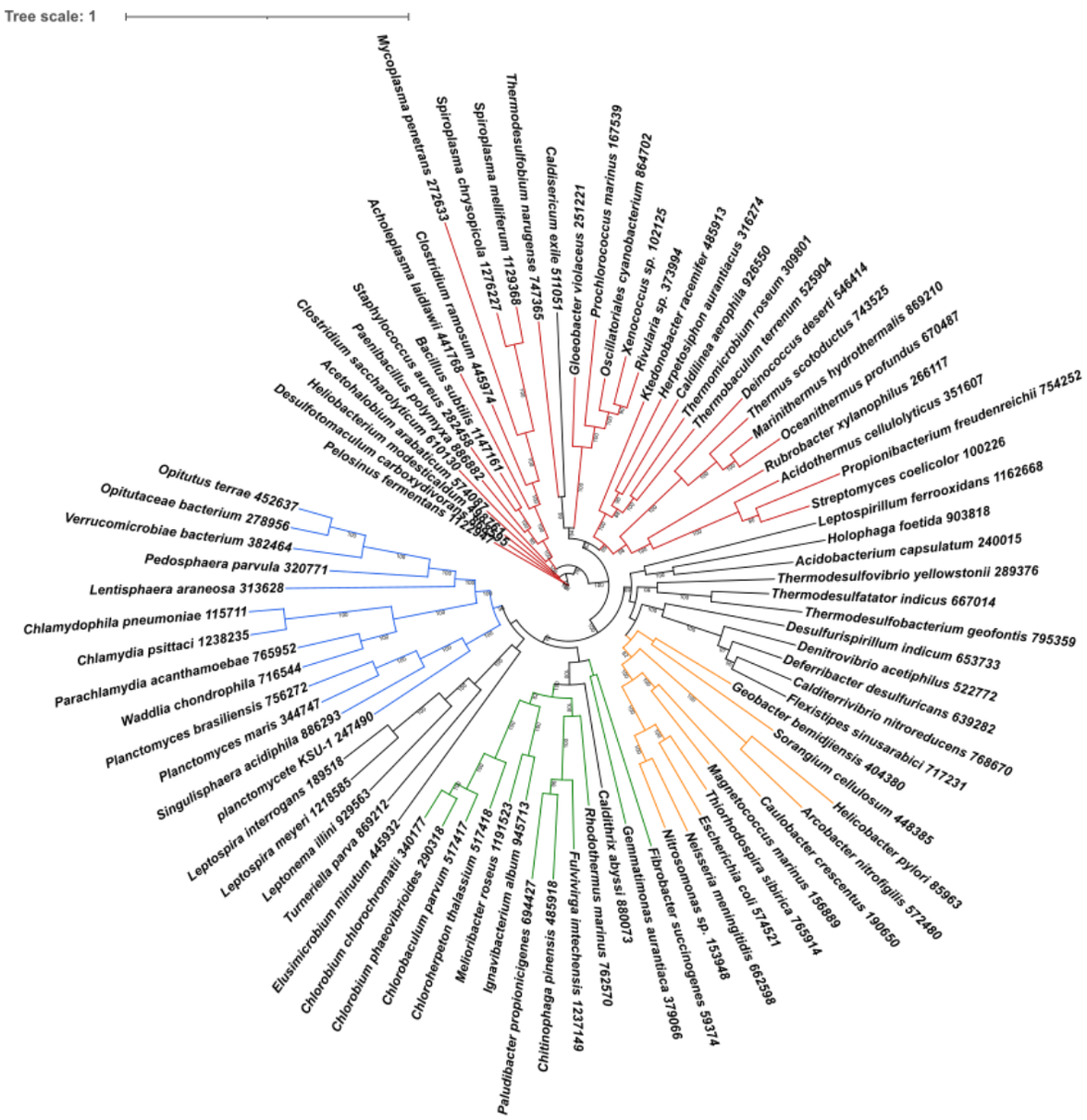


Figure S2: Phylogenomic tree of the bacterial domain based on a supermatrix concatenating 117 single-copy orthologous genes chosen for their broad conservation across Bacteria. The supermatrix contained 85 species and 19,959 unambiguously aligned amino-acid positions (< 5% missing character states). “Cell-by-cell outliers” (62 sequences) identified by Phylo-MCOA v1.4 were removed from the supermatrix. The tree was inferred from amino-acid sequences using RAxML v8.1.17 with the LG+F+Γ model of sequence evolution.

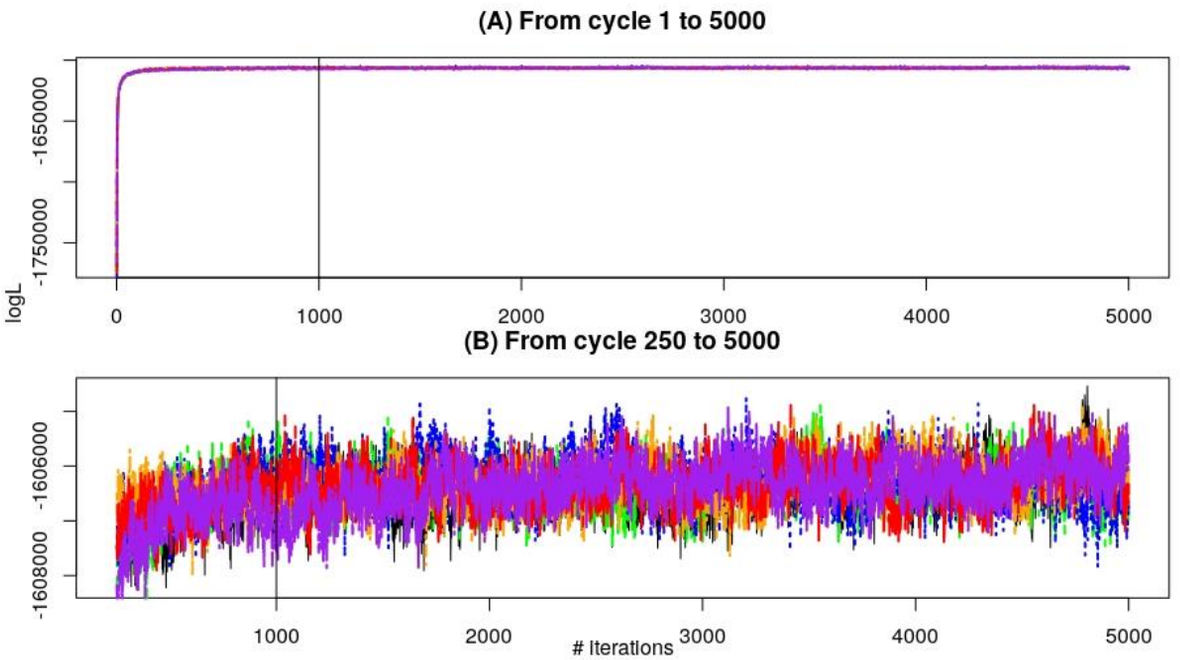


Figure S3: Evolution of the log likelihood of six PhyloBayes MCMC chains running under the CAT+GTR+Γ model of sequence evolution. The vertical line at cycle 1000 marks the end of the burnin. The supermatrix is the one of Figure 1.

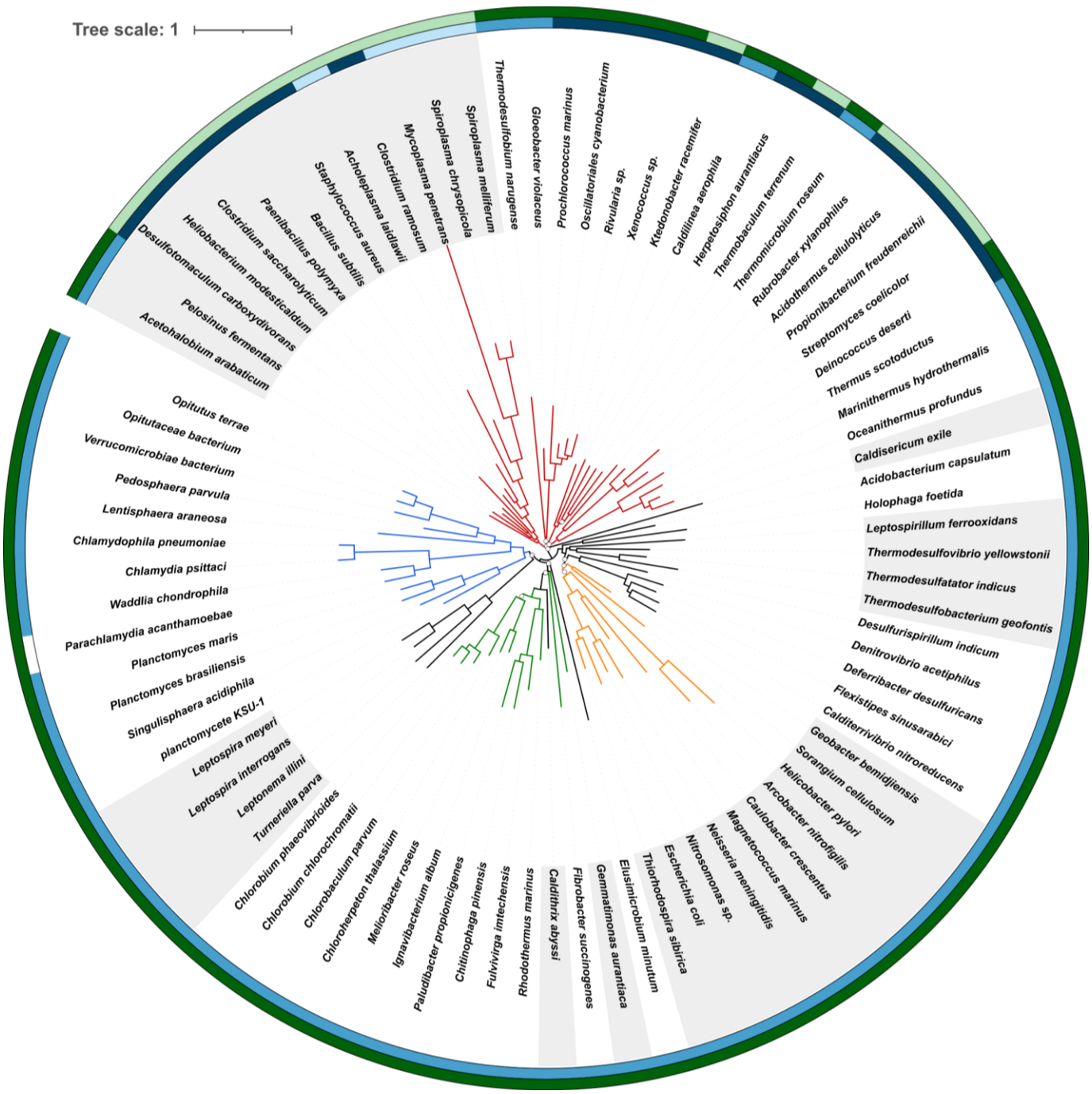
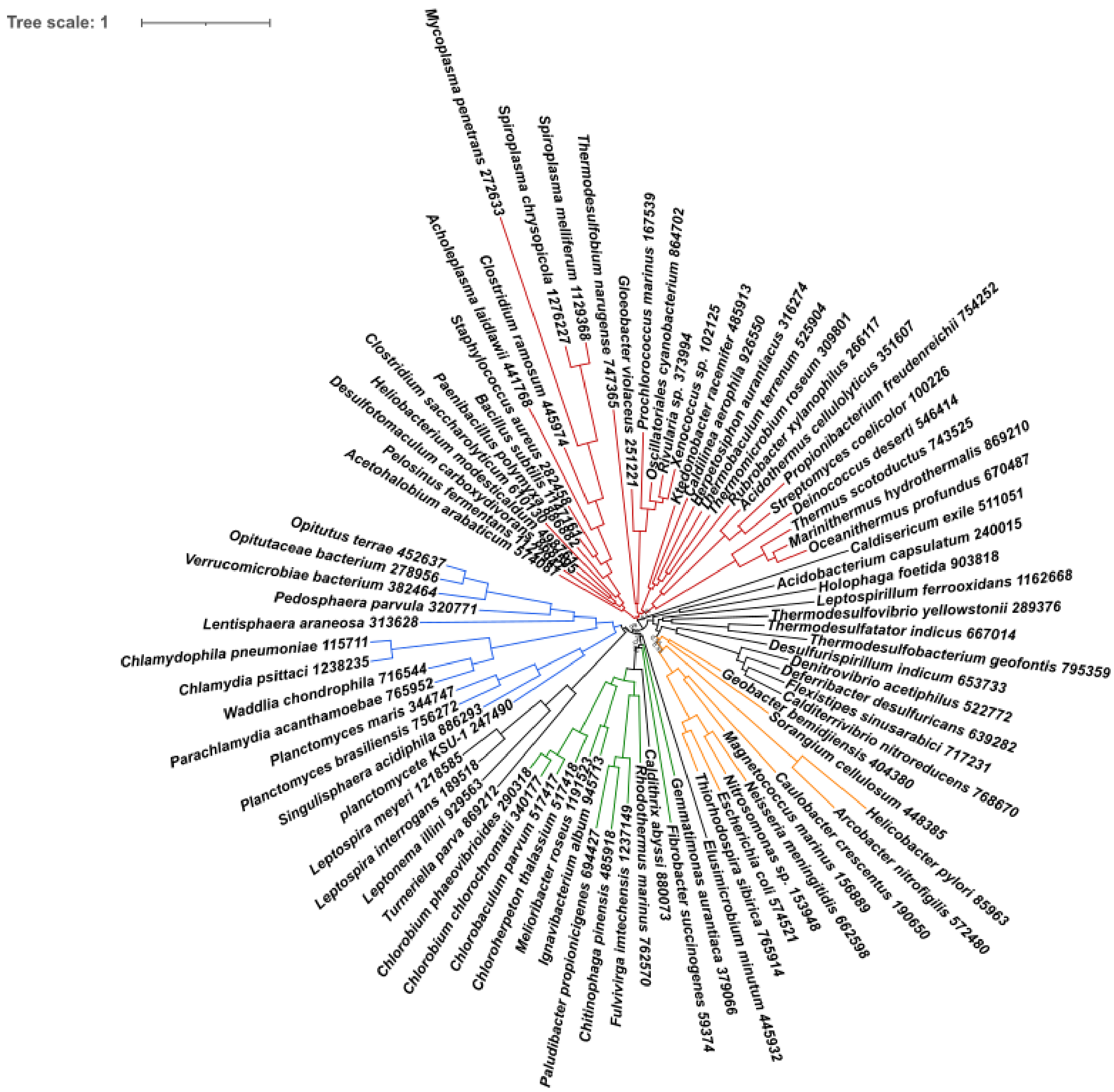


Figure S4: Phylogenomic tree of the bacterial domain based on a supermatrix concatenating 117 single-copy orthologous genes chosen for their broad conservation across Bacteria. This tree is the consensus of the six MCMC chains (A to F) shown in Figure S5. Otherwise, it is equivalent to Figure 1.



Figures S5A to S5F: Trees inferred by the six individual MCMC chains running under the CAT+GTR+Γ model of sequence evolution. Only PP <1.0 are shown. Figure 1 is the consensus of chains A and C. For a consensus built from all six chains, see Figure S4.

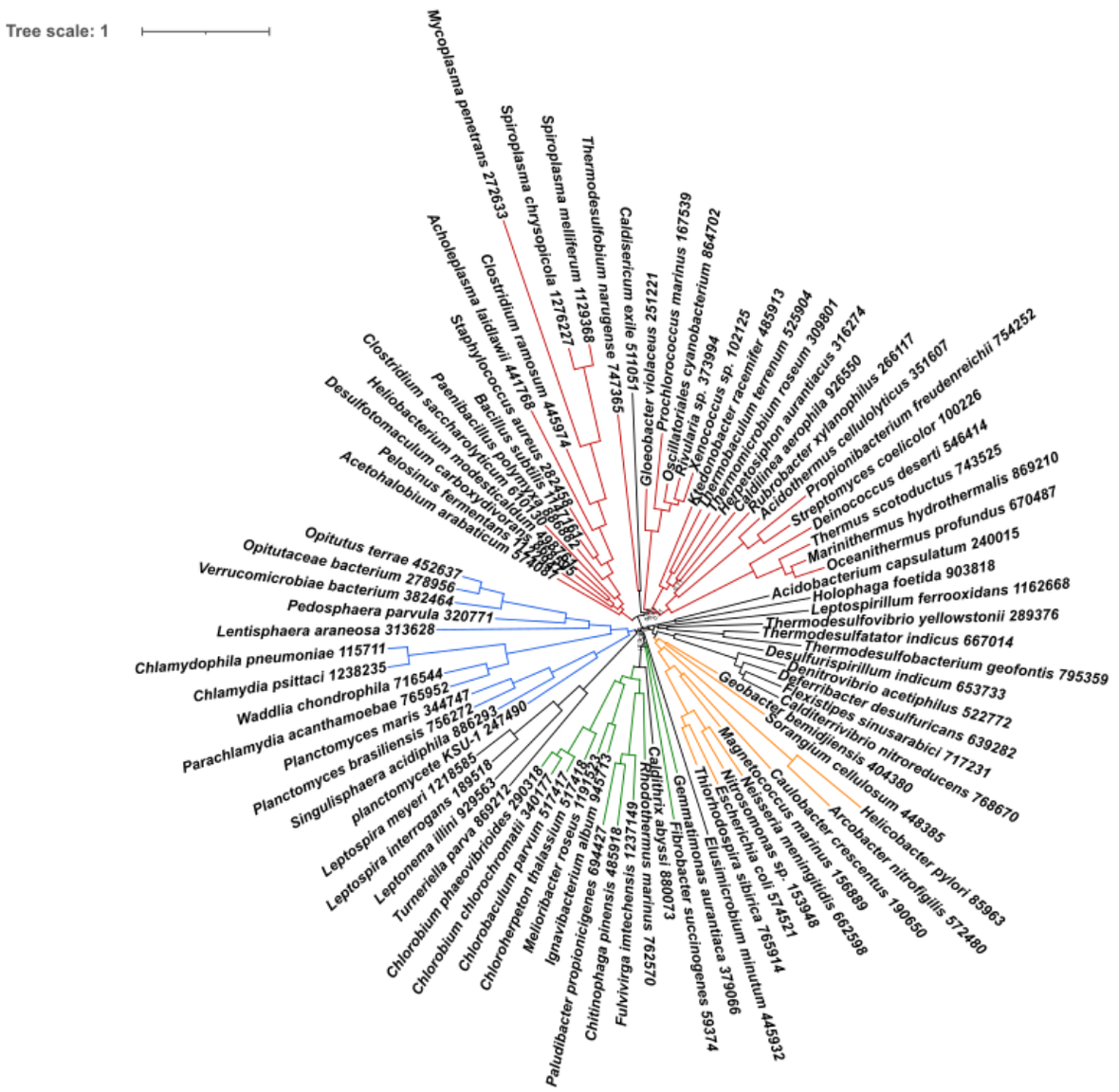

Tree scale: 1
Mycoplasma penetrans 272633
Spiroplasma chrysopicola 1276227
Spiroplasma melliferum 1129368
Thermodesulfobium narugense 747365
Caldisericum exile 511051
Gloeobacter violaceus 251221
Prochlorococcus marinus 167539
Oscillatoriales cyanobacterium 864702
Rivularia sp. 373994
Xenococcus sp. 102125
Kiedonobacter racemifer 485913
Thermomicrobium terrenum 525904
Thermobaculum roseum 309801
Herpetosiphon aurantiacus 316274
Caldilinea aerophila 926550
Rubrobacter xylanophilus 266117
Acidothermus cellulolyticus 351607
Propionibacterium freudenreichii 754252
Streptomyces coelicolor 100226
Deinococcus deserti 546414
Thermus scotoductus 743525
Marinithermus hydrothermalis 869210
Oceanithermus profundus 670487
Acidobacterium capsulatum 240015
Holophaga foetida 903818
Leptospirillum ferrooxidans 1162668
Thermodesulfovibrio yellowstonii 289376
Thermodesulfatator indicus 667014
Thermodesulfobacterium geofontis 795359
Desulfurispirillum indicum 653733
Denitrovibrio acetiphilus 522772
Deferribacter desulfuricans 639282
Flexistipes sinusarabici 717231
Calditerrivibrio nitroreducens 768670
Sorangium cellulosum 448385
Helicobacter pylori 85963
Arcobacter crescentus 190650
Caulobacter nitrofigilis 572480
Geobacter bemidjiensis 404380
Magnetococcus marinus 156889
Neisseria meningitidis 662598
Nitrosomonas sp. 153948
Escherichia coli 514521
Thiorhodospira sibirica 765914
Elusimicrobium minutum 445932
Thiorhodovibrio winogradskyi 379066
Gemmatimonas aurantiaca 59374
Fibrobacter succinogenes 59374
Calothrix abyssi 880073
Rhodothermus marinus 762570
Fulvivirga imtechensis 1232149
Chitinophaga pinensis 485918
Paludibacter propionicigenes 694427
Ignavibacterium album 945713
Meiothermus roseus 1191523
Chlorobaculum parvum 517417
Chlorobium chlorochromatii 340177
Chlorobium phaeovibrioides 290318
Chlorobium thalassium 517418
Chlorotherpeton thalassium 517418
Leptonema illini 929563
Turneriella parva 869212
Leptospira interrogans 189518
Leptospira meyeri 1218585
Planctomycete KSU-1 247490
Singulisphaera acidiphila 886293
Planctomyces brasiliensis 756272
Planctomyces maris 344747
Parachlamydia acanthamoebae 765952
Waddlia chondrophila 716544
Chlamydia psittaci 1238235
Chlamydophila pneumoniae 115711
Lentisphaera araneosa 313628
Pedosphaera parvula 320771
Verrucomicrobiae bacterium 382464
Opitutaceae bacterium 278956
Opitutus terrae 452637
Acetohalobium arabaticum 574087
Pelosinus fermentans 1121866
Desulfotomaculum carboxydivorans 1121416
Heliobacterium modesticaldum 498761
Clostridium saccharolyticum 610130
Bacillus subtilis 1423
Paenibacillus polymyxa 1069682
Staphylococcus aureus 282458
Staphylococcus laidlawii 441768
Clostridium ramosum 445974
Acholeplasma laidlawii 441768

Figure S5B.

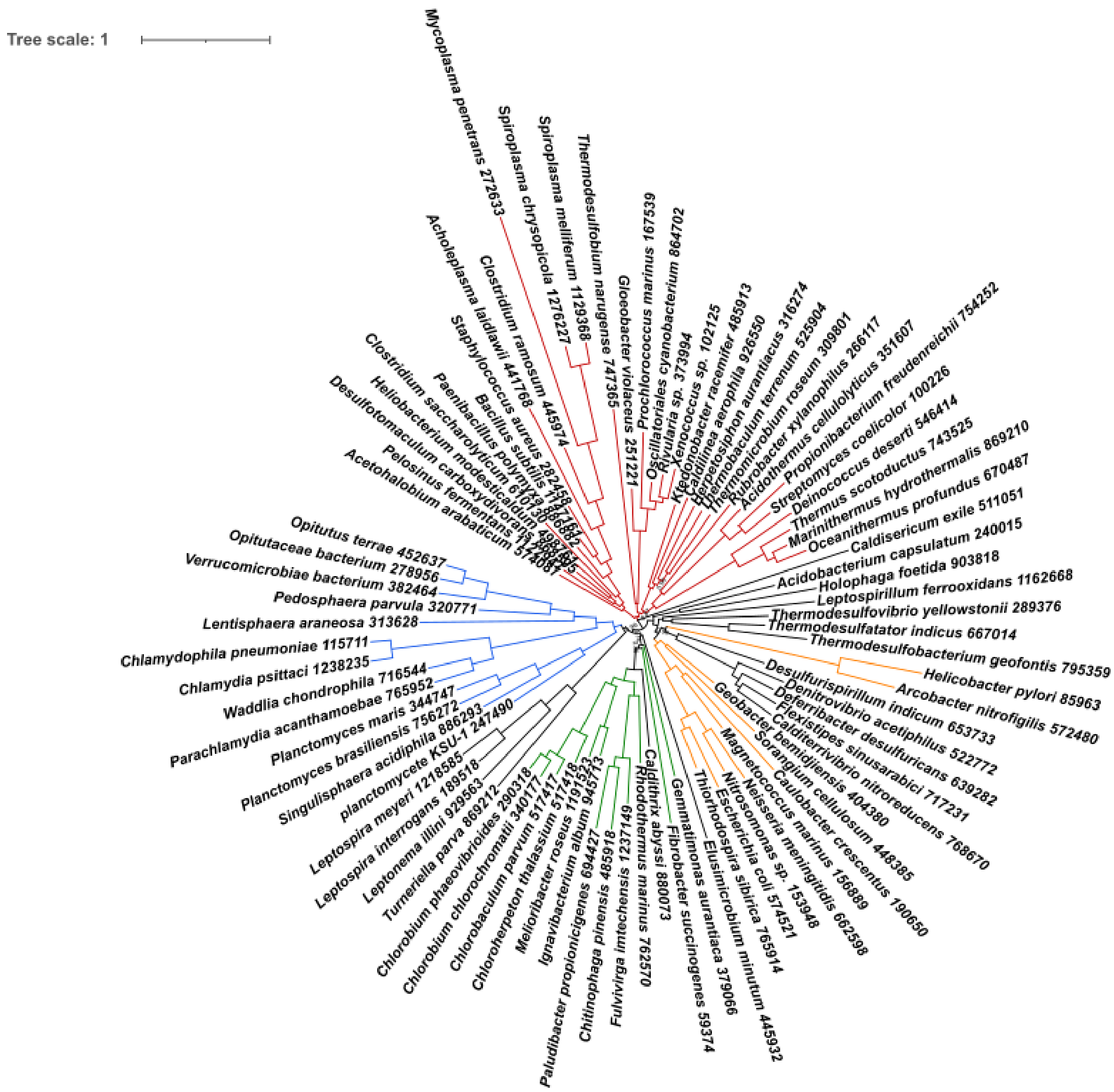


Figure S5C.

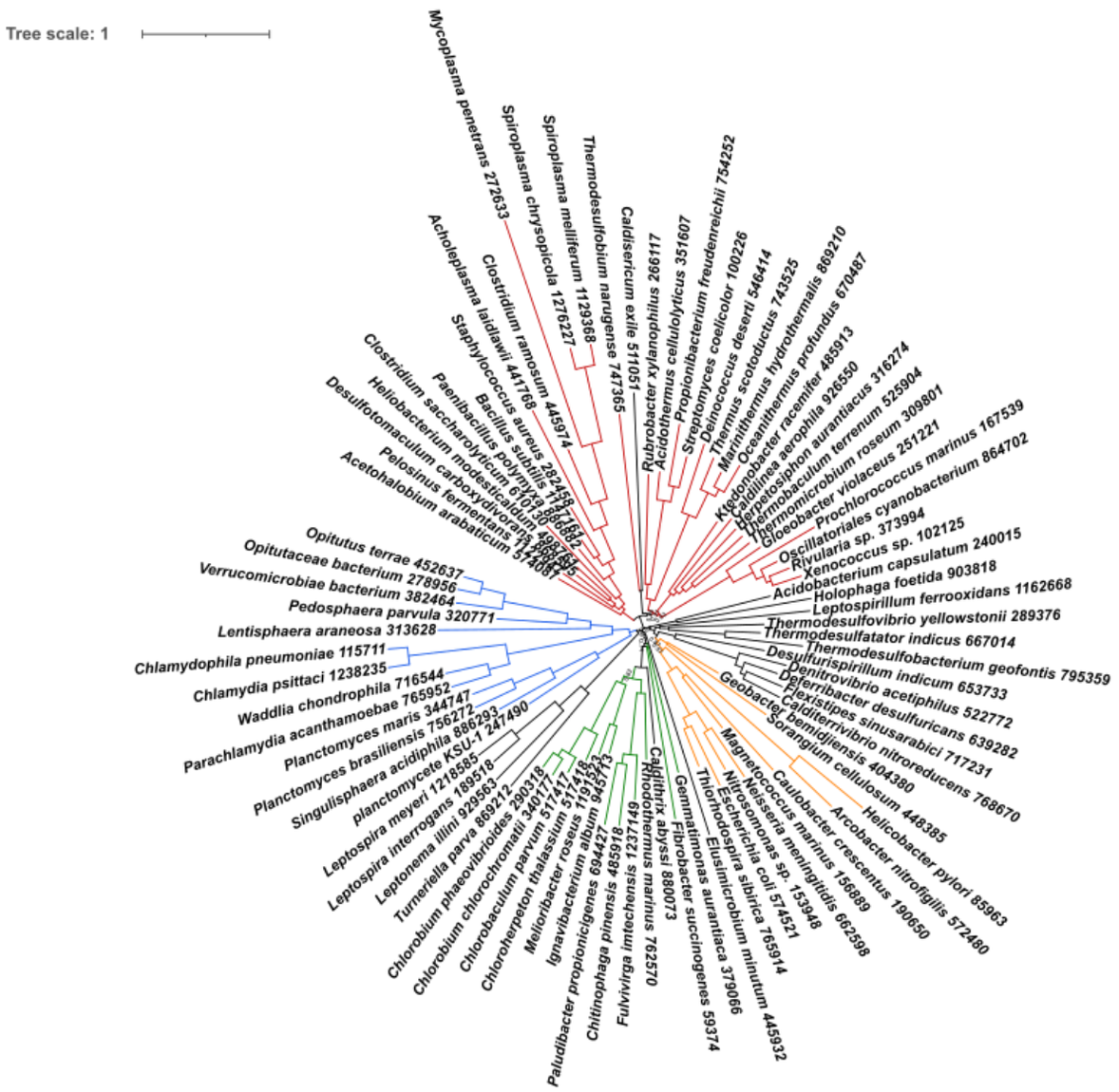

Tree scale: 1
Mycoplasma penetrans 272633
Spiroplasma chrysopicola 1276227
Spiroplasma melliferum 1129368
Thermodesulfobium narugense 747365
Caldisericum exile 511051
Rubrobacter xylanophilus 266117
Acidothermus cellulolyticus 351607
Propionibacterium freudenreichii 754252
Streptomyces coelicolor 100226
Deinococcus deserti 546414
Thermus scotoductus 743525
Marinithermus hydrothermalis 869210
Oceanithermus profundus 670487
Caldilinea aerophila 485913
Herpetosiphon aurantiacus 926550
Thermobaculum terrenum 316274
Thermomicrobium roseum 525904
Roseobacter violaceus 309801
Prochlorococcus marinus 251221
Gloeobacter violaceus 864702
Oscillatoriales cyanobacterium 373994
Rivularia sp. 102125
Xenococcus sp. 102125
Acidobacterium capsulatum 240015
Holophaga foetida 903818
Leptospirillum ferrooxidans 1162668
Thermodesulfovibrio yellowstonii 289376
Thermodesulfatator indicus 667014
Thermodesulfobacterium geofontis 795359
Desulfurispirillum indicum 653733
Denitrovibrio acetiphilus 522772
Deferribacter desulfuricans 639282
Flexistipes sinusarabici 717231
Calditerrivibrio nitroreducens 768670
Sorangium cellulosum 448385
Geobacter bemidjiensis 404380
Helicobacter pylori 85963
Caulobacter crescentus 190650
Arcobacter nitrofigilis 572480
Magnetococcus marinus 156889
Neisseria meningitidis 662598
Nitrosomonas sp. 153948
Escherichia coli 514521
Thiorhodospira sibirica 765914
Ectothiorhodospira mobilis 445932
Thiorhodococcus drewsii 379066
Elusimicrobium minutum 59374
Gemmatimonas aurantiaca 59374
Fibrobacter succinogenes 880073
Calothrix abyssi 880073
Rhodothermus marinus 762570
Fulvivirga imtechensis 1237146
Chitinophaga pinensis 485918
Paludibacter propionicigenes 694427
Ignavibacterium album 945713
Melioribacter roseus 1191523
Chloroherpeton thalassium 517417
Chlorobaculum parvum 517417
Chlorobium chlorochromatii 340177
Chlorobium phaeovibrioides 290318
Leptonema illini 929563
Turneriella parva 869212
Leptospira interrogans 189518
Leptospira meyeri 1218585
planctomycete KSU-1 247490
Planctomyces brasiliensis 756272
Singulisphaera acidiphila 886293
Planctomyces maris 344747
Parachlamydia acanthamoebae 765952
Waddlia chondrophila 716544
Chlamydia psittaci 1238235
Chlamydophila pneumoniae 115711
Lentisphaera araneosa 313628
Pedosphaera parvula 320771
Verrucomicrobiae bacterium 382464
Opitutaceae bacterium 278956
Opitutus terrae 452637
Acetohalobium arabaticum 574087
Pelosinus fermentans 749784
Desulfotomaculum carboxydivorans 484770
Heliobacterium modesticaldum 498761
Clostridium saccharolyticum 610136
Paenibacillus polymyxa 886882
Bacillus subtilis 1149
Staphylococcus aureus 282458
Clostridium ramosum 445974
Acholeplasma laidlawii 441768

Figure S5D.

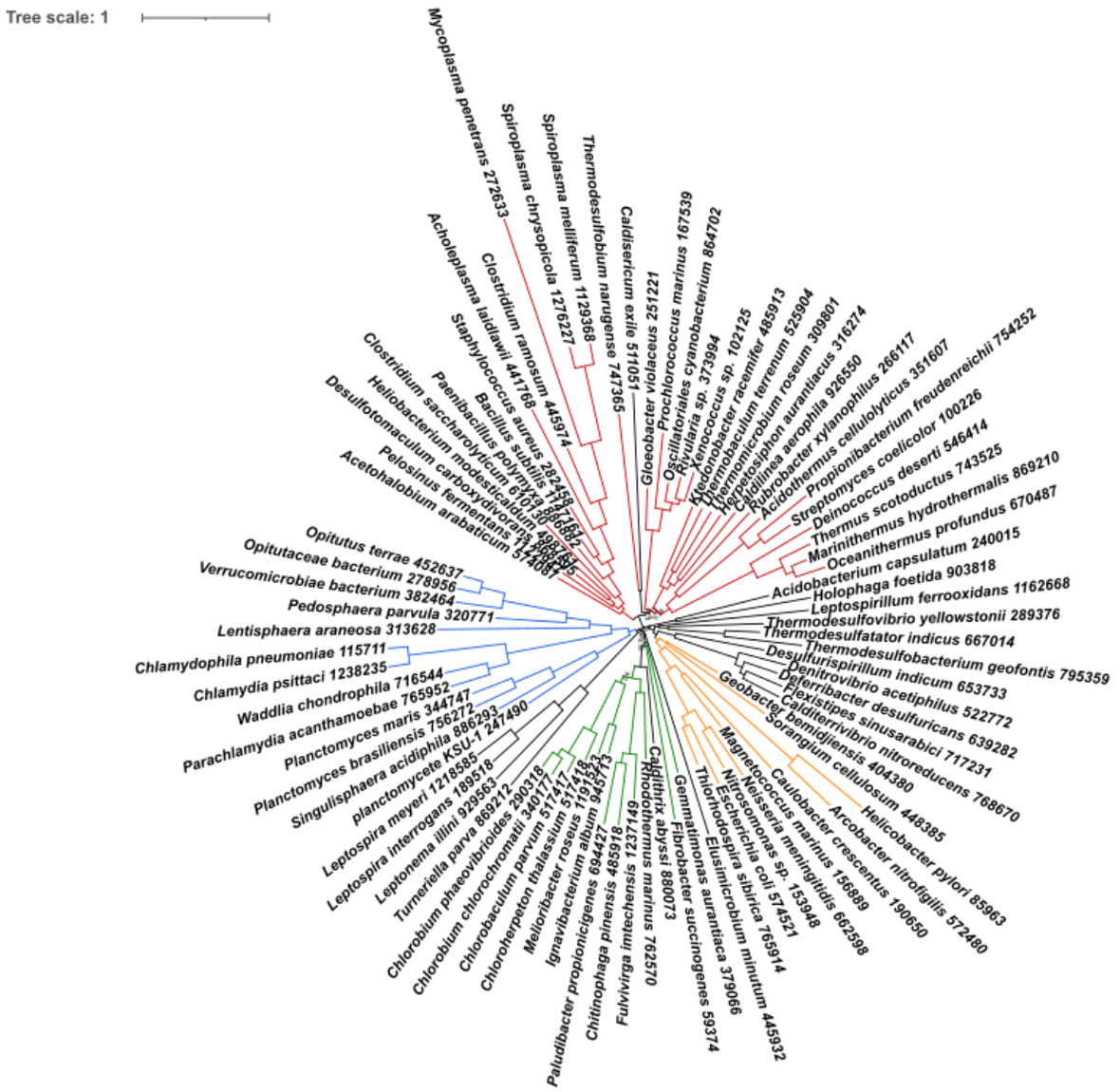

Tree scale: 1
Mycoplasma penetrans 272633
Spiroplasma chrysopicola 1276227
Spiroplasma melliferum 1129368
Thermodesulfobium narugense 747365
Caldisericum exile 511051
Gloeobacter violaceus 251221
Prochlorococcus marinus 167539
Oscillatoriales cyanobacterium 864702
Rivularia sp. 373984
Xenococcus sp. 102125
Kladonobacter racemifer 485913
Thermomicrobium terrenum 525904
Thermobaculum roseum 309801
Herpetosiphon aurantiacus 316274
Chloroflexus aggregans 926550
Rubrobacter xylanophilus 266117
Acidothermus cellulolyticus 351607
Propionibacterium freudenreichii 754252
Streptomyces coelicolor 100226
Deinococcus deserti 546414
Thermus scotoductus 743525
Marinithermus hydrothermalis 869210
Oceanithermus profundus 670487
Acidobacterium capsulatum 240015
Holophaga foetida 903818
Leptospirillum ferrooxidans 1162668
Thermodesulfovibrio yellowstonii 289376
Thermodesulfatator indicus 667014
Thermodesulfobacterium geofontis 795359
Desulfurispirillum indicum 653733
Denitrovibrio acetiphilus 522772
Deferribacter desulfuricans 639282
Flexistipes sinusarabici 717231
Calditerrivibrio nitroreducens 768670
Geobacter bemidjiensis 404380
Sorangium cellulosum 448385
Helicobacter pylori 85963
Arcobacter crescentus 190650
Caulobacter nitrofigilis 572480
Magnetococcus marinus 156889
Neisseria meningitidis 662598
Nitrosomonas sp. 153948
Escherichia coli 574521
Thiorhodospira sibirica 765914
Elusimicrobium minutum 445932
Thiorhodococcus drewsii 379066
Gemmatimonas aurantiaca 59374
Fibrobacter succinogenes 60488
Calditrichia abyssi 880073
Rhodothermus marinus 762570
Fulvivirga imtechensis 1237149
Chitinophaga pinensis 485918
Paludibacter propionicigenes 694427
Ignavibacterium album 945713
Melioribacter roseus 1191523
Chloroherpeton thalassium 517417
Chlorobaculum parvum 517418
Chlorobium chlorochromatii 340177
Chlorobium phaeovibrioides 290318
Turneriella parva 869212
Leptonema illini 929563
Leptospira interrogans 189518
Leptospira meyeri 1218585
planctomycete KSU-1 247490
Singulisphaera acidiphila 886293
Planctomyces brasiliensis 756272
Planctomyces maris 344747
Parachlamydia acanthamoebae 765952
Waddlia chondrophila 716544
Chlamydia psittaci 1238235
Chlamydophila pneumoniae 115711
Lentisphaera araneosa 313628
Pedosphaera parvula 320771
Verrucomicrobiae bacterium 382464
Opitutaceae bacterium 278956
Opitutus terrae 452637
Acetohalobium arabaticum 574087
Pelosinus fermentans 1121486
Desulfotomaculum carboxydivorans 1121416
Heliobacterium modesticaldum 498761
Clostridium saccharolyticum 610130
Paenibacillus polymyxa 1069682
Bacillus subtilis 1149
Staphylococcus aureus 282458
Acholeplasma laidlawii 441768
Clostridium ramosum 445974

Figure S5E.

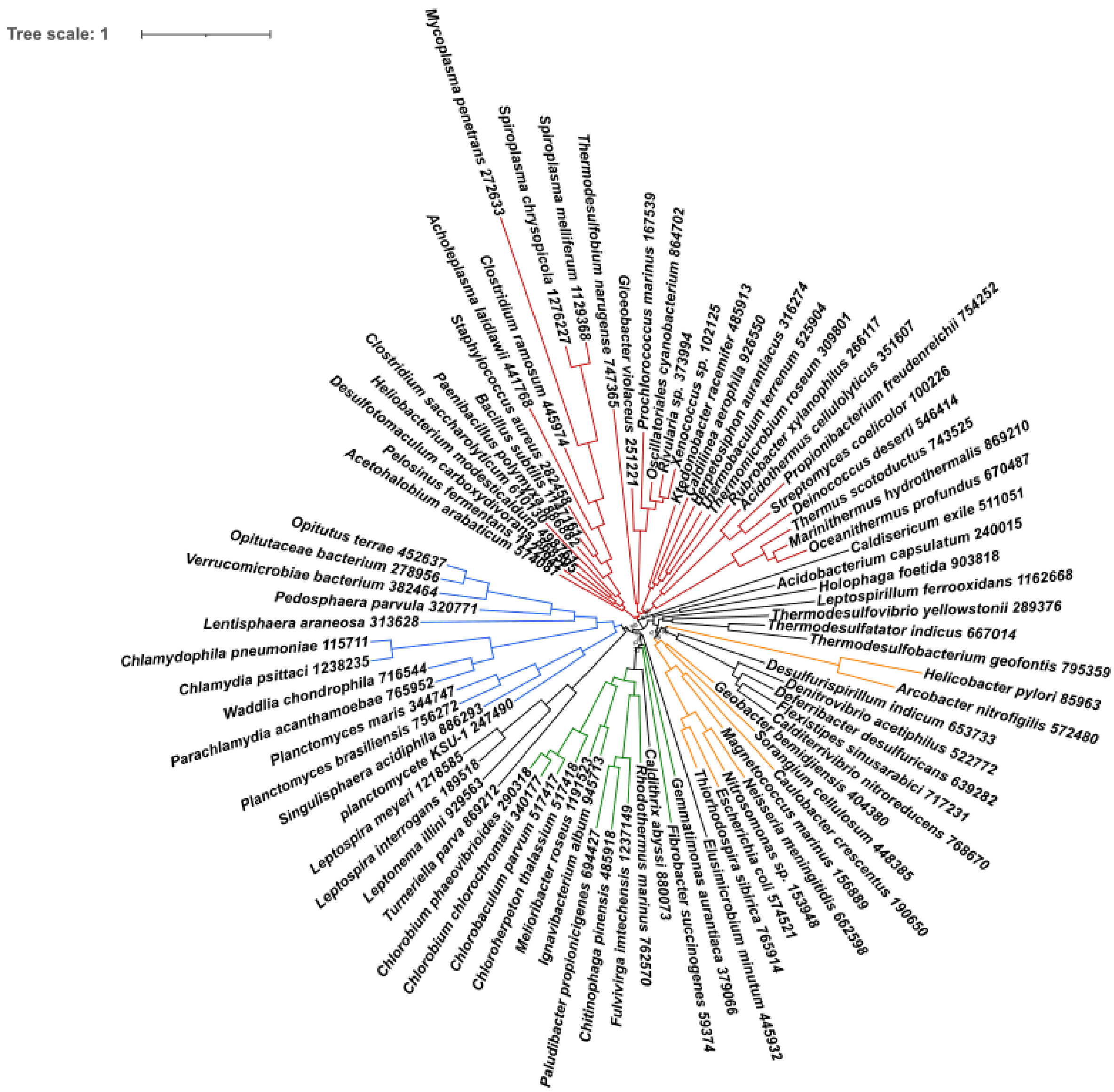

Tree scale: 1
Mycoplasma penetrans 272633
Spiroplasma chrysopicola 1276227
Spiroplasma melliferum 1129368
Thermodesulfobium narugense 747365
Acholeplasma laidlawii 441768
Clostridium ramosum 445974
Staphylococcus aureus 282458
Bacillus subtilis
Paenibacillus polymyxa
Clostridium saccharolyticum
Heliobacterium modesticaldum
Desulfotomaculum carboxydivorans
Pelosinus fermentans
Acetohalobium arabaticum 574087
Gloeobacter violaceus 251221
Prochlorococcus marinus 167539
Oscillatoriales cyanobacterium 864702
Rivularia sp. 373994
Xenococcus sp. 102125
Kiedonobacter racemifer 485913
Caldilinea aerophila 926550
Herpetosiphon aurantiacus 316274
Thermobaculum terrenum 525904
Thermomicrobium roseum 309801
Rubrobacter xylanophilus 266117
Acidothermus cellulolyticus 351607
Propionibacterium freudenreichii 754252
Streptomyces coelicolor 100226
Deinococcus deserti 546414
Thermus scotoductus 743525
Marinithermus hydrothermalis 869210
Oceanithermus profundus 670487
Caldisericum exile 511051
Acidobacterium capsulatum 240015
Holophaga foetida 903818
Leptospirillum ferrooxidans 1162668
Thermodesulfovibrio yellowstonii 289376
Thermodesulfatator indicus 667014
Thermodesulfobacterium geofontis 795359
Helicobacter pylori 85963
Arcobacter nitrofigilis 572480
Desulfurispirillum indicum 653733
Denitrovibrio acetiphilus 522772
Deferribacter desulfuricans 639282
Flexistipes sinusarabici 717231
Calditerrivibrio nitroreducens 768670
Geobacter bemidjiensis 404380
Sorangium cellulosum 448385
Caulobacter crescentus 190650
Magnetococcus marinus 156889
Neisseria meningitidis 662598
Nitrosomonas sp. 153948
Escherichia coli 574521
Thiorhodospira sibirica 765914
Elusimicrobium minutum 445932
Gemmatimonas aurantiaca 379066
Fibrobacter succinogenes 59374
Calditrix abyssi 880073
Rhodothermus marinus 762570
Fulvivirga imtechensis 1237149
Chitinophaga pinensis 485918
Paludibacter propionicigenes 694427
Ignavibacterium album 945713
Melioribacter roseus 1191523
Chloroherpeton thalassium 517418
Chlorobaculum parvum 517417
Chlorobium chlorochromatii 340177
Chlorobium phaeovibrioides 290318
Turneriella parva 869212
Leptonema illini 929563
Leptospira interrogans 189518
Leptospira meyeri 1218585
planctomycete KSU-1 247490
Singulisphaera acidiphila 886293
Planctomyces brasiliensis 756272
Planctomyces maris 344747
Parachlamydia acanthamoebae 765952
Waddlia chondrophila 716544
Chlamydia psittaci 1238235
Chlamydophila pneumoniae 115711
Lentisphaera araneosa 313628
Pedosphaera parvula 320771
Verrucomicrobiae bacterium 382464
Opitutaceae bacterium 278956
Opitutus terrae 452637

Figure S5F.

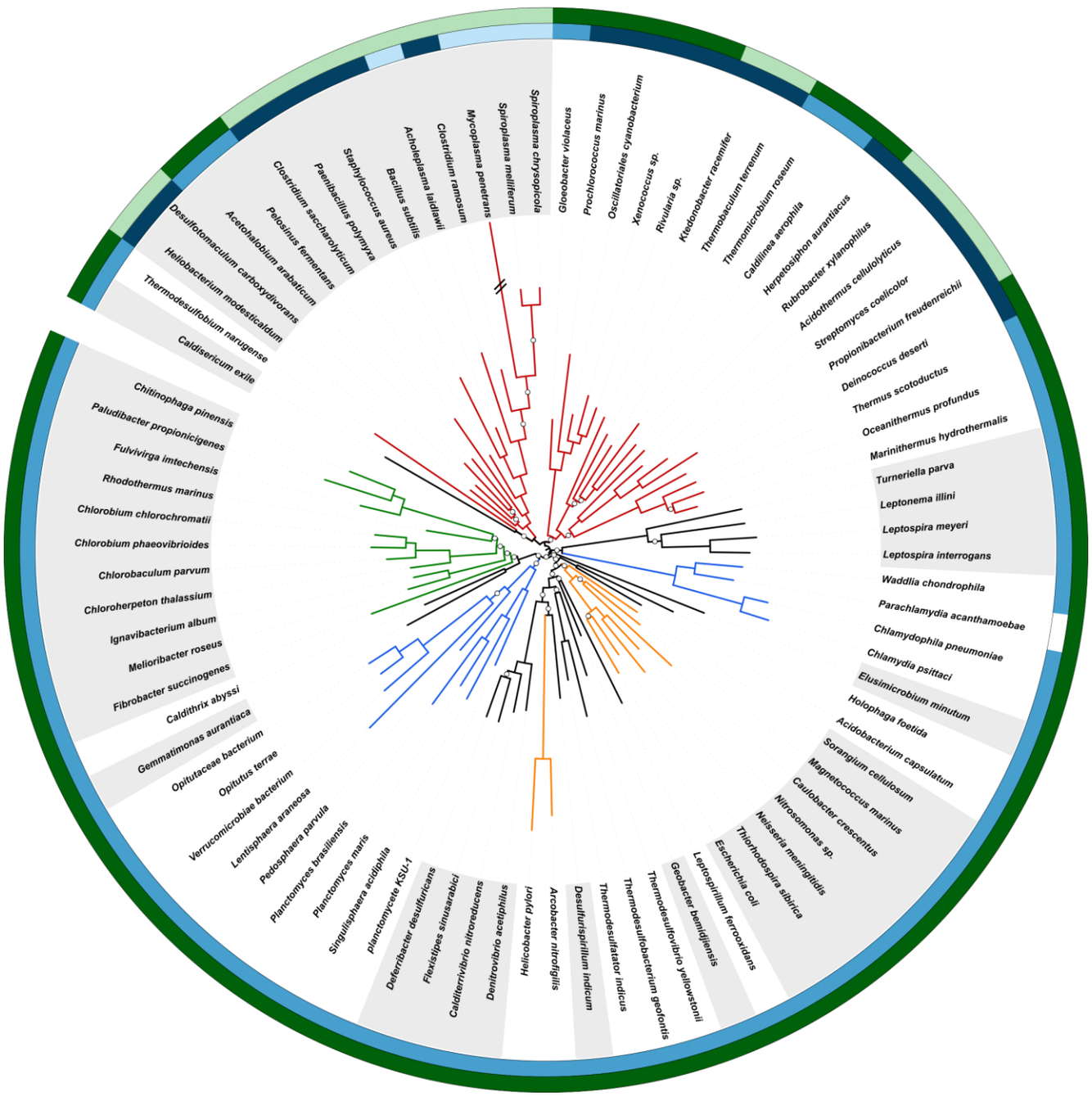


Figure S6: Phylogenomic tree of the bacterial domain based on a supermatrix concatenating 117 single-copy orthologous genes chosen for their broad conservation across Bacteria. The supermatrix contained 101 species and 19,959 unambiguously aligned amino-acid positions (4.72% missing character states). The tree was inferred from amino-acid sequences using PhyloBayes MPI and the CAT+Γ model of sequence evolution. Tree annotations are as in Figure 1: the circles at the nodes are posterior probabilities (PP) which are below the maximum statistical support (PP of 1.0). Nodes without a circle correspond to maximum statistical support for phylogenetic inference (PP of 1.0). The branch with "//" means that this branch has been cropped of half its length for clarity. The outer circles represent the status of the peptidoglycan (PG) and the outer membrane (OM) in the organism. Dark blue = thick peptidoglycan (PG), blue = thin PG, light blue = no PG. Dark green = diderm, light green = monoderm. White = no information.

Source	Roots
Battistuzzi & Hedges 2009	Deinococcus-Thermus
Battistuzzi & Hedges 2009	Terrabacteria
Forterre 2015	PVC
Yutin et al. 2011	Proteobacteria
Whidden et al. 2014	Deferribacteres/Nitrospira
Raymann et al. 2015	Terrabacteria
Ciccarelli et al. 2006	Firmicutes
Hug et al. 2019	Cyanobacteria

Table S2: Possible roots for the bacterial domain reported in the phylogenomic literature since 2006.

organism	phylum	PG	MBN (full)	spore	MBN (simple)	ref	taxonomic level of the information
Acetohalobium_arabaticum_574087	Firmicutes	1	2	1	1	59;60	species
Acholeplasma_laidlawii_441768	Tenericutes	0	0	0	0	87	class
Acidobacterium_capsulatum_240015	Acidobacteria	1	12	0	1	3;4	species
Acidothermus_cellulolyticus_351607	Actinobacteria	2	0	0	0	9	species
Arcobacter_nitrofigilis_572480	Proteobacteria	1	12	0	1	79	species
Bacillus_subtilis_1147161	Firmicutes	2	0	1	0	50-53	species
Caldilinea_aerophila_926550	Chloroflexi	1	12	0	1	32	species
Caldisericum_exile_511051	Caldiserica	1	12	0	1	16	species
Calditerrivibrio_nitroreducens_768670	Deferribacteres	1	12	0	1	42	species
Caldithrix_abyssi_880073	undef	1	12	0	1	93	species
Caulobacter_crescentus_190650	Proteobacteria	1	12	0	1	80	species
Chitinophaga_pinensis_485918	Bacteroidetes	1	12	3	1	12	species
Chlamydia_psittaci_1238235	Chlamydiae	1	2	0	1	17-20;25	species
Chlamydophila_pneumoniae_115711	Chlamydiae	1	2	0	1	17-20	phylum
Chlorobaculum_parvum_517417	Chlorobi	1	12	0	1	26	phylum
Chlorobium_chlorochromatii_340177	Chlorobi	1	12	0	1	26;29	species

Chlorobium_phaeovibrioides_290318	Chlorobi	1	12	0	1	26;28	species
Chloroherpeton_thalassium_517418	Chlorobi	1	12	0	1	26;27	species
Clostridium_ramosum_445974	Firmicutes	2	0	1	0	65	species
Clostridium_saccharolyticum_610130	Firmicutes	2	0	1	0	62;63	species
Deferribacter_desulfuricans_639282	Deferribacteres	1	12	0	1	41	species
Deinococcus_deserti_546414	Deinococcus-Thermus	2	2	0	1	47	species
Denitrovibrio_acetiphilus_522772	Deferribacteres	1	12	0	1	40	species
Desulfotomaculum_carboxydivorans_868595	Firmicutes	2	0	1	0	54;55	species
Desulfurispirillum_indicum_653733	Chrysiogenetes	1	12	0	1	37	species
Elusimicrobium_minutum_445932	Elusimicrobia	1	12	0	1	48	species
Escherichia_coli_574521	Proteobacteria	1	2	0	1	76	species
Fibrobacter_succinogenes_59374	Fibrobacteres	1	12	0	1	49	species
Flexistipes_sinusarabici_717231	Deferribacteres	1	12	0	1	43	species
Fulvivirga_imtechensis_1237149	Bacteroidetes	1	12	0	1	14	species
Gemmatimonas_aurantiaca_379066	Gemmatimonadetes	1	12	0	1	66	species
Geobacter_bemidjiensis_404380	Proteobacteria	1	12	0	1	73	species
Gloeobacter_violaceus_251221	Cyanobacteria	1	2	0	1	38;39	species
Helicobacter_pylori_85963	Proteobacteria	1	12	0	1	82	species
Heliobacterium_modesticaldum_498761	Firmicutes	2	0	1	0	58	species
Herpetosiphon_aurantiacus_316274	Chloroflexi	2	12	0	1	30;31	species
Holophaga_foetida_903818	Acidobacteria	1	12	0	1	1;2	species
Ignavibacterium_album_945713	Ignavibacteriae	1	12	0	1	68	species
Ktedonobacter_racemifer_485913	Chloroflexi	2	0	2	0	33;34	species
Lentisphaera_araneosa_313628	Lentisphaerae	1	12	0	1	69	species
Leptonema_illini_929563	Spirochaetes	1	12	0	1	83	species
Leptospira_interrogans_189518	Spirochaetes	1	2	0	1	84	species
Leptospira_meyeri_1218585	Spirochaetes	1	2	0	1	86	species
Leptospirillum_ferrooxidans_1162668	Nitrospirae	1	12	0	1	71	species
Magnetococcus_marinus_156889	Proteobacteria	1	12	0	1	74	species
Marinithermus_hydrothermalis_869210	Deinococcus-Thermus	1	12	0	1	45	species

Melioribacter_roseus_1191523	Ignavibacteriae	1	12	0	1	67	species
Mycoplasma_penetrans_272633	Tenericutes	0	0	0	0	87	class
Neisseria_meningitidis_662598	Proteobacteria	1	2	0	1	81	species
Nitrosomonas_sp._153948	Proteobacteria	1	12	0	1	78	species
Oceanithermus_profundus_670487	Deinococcus-Thermus	1	12	0	1	44	species
Opitutaceae_bacterium_278956	Verrucomicrobia	1	12	0	1	92	other species
Opitutus_terrae_452637	Verrucomicrobia	1	12	0	1	92	species
Oscillatoriales_cyanobacterium_864702	Cyanobacteria	2	2	0	1	38	phylum
Paenibacillus_polymyxa_886882	Firmicutes	2	0	0	0	56	species
Paludibacter_propionicigenes_694427	Bacteroidetes	1	12	0	1	15	species
Parachlamydia_acanthamoebae_765952	Chlamydiae	12	2	0	1	17-23	species
Pedosphaera_parvula_320771	Verrucomicrobia	1	12	0	1	90	species
Pelosinus_fermentans_1122947	Firmicutes	1	12	1	1	57	species
Planctomyces_brasiliensis_756272	Planctomycetes	1	12	0	1	72	phylum
Planctomyces_maris_344747	Planctomycetes	1	12	0	1	72	phylum
Prochlorococcus_marinus_167539	Cyanobacteria	2	2	0	1	38	phylum
Propionibacterium_freudenreichii_754252	Actinobacteria	2	0	0	0	10;11	species
Rhodothermus_marinus_762570	Bacteroidetes	1	12	0	1	13	species
Rivularia_sp._373994	Cyanobacteria	2	2	0	1	38	phylum
Rubrobacter_xylanophilus_266117	Actinobacteria	2	0	0	0	5;6	species
Singulisphaera_acidiphila_886293	Planctomycetes	1	12	0	1	72	phylum
Sorangium_cellulosum_448385	Proteobacteria	1	1	0	1	77	species
Spiroplasma_chrysopicola_1276227	Tenericutes	0	0	0	0	87	class
Spiroplasma_melliferum_1129368	Tenericutes	0	0	0	0	87	class
Staphylococcus_aureus_282458	Firmicutes	2	0	0	0	64	species
Streptomyces_coelicolor_100226	Actinobacteria	2	0	2	0	7;8	species
Thermobaculum_terrenum_525904	undef	2	0	0	0	94	species
Thermodesulfatator_indicus_667014	Thermodesulfobacteria	1	12	0	1	88	species
Thermodesulfobacterium_geofontis_795359	Thermodesulfobacteria	1	12	0	1	89	species

Thermodesulfobium_narugense_747365	Firmicutes	1	12	0	1	61	species
Thermodesulfovibrio_yellowstonii_289376	Nitrospirae	1	12	0	1	70	species
Thermomicrobium_roseum_309801	Chloroflexi	1	12	0	1	35;36	species
Thermus_scotoductus_743525	Deinococcus-Thermus	1	12	0	1	46	species
Thiorhodospira_sibirica_765914	Proteobacteria	1	12	0	1	75	species
Turneriella_parva_869212	Spirochaetes	1	12	0	1	85	species
Verrucomicrobiae_bacterium_382464	Verrucomicrobia	1	12	0	1	91	species
Waddlia_chondrophila_716544	Chlamydiae	1	12	0	1	17-20;24	species
Xenococcus_sp._102125	Cyanobacteria	2	2	0	1	38	phylum
planctomycete_KSU-1_247490	Planctomycetes	1	12	0	1	72	phylum

Table S3: Details of the data given to BayesTraits for the ancestral trait reconstruction. Trailing numbers after organism names are NCBI Taxonomy identifiers. In the reference column, the reference corresponding to the number can be found in Table S4. Peptidoglycan (PG): 0 = no PG, 1= thin PG, 2 = thick PG; membrane (MBN full): 0 = monoderm, 1 = diderm without LPS, 2 = diderm with LPS; spore: 0 = no spore, 1 = endospore, 2 = exospore, 3 = myxospore; membrane (MBN simple): 0 = monoderm, 1 = diderm.

1	Anderson, Iain et al. “Genome Sequence of the Homoacetogenic Bacterium Holophaga Foetida Type Strain (TMBS4T).” Standards in Genomic Sciences 6.2 (2012): 174–184. PMC. Web. 23 May 2016.
2	Liesack, Werner, et al. "Holophaga foetida gen. nov., sp. nov., a new, homoacetogenic bacterium degrading methoxylated aromatic compounds." Archives of Microbiology 162.1-2 (1994): 85-90.
3	Pankratov, Timofei A., et al. "Substrate-induced growth and isolation of Acidobacteria from acidic Sphagnum peat." The ISME journal 2.5 (2008): 551-560.
4	Kishimoto, Noriaki, Yoshimasa Kosako, and Tatsuo Tano. "Acidobacterium capsulatum gen. nov., sp. nov.: an acidophilic chemoorganotrophic bacterium containing menaquinone from acidic mineral environment." Current Microbiology 22.1 (1991): 1-7.
5	Carreto, Laura, et al. "Rubrobacter xylanophilus sp. nov., a new thermophilic species isolated from a thermally polluted effluent." International Journal of Systematic and Evolutionary Microbiology 46.2 (1996): 460-465.
6	Albuquerque, Luciana, and Milton S. da Costa. "The Family Rubrobacteraceae." The Prokaryotes. Springer Berlin Heidelberg, 2014. 861-866.
7	Del Sol, Ricardo, et al. "Characterization of changes to the cell surface during the life cycle of Streptomyces coelicolor: atomic force microscopy of living cells." Journal of bacteriology 189.6 (2007): 2219-2225.
8	Bentley, S. D., et al. "SCP1, a 356 023 bp linear plasmid adapted to the ecology and developmental biology of its host, Streptomyces coelicolor A3 (2)." Molecular microbiology 51.6 (2004): 1615-1628.

9	Mohangheghi A., Grohmann K., Himmel M., Leighton L., Updegraff D.M. 1986. Isolation and characterization of Acidothermus cellulolyticus, a new genus of Thermophilic, Acidophlic, and Cellulolytic bacteria. USEM 36(3): 435-443.
10	Brüggemann, Holger, et al. "The complete genome sequence of Propionibacterium acnes, a commensal of human skin." Science 305.5684 (2004): 671-673.
11	Koskinen, Patrik, et al. "Complete genome sequence of Propionibacterium freudenreichii DSM 20271T." Standards in genomic sciences 10.1 (2015): 1-6.
12	Glavina Del Rio, Tijana et al. “Complete Genome Sequence of Chitinophaga Pinensis Type Strain (UQM 2034T).” Standards in Genomic Sciences 2.1 (2010): 87–95. PMC. Web. 24 May 2016.
13	ALFREDSSON, GUDNI A., et al. "Rhodothermus marinus, gen. nov., sp. nov., a thermophilic, halophilic bacterium from submarine hot springs in Iceland." Microbiology 134.2 (1988): 299-306.
14	Sharma, Shalley, et al. "Fulvivirga imtechensis sp. nov., a member of the phylum Bacteroidetes." International journal of systematic and evolutionary microbiology 62.9 (2012): 2213-2217.
15	Ueki, Atsuko, et al. "Paludibacter propionicigenes gen. nov., sp. nov., a novel strictly anaerobic, Gram-negative, propionate-producing bacterium isolated from plant residue in irrigated rice-field soil in Japan." International journal of systematic and evolutionary microbiology 56.1 (2006): 39-44.
16	Mori, Koji, et al. "Caldisericum exile gen. nov., sp. nov., an anaerobic, thermophilic, filamentous bacterium of a novel bacterial phylum, Caldiserica phyl. nov., originally called the candidate phylum OP5, and description of Caldisericaceae fam. nov., Caldisericales ord. nov. and Caldisericia classis nov." International journal of systematic and evolutionary microbiology 59.11 (2009): 2894-2898.
17	Matsumoto, A. (1988). Structural characteristics of chlamydial bodies. In: Microbiology of Chlamydia , A. L. Barron, Ed., CRC Press, Boca Raton, Florida, pp. 21±45.
18	Frohlich, Kyla M. et al. “Membrane Vesicle Production by Chlamydia Trachomatis as an Adaptive Response.” Frontiers in Cellular and Infection Microbiology 4 (2014): 73. PMC. Web. 24 May 2016.
19	Everett, Karin DE, Robin M. Bush, and Arthur A. Andersen. "Emended description of the order Chlamydiales, proposal of Parachlamydiaceae fam. nov. and Simkaniaceae fam. nov., each containing one monotypic genus, revised taxonomy of the family Chlamydiaceae, including a new genus and five new species, and standards for the identification of organisms." International Journal of Systematic and Evolutionary Microbiology 49.2 (1999): 415-440.
20	van Teeseling, Muriel CF, et al. "Anammox Planctomycetes have a peptidoglycan cell wall." Nature communications 6 (2015).
21	Greub, G. (2009). Parachlamydia acanthamoebae, an emerging agent of pneumonia. Clinical Microbiology and Infection, 15(1), 18-28.
22	Michel, Rolf, et al. "Acanthamoeba from human nasal mucosa infected with an obligate intracellular parasite." European journal of protistology 30.1 (1994): 104-110.
23	Amann, R., Springer, N., Schönhuber, W., Ludwig, W., Schmid, E. N., Müller, K. D., & Michel, R. (1997). Obligate intracellular bacterial parasites of acanthamoebae related to Chlamydia spp. Applied and environmental microbiology, 63(1), 115-121.
24	Jacquier, Nicolas, et al. "Cell wall precursors are required to organize the chlamydial division septum." Nature communications 5 (2014).
25	Everett, K D, and T P Hatch. “Architecture of the Cell Envelope of Chlamydia Psittaci 6BC.” Journal of Bacteriology 177.4 (1995): 877–882. Print.

26	Larsen, Helge. "On the culture and general physiology of the green sulfur bacteria." Journal of bacteriology 64.2 (1952): 187.
27	Gibson, Jane, Norbert Pfennig, and John B. Waterbury. "Chloroherpeton thalassium gen. nov. et spec. nov., a non-filamentous, flexing and gliding green sulfur bacterium." Archives of microbiology 138.2 (1984): 96-101.
28	Pfennig, Norbert. "Chlorobium phaeobacteroides nov. spec. und C. phaeovibrioides nov. spec., zwei neue Arten der grünen Schwefelbakterien." Archiv für Mikrobiologie 63.3 (1968): 224-226.
29	Vogl, Kajetan, et al. "Chlorobium chlorochromatii sp. nov., a symbiotic green sulfur bacterium isolated from the phototrophic consortium “Chlorochromatium aggregatum”." Archives of microbiology 185.5 (2006): 363-372.
30	Reichenbach, Hans, and Jochen R. Golecki. "The fine structure of Herpetosiphon, and a note on the taxonomy of the genus." Archives of microbiology 102.1 (1975): 281-291.
31	Holt, J. G., and R. A. Lewin. "Herpetosiphon aurantiacus gen. et sp. n., a new filamentous gliding organism." Journal of bacteriology 95.6 (1968): 2407.
32	Sekiguchi, Yuji, et al. "Anaerolinea thermophila gen. nov., sp. nov. and Caldilinea aerophila gen. nov., sp. nov., novel filamentous thermophiles that represent a previously uncultured lineage of the domain Bacteria at the subphylum level." International journal of systematic and evolutionary microbiology 53.6 (2003): 1843-1851.
33	Chang, Yun-Juan, et al. "Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium Ktedonobacter racemifer type strain (SOSP1-21T)." Standards in genomic sciences 5.1 (2011): 97-111.
34	Cavaletti, Linda, et al. "New lineage of filamentous, spore-forming, gram-positive bacteria from soil." Applied and environmental microbiology 72.6 (2006): 4360-4369.
35	Wu, Dongying, et al. "Complete genome sequence of the aerobic CO-oxidizing thermophile Thermomicrobium roseum." PLoS One 4.1 (2009): e4207.
36	Merkel, G. J., D. R. Durham, and J. J. Perry. "The atypical cell wall composition of Thermomicrobium roseum." Canadian journal of microbiology 26.4 (1980): 556-559.
37	Rauschenbach, Ines, Priya Narasingarao, and Max M. Häggblom. "Desulfurispirillum indicum sp. nov., a selenate-and selenite-respiring bacterium isolated from an estuarine canal." International journal of systematic and evolutionary microbiology 61.3 (2011): 654-658.
38	Hoiczyk, Egbert, and Alfred Hansel. "Cyanobacterial cell walls: news from an unusual prokaryotic envelope." Journal of Bacteriology 182.5 (2000): 1191-1199.
39	Schneider, Sabine, and Uwe J. Jürgens. "Cell wall and sheath constituents of the cyanobacterium Gloeobacter violaceus." Archives of microbiology 156.4 (1991): 312-318.
40	Myhr, Siri, and Terje Torsvik. "Denitrovibrio acetiphilus, a novel genus and species of dissimilatory nitrate-reducing bacterium isolated from an oil reservoir model column." International journal of systematic and evolutionary microbiology 50.4 (2000): 1611-1619.
41	Takai, Ken, et al. "Deferribacter desulfuricans sp. nov., a novel sulfur-, nitrate-and arsenate-reducing thermophile isolated from a deep-sea hydrothermal vent." International journal of systematic and evolutionary microbiology 53.3 (2003): 839-846.
42	Iino, Takao, et al. "Calditerrivibrio nitroreducens gen. nov., sp. nov., a thermophilic, nitrate-reducing bacterium isolated from a terrestrial hot spring in Japan." International journal of systematic and evolutionary microbiology 58.7 (2008): 1675-1679.
43	Fiala, Gerhard, et al. "Flexistipes sinusarabici, a novel genus and species of eubacteria occurring in the Atlantis II Deep brines of the Red Sea." Archives of Microbiology 154.2 (1990): 120-126.
44	Miroshnichenko, M. L., et al. "Oceanithermus profundus gen. nov., sp. nov., a thermophilic,

	microaerophilic, facultatively chemolithoheterotrophic bacterium from a deep-sea hydrothermal vent." International journal of systematic and evolutionary microbiology 53.3 (2003): 747-752.
45	Sako, Yoshihiko, et al. "Marinithermus hydrothermalis gen. nov., sp. nov., a strictly aerobic, thermophilic bacterium from a deep-sea hydrothermal vent chimney." International Journal of Systematic and Evolutionary Microbiology 53.1 (2003): 59-65.
46	Henne, Anke, et al. "The genome sequence of the extreme thermophile Thermus thermophilus." Nature biotechnology 22.5 (2004): 547-553.
47	Brooks, B. W., and R. G. E. Murray. "Nomenclature for “Micrococcus radiodurans” and other radiation-resistant cocci: Deinococcaceae fam. nov. and Deinococcus gen. nov., including five species." International Journal of Systematic and Evolutionary Microbiology 31.3 (1981): 353-360.
48	Geissinger, Oliver, et al. "The ultramicrobacterium “Elusimicrobium minutum” gen. nov., sp. nov., the first cultivated representative of the termite group 1 phylum." Applied and environmental microbiology 75.9 (2009): 2831-2840.
49	Jun, H. S., et al. "Fibrobacter succinogenes, a dominant fibrolytic ruminal bacterium: transition to the post genomic era." ASIAN AUSTRALASIAN JOURNAL OF ANIMAL SCIENCES 20.5 (2007): 802.
50	Driks, A. "Overview: development in bacteria: spore formation in Bacillus subtilis." Cellular and Molecular Life Sciences CMLS 59.3 (2002): 389-391.
51	Sekiguchi, Junichi, and Hiroki Yamamoto. "4-3. Cell wall structure of E. coli and B. subtilis."
52	Kay, D., and S. C. Warren. "Sporulation in Bacillus subtilis. Morphological changes." Biochemical Journal 109.5 (1968): 819-824.
53	Tocheva, Elitza I., et al. "Peptidoglycan transformations during Bacillus subtilis sporulation." Molecular microbiology 88.4 (2013): 673-686.
54	Parshina, Sofiya N., et al. "Desulfotomaculum carboxydivorans sp. nov., a novel sulfate-reducing bacterium capable of growth at 100% CO." International Journal of Systematic and Evolutionary Microbiology 55.5 (2005): 2159-2165.
55	Campbell, L. LEON, and JOHN R. Postgate. "Classification of the spore-forming sulfate-reducing bacteria." Bacteriological reviews 29.3 (1965): 359.
56	He, Zengguo et al. “Isolation and Identification of a Paenibacillus Polymyxa Strain That Coproduces a Novel Lantibiotic and Polymyxin .” Applied and Environmental Microbiology 73.1 (2007): 168–178. PMC. Web. 25 May 2016.
57	Shelobolina, Evgenya S., et al. "Geobacter pickeringii sp. nov., Geobacter argillaceus sp. nov. and Pelosinus fermentans gen. nov., sp. nov., isolated from subsurface kaolin lenses." International Journal of Systematic and Evolutionary Microbiology 57.1 (2007): 126-135.
58	Tang, Kuo-Hsiang, Hai Yue, and Robert E. Blankenship. "Energy metabolism of Heliobacterium modesticaldum during phototrophic and chemotrophic growth." BMC microbiology 10.1 (2010): 1.
59	Sikorski, Johannes et al. “Complete Genome Sequence of Acetohalobium Arabaticum Type Strain (Z-7288T).” Standards in Genomic Sciences 3.1 (2010): 57–65. PMC. Web. 25 May 2016.
60	Tocheva, Elitza I., et al. "Peptidoglycan remodeling and conversion of an inner membrane into an outer membrane during sporulation." Cell 146.5 (2011): 799-812.
61	Mori, Koji, et al. "A novel lineage of sulfate-reducing microorganisms: Thermodesulfobiaceae fam. nov., Thermodesulfobium narugense, gen. nov., sp. nov., a new thermophilic isolate from a hot spring." Extremophiles 7.4 (2003): 283-290.
62	MURRAY, WILLIAM D., and A. W. Khan. "Clostridium saccharolyticum sp. nov., a Saccharolytic Species from Sewage Sludge†." International Journal of Systematic and Evolutionary Microbiology 32.1 (1982): 132-135.

63	Van Gylswyk, N. O., E. JANE MORRIS, and H. J. Els. "Sporulation and cell wall structure of Clostridium polysaccharolyticum comb. nov.(formerly Fusobacterium polysaccharolyticum)." Microbiology 121.2 (1980): 491-493.
64	Umeda, A. K. I. K. O., Y. U. J. I. Ueki, and K. A. Z. U. N. O. B. U. Amako. "Structure of the Staphylococcus aureus cell wall determined by the freeze-substitution method." Journal of bacteriology 169.6 (1987): 2482-2487.
65	Kosowska, Klaudia, et al. "The Clostridium ramosum IgA proteinase represents a novel type of metalloendopeptidase." Journal of Biological Chemistry 277.14 (2002): 11987-11994.
66	Zhang, Hui, et al. "Gemmatimonas aurantiaca gen. nov., sp. nov., a Gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum Gemmatimonadetes phyl. nov." International journal of systematic and evolutionary microbiology 53.4 (2003): 1155-1163.
67	Podosokorskaya, Olga A., et al. "Characterization of Melioribacter roseus gen. nov., sp. nov., a novel facultatively anaerobic thermophilic cellulolytic bacterium from the class Ignavibacteria, and a proposal of a novel bacterial phylum Ignavibacteriae." Environmental microbiology 15.6 (2013): 1759-1771.
68	Iino, Takao, et al. "Ignavibacterium album gen. nov., sp. nov., a moderately thermophilic anaerobic bacterium isolated from microbial mats at a terrestrial hot spring and proposal of Ignavibacteria classis nov., for a novel lineage at the periphery of green sulfur bacteria." International journal of systematic and evolutionary microbiology 60.6 (2010): 1376-1382.
69	Cho, Jang-Cheon, et al. "Lentisphaera araneosa gen. nov., sp. nov, a transparent exopolymer producing marine bacterium, and the description of a novel bacterial phylum, Lentisphaerae." Environmental Microbiology 6.6 (2004): 611-621.
70	Henry, E. A., et al. "Characterization of a new thermophilic sulfate-reducing bacterium." Archives of Microbiology 161.1 (1994): 62-69.
71	Hippe, Hans. "Leptospirillum gen. nov.(ex Markosyan 1972), nom. rev., including Leptospirillum ferrooxidans sp. nov.(ex Markosyan 1972), nom. rev. and Leptospirillum thermoferrooxidans sp. nov.(Golovacheva et al. 1992)." International journal of systematic and evolutionary microbiology 50.2 (2000): 501-503.
72	van Teeseling, Muriel CF, et al. "Anammox Planctomycetes have a peptidoglycan cell wall." Nature communications 6 (2015).
73	Nevin, Kelly P., et al. "Geobacter bemidjiensis sp. nov. and Geobacter psychrophilus sp. nov., two novel Fe (III)-reducing subsurface isolates." International Journal of Systematic and Evolutionary Microbiology 55.4 (2005): 1667-1674.
74	Bazylinski, Dennis A., et al. "Magnetococcus marinus gen. nov., sp. nov., a marine, magnetotactic bacterium that represents a novel lineage (Magnetococcaceae fam. nov., Magnetococcales ord. nov.) at the base of the Alphaproteobacteria." International journal of systematic and evolutionary microbiology 63.3 (2013): 801-808.
75	Bryantseva, Irina, et al. "Thiorhodospira sibirica gen. nov., sp. nov., a new alkaliphilic purple sulfur bacterium from a Siberian soda lake." International Journal of Systematic and Evolutionary Microbiology 49.2 (1999): 697-703.
76	Matias, Valério RF, et al. "Cryo-transmission electron microscopy of frozen-hydrated sections of Escherichia coli and Pseudomonas aeruginosa." Journal of Bacteriology 185.20 (2003): 6112-6118.
77	Keck, Matthias, et al. "Unusual outer membrane lipid composition of the gram-negative, lipopolysaccharide-lacking myxobacterium Sorangium cellulosum So ce56." Journal of Biological Chemistry 286.15 (2011): 12850-12859.

78	Koops, H. P., et al. "Classification of eight new species of ammonia-oxidizing bacteria: Nitrosomonas communis sp. nov., Nitrosomonas ureae sp. nov., Nitrosomonas aestuarii sp. nov., Nitrosomonas marina sp. nov., Nitrosomonas nitrosa sp. nov., Nitrosomonas eutropha sp. nov., Nitrosomonas oligotropha sp. nov. and Nitrosomonas halophila sp. nov." Microbiology 137.7 (1991): 1689-1699.
79	Pati, Amrita, et al. "Complete genome sequence of Arcobacter nitrofigilis type strain (CIT)." Standards in genomic sciences 2.3 (2010): 300-308.
80	Curtis, Patrick D., and Yves V. Brun. "Getting in the loop: regulation of development in Caulobacter crescentus." Microbiology and Molecular Biology Reviews 74.1 (2010): 13-41.
81	Hill, Darryl J., et al. "Cellular and molecular biology of Neisseria meningitidis colonization and invasive disease." Clinical Science 118.9 (2010): 547-564.
82	O'Toole PW, Clyne M. Cell Envelope. In: Mobley HLT, Mendz GL, Hazell SL, editors. Helicobacter pylori: Physiology and Genetics. Washington (DC): ASM Press
83	Huntemann, Marcel, et al. "Genome sequence of the phylogenetically isolated spirochete Leptonema illini type strain (3055T)." Standards in genomic sciences 8.2 (2013): 177-187.
84	Johnson RC. Leptospira. In: Baron S, editor. Medical Microbiology. 4th edition. Galveston (TX): University of Texas Medical Branch at Galveston
85	Stackebrandt, Erko, et al. "Genome sequence of the free-living aerobic spirochete Turneriella parva type strain (HT), and emendation of the species Turneriella parva." Standards in genomic sciences 8.2 (2013): 228-238.
86	Evangelista, Karen V, and Jenifer Coburn. “Leptospira as an Emerging Pathogen: A Review of Its Biology, Pathogenesis and Host Immune Responses.” Future microbiology 5.9 (2010): 1413–1425. PMC. Web. 25 May 2016.
87	Razin, Shmuel. Mycoplasma. John Wiley & Sons, Ltd, 1979.
88	Moussard, H., et al. "Thermodesulfatator indicus gen. nov., sp. nov., a novel thermophilic chemolithoautotrophic sulfate-reducing bacterium isolated from the Central Indian Ridge." International journal of systematic and evolutionary microbiology 54.1 (2004): 227-233.
89	Hamilton-Brehm, Scott D., et al. "Thermodesulfobacterium geofontis sp. nov., a hyperthermophilic, sulfate-reducing bacterium isolated from Obsidian Pool, Yellowstone National Park." Extremophiles 17.2 (2013): 251-263.
90	Kant, Ravi, et al. "Genome sequence of Pedosphaera parvula Ellin514, an aerobic verrucomicrobial isolate from pasture soil." Journal of bacteriology (2011).
91	Schlesner, Heinz. The Prokaryotes: An Evolving Electronic Resource for the Mircobiological Community: The Genus Verrucomicrobium. 2004. Springer-Verlag New York, LLC.
92	Chin, Kuk-Jeong, Werner Liesack, and Peter H. Janssen. "Opitutus terrae gen. nov., sp. nov., to accommodate novel strains of the division'Verrucomicrobia'isolated from rice paddy soil." International journal of systematic and evolutionary microbiology 51.6 (2001): 1965-1968.
93	Miroshnichenko, Margarita L., et al. "Caldithrix abyssi gen. nov., sp. nov., a nitrate-reducing, thermophilic, anaerobic bacterium isolated from a Mid-Atlantic Ridge hydrothermal vent, represents a novel bacterial lineage." International journal of systematic and evolutionary microbiology 53.1 (2003): 323-329.
94	Botero, Lina M., et al. "Thermobaculum terrenum gen. nov., sp. nov.: a non-phototrophic gram-positive thermophile representing an environmental clone group related to the Chloroflexi (green non-sulfur bacteria) and Thermomicrobia." Archives of microbiology 181.4 (2004): 269-277.

Table S4: List of references used to determine the cell-wall architecture for the 85 representative organisms of Figure 1. To be used with Table S3.

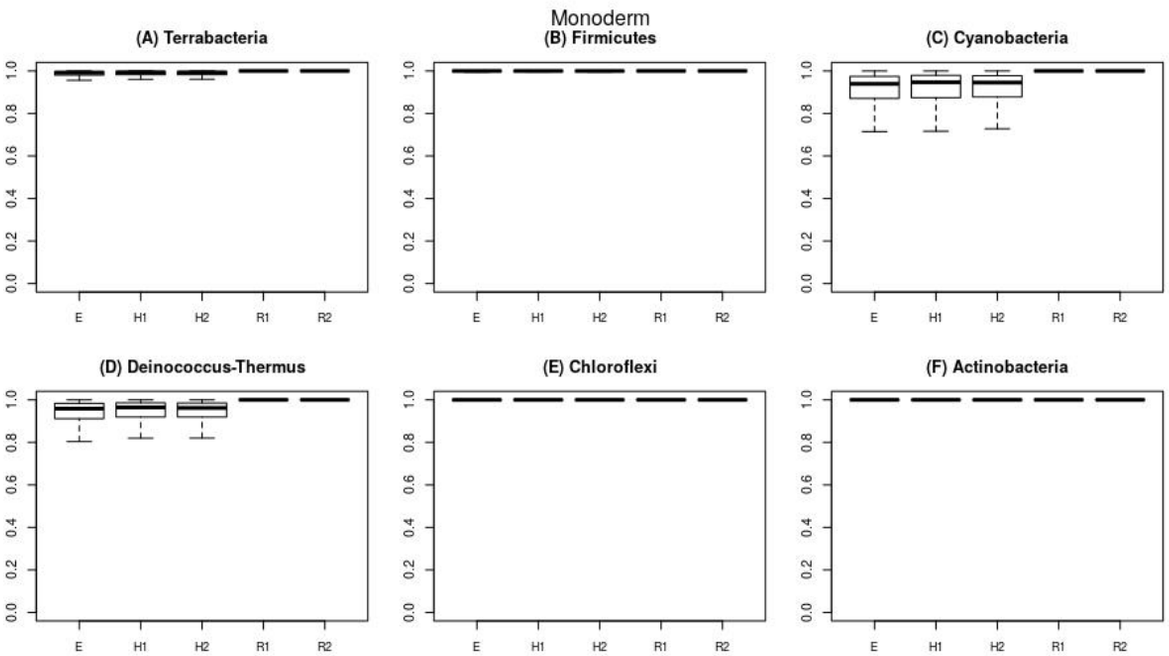


Figure S7: Posterior probabilities for a monoderm LBCA according to five different models, prior exponential of 10 (E), hyperprior exponential 0 to 10 (H1), hyperprior exponential 0 to 100 (H2), reverse jump hyperprior exponential 0 to 10 (R1) and reverse jump hyperprior exponential 0 to 100 (R2), and six possible roots for the bacterial domain (Terrabacteria, Firmicutes, Cyanobacteria, Deinococcus-Thermus, Chloroflexi and Actinobacteria).

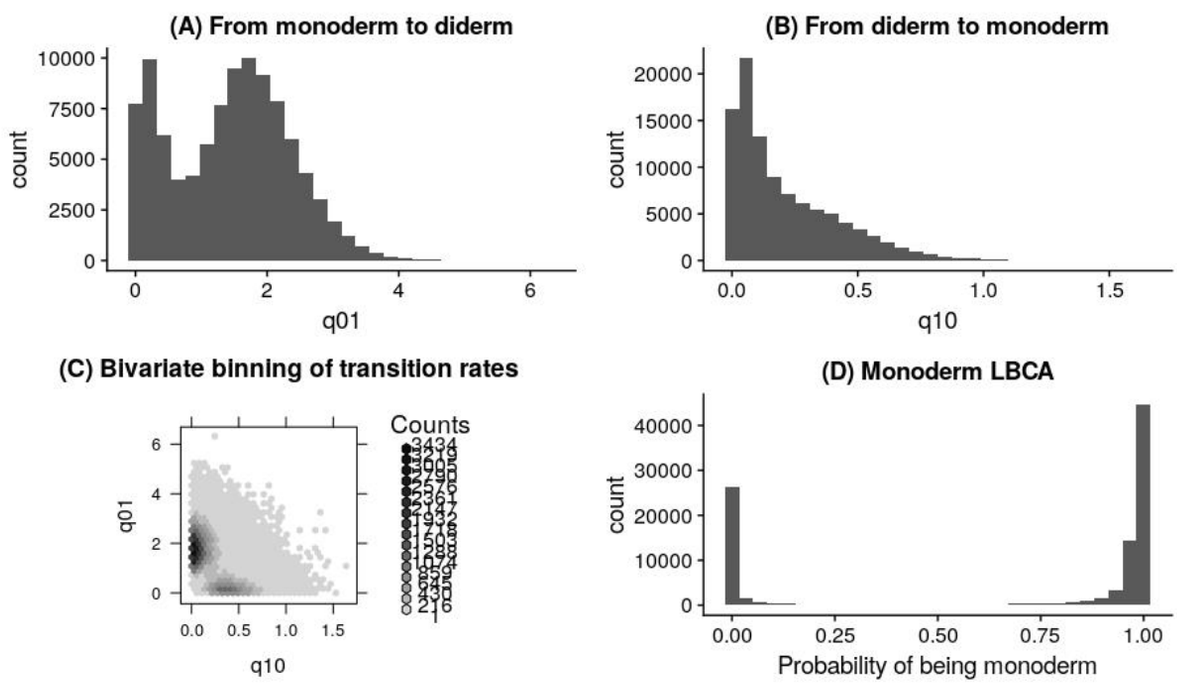


Figure S8: Posterior transition rates and posterior probability of being monoderm for the model where the hyper-prior was purposely biased towards the “diderm-first” hypothesis. “q01” is the transition rate from monoderm to diderm (limited) and “q10” is the transition rate from diderm to monoderm (favored). LBCA = Last Bacterial Common Ancestor.

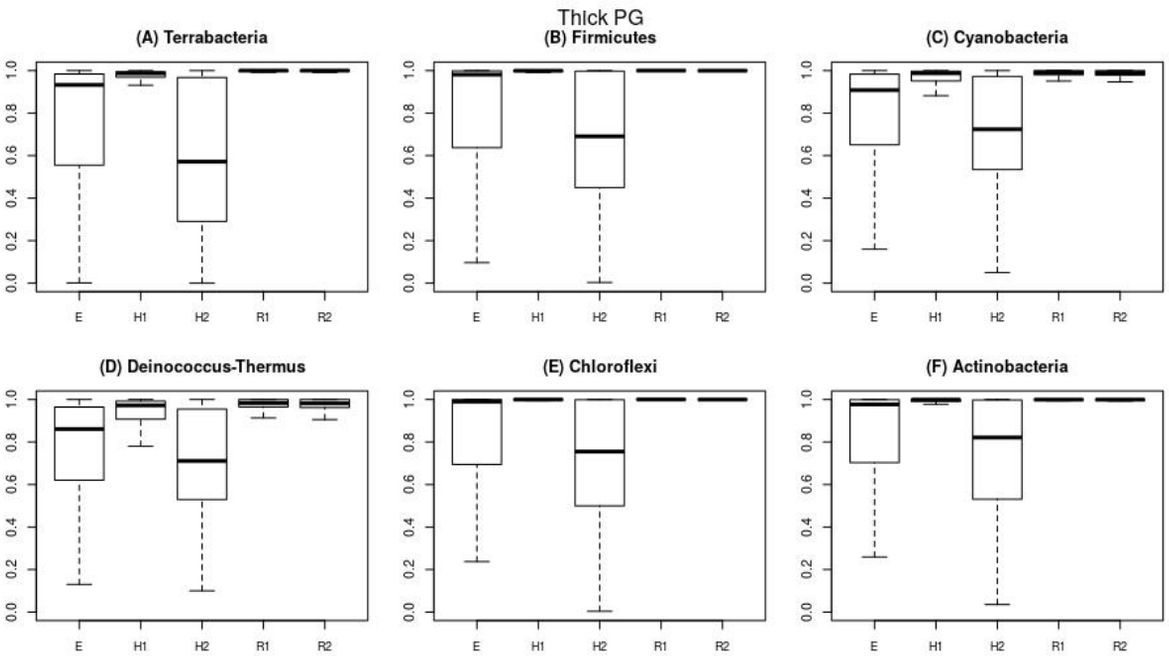


Figure S9: Posterior probabilities for a LBCA featuring a thick peptidoglycan (PG) layer according to the five different models and the six possible bacterial roots.

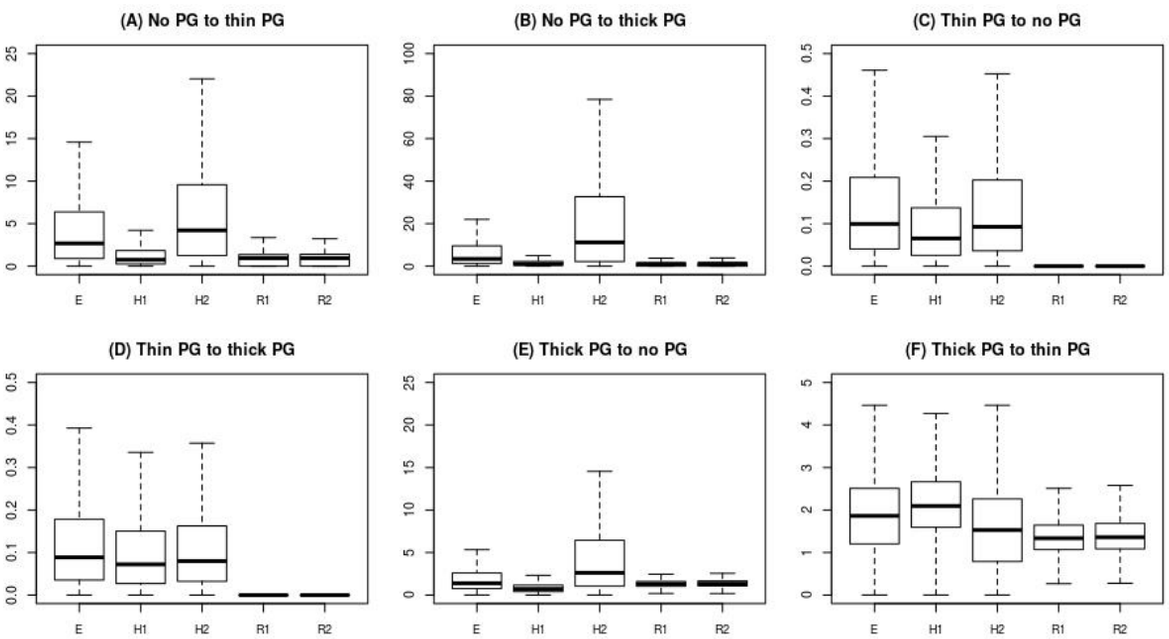


Figure S10: Posterior transition rates for the peptidoglycan (PG) trait. The Terrabacteria root was used for the five models.

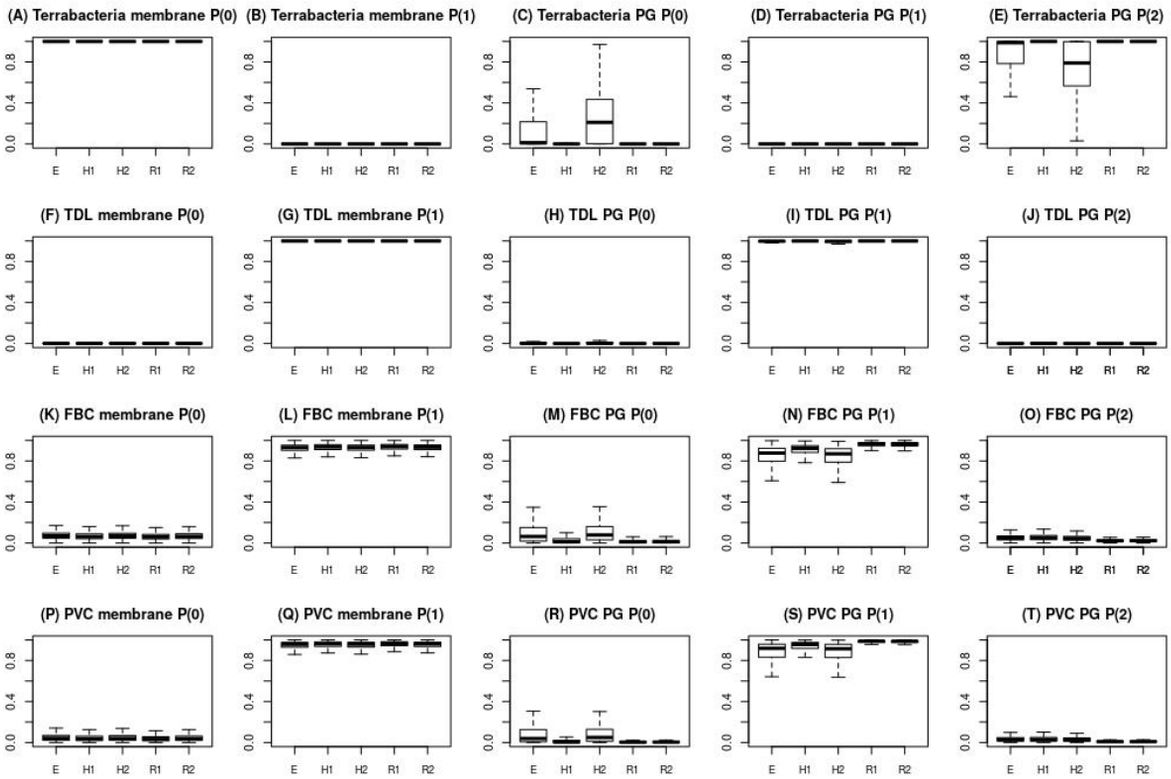


Figure S11: Posterior probabilities for the peptidoglycan (PG) and membrane state in the LCA of four bacterial groups. Membrane P(0) and P(1) correspond to one and two membranes, respectively, whereas PG P(0), P(1) and P(2) correspond to no PG, thin PG and thick PG, respectively. The Terrabacteria root was used for the five models.

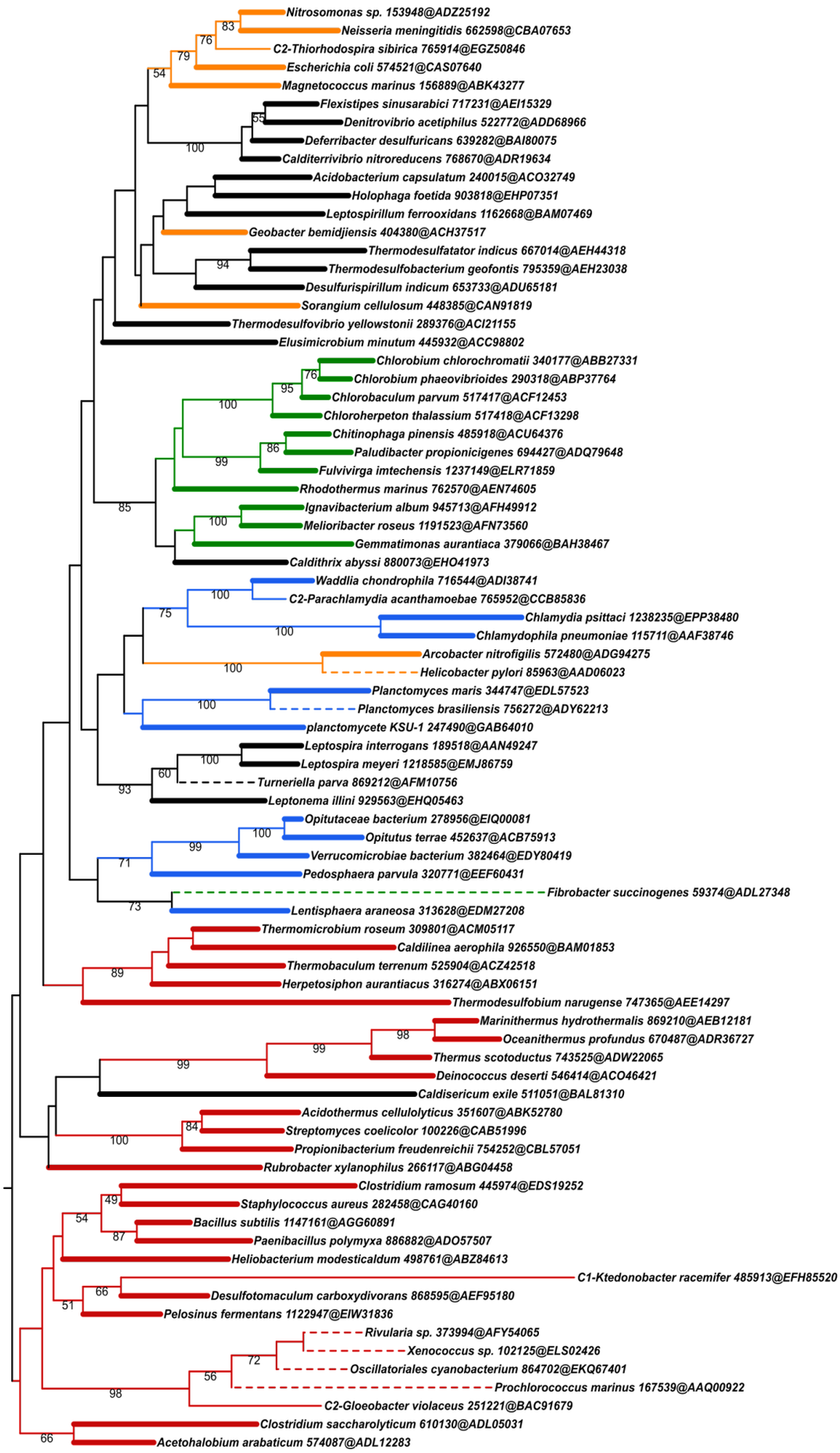

Nitrosomonas sp. 153948@ADZ25192
Neisseria meningitidis 662598@CBA07653
C2-Thiorhodospira sibirica 765914@EGZ50846
Escherichia coli 574521@CAS07640
Magnetococcus marinus 156889@ABK43277
Flexistipes sinusarabici 717231@AEI15329
Denitrovibrio acetiphilus 522772@ADD68966
Deferribacter desulfuricans 639282@BAI80075
Calditerrivibrio nitroreducens 768670@ADR19634
Acidobacterium capsulatum 240015@ACO32749
Holophaga foetida 903818@EHP07351
Leptospirillum ferrooxidans 1162668@BAM07469
Geobacter bemidjiensis 404380@ACH37517
Thermodesulfatator indicus 667014@AEH44318
Thermodesulfobacterium geofontis 795359@AEH23038
Desulfurispirillum indicum 653733@ADU65181
Sorangium cellulosum 448385@CAN91819
Thermodesulfovibrio yellowstonii 289376@ACI21155
Elusimicrobium minutum 445932@ACC98802
Chlorobium chlorochromatii 340177@ABB27331
Chlorobium phaeovibrioides 290318@ABP37764
Chlorobaculum parvum 517417@ACF12453
Chloroherpeton thalassium 517418@ACF13298
Chitinophaga pinensis 485918@ACU64376
Paludibacter propionicigenes 694427@ADQ79648
Fulvivirga imtechensis 1237149@ELR71859
Rhodothermus marinus 762570@AEN74605
Ignavibacterium album 945713@AFH49912
Melioribacter roseus 1191523@AFN73560
Gemmatimonas aurantiaca 379066@BAH38467
Caldithrix abyssi 880073@EHO41973
Waddlia chondrophila 716544@ADI38741
C2-Parachlamydia acanthamoebae 765952@CCB85836
Chlamydia psittaci 1238235@EPP38480
Chlamydophila pneumoniae 115711@AAF38746
Arcobacter nitrofigilis 572480@ADG94275
Helicobacter pylori 85963@AAD06023
Planctomyces maris 344747@EDL57523
Planctomyces brasiliensis 756272@ADY62213
planctomycete KSU-1 247490@GAB64010
Leptospira interrogans 189518@AAN49247
Leptospira meyeri 1218585@EMJ86759
Turneriella parva 869212@AFM10756
Leptonema illini 929563@EHQ05463
Opitutaceae bacterium 278956@EIQ00081
Opitutus terrae 452637@ACB75913
Verrucomicrobiae bacterium 382464@EDY80419
Pedosphaera parvula 320771@EEF60431
Fibrobacter succinogenes 59374@ADL27348
Lentisphaera araneosa 313628@EDM27208
Thermomicrobium roseum 309801@ACM05117
Caldilinea aerophila 926550@BAM01853
Thermobaculum terrenum 525904@ACZ42518
Herpetosiphon aurantiacus 316274@ABX06151
Thermodesulfobium narugense 747365@AEE14297
Marinithermus hydrothermalis 869210@AEB12181
Oceanithermus profundus 670487@ADR36727
Thermus scotoductus 743525@ADW22065
Deinococcus deserti 546414@ACO46421
Caldisericum exile 511051@BAL81310
Acidothermus cellulolyticus 351607@ABK52780
Streptomyces coelicolor 100226@CAB51996
Propionibacterium freudenreichii 754252@CBL57051
Rubrobacter xylanophilus 266117@ABG04458
Clostridium ramosum 445974@EDS19252
Staphylococcus aureus 282458@CAG40160
Bacillus subtilis 1147161@AGG60891
Paenibacillus polymyxa 886882@ADO57507
Heliobacterium modesticaldum 498761@ABZ84613
C1-Ktedonobacter racemifer 485913@EFH85520
Desulfotomaculum carboxydivorans 868595@AEF95180
Pelosinus fermentans 1122947@EIW31836
Rivularia sp. 373994@AFY54065
Xenococcus sp. 102125@ELS02426
Oscillatoriales cyanobacterium 864702@EKQ67401
Prochlorococcus marinus 167539@AAQ00922
C2-Gloeobacter violaceus 251221@BAC91679
Clostridium saccharolyticum 610130@ADL05031
Acetohalobium arabaticum 574087@ADL12283

Figure S12: MraY tree inferred using RAxML v8.1.17 under the LG+F+Γ model of sequence evolution (see "DCW_17_SG.pdf" available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder Trees, for the remaining *dcw* trees). Thick branches indicate a gene present in the main cluster (longest cluster), while thin branches indicate a gene present in a sub-cluster (the different sub-clusters are numbered following the nomenclature "Cn" with "n" being the number of the cluster) and dotted branches indicate a gene located outside of any cluster.

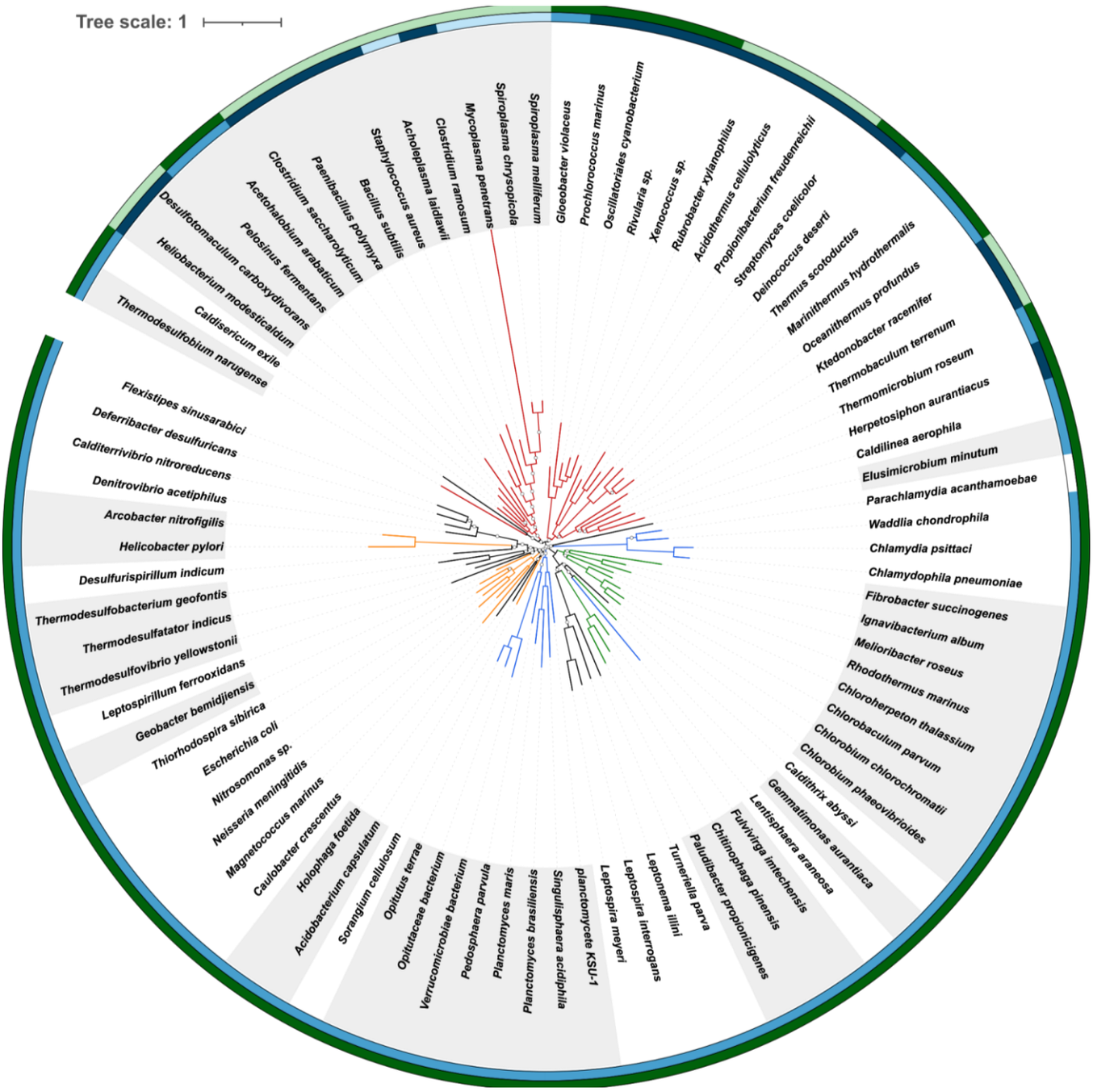


Figure S13: Phylogenomic tree based on a supermatrix of 85 species x 4571 unambiguously aligned amino-acid positions (8.47% missing character states) using 15 of the *dcw* cluster genes. PhyloBayes MPI v1.4 was used to run two MCMC chains under the CAT+Γ model for 50,000 cycles. Both chains were used to compute the consensus tree (maxdiff = 0.284; meandiff = 0.007).

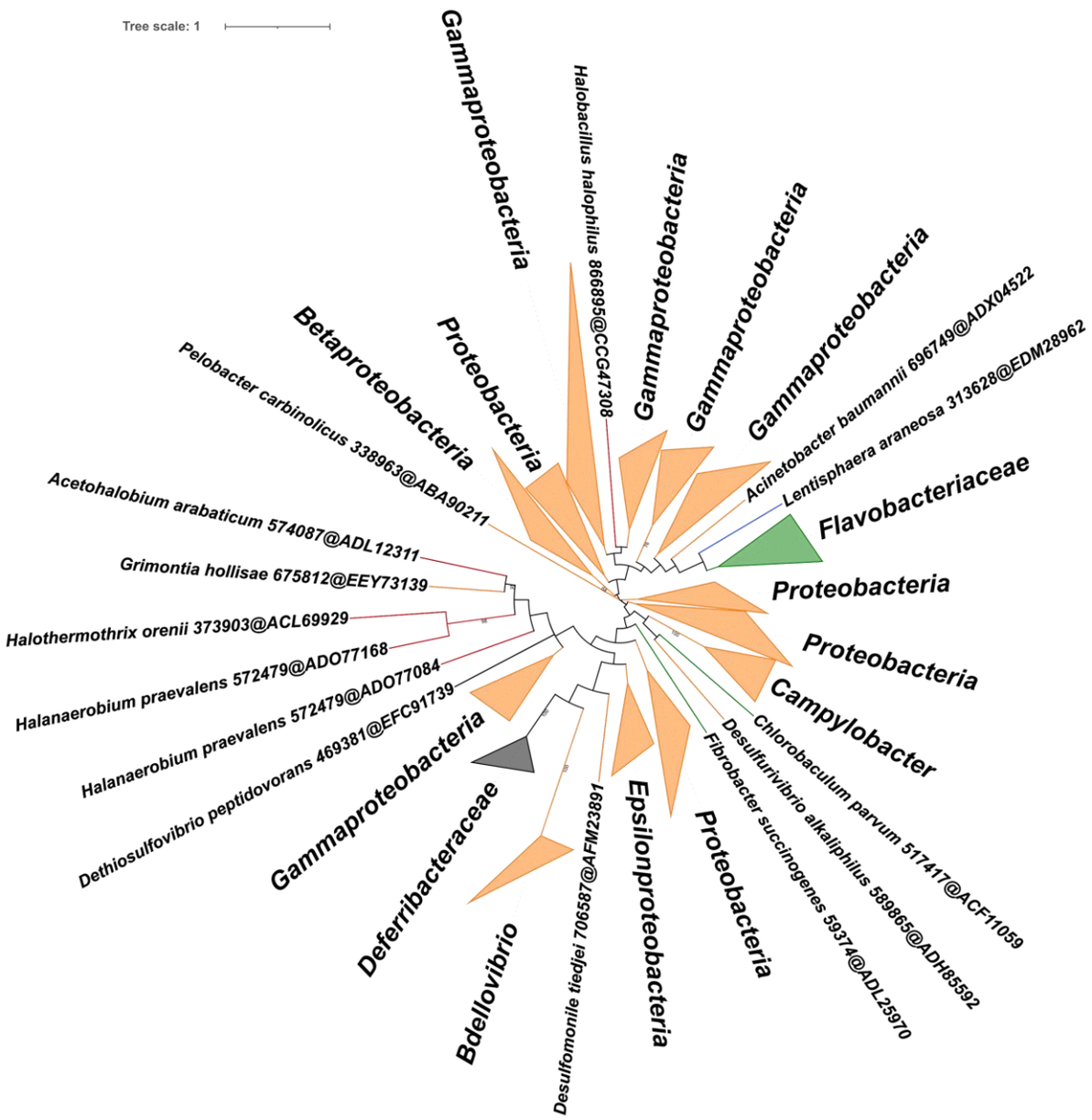


Figure S14: TolA tree inferred using RAxML under the LG+F+Γ model.

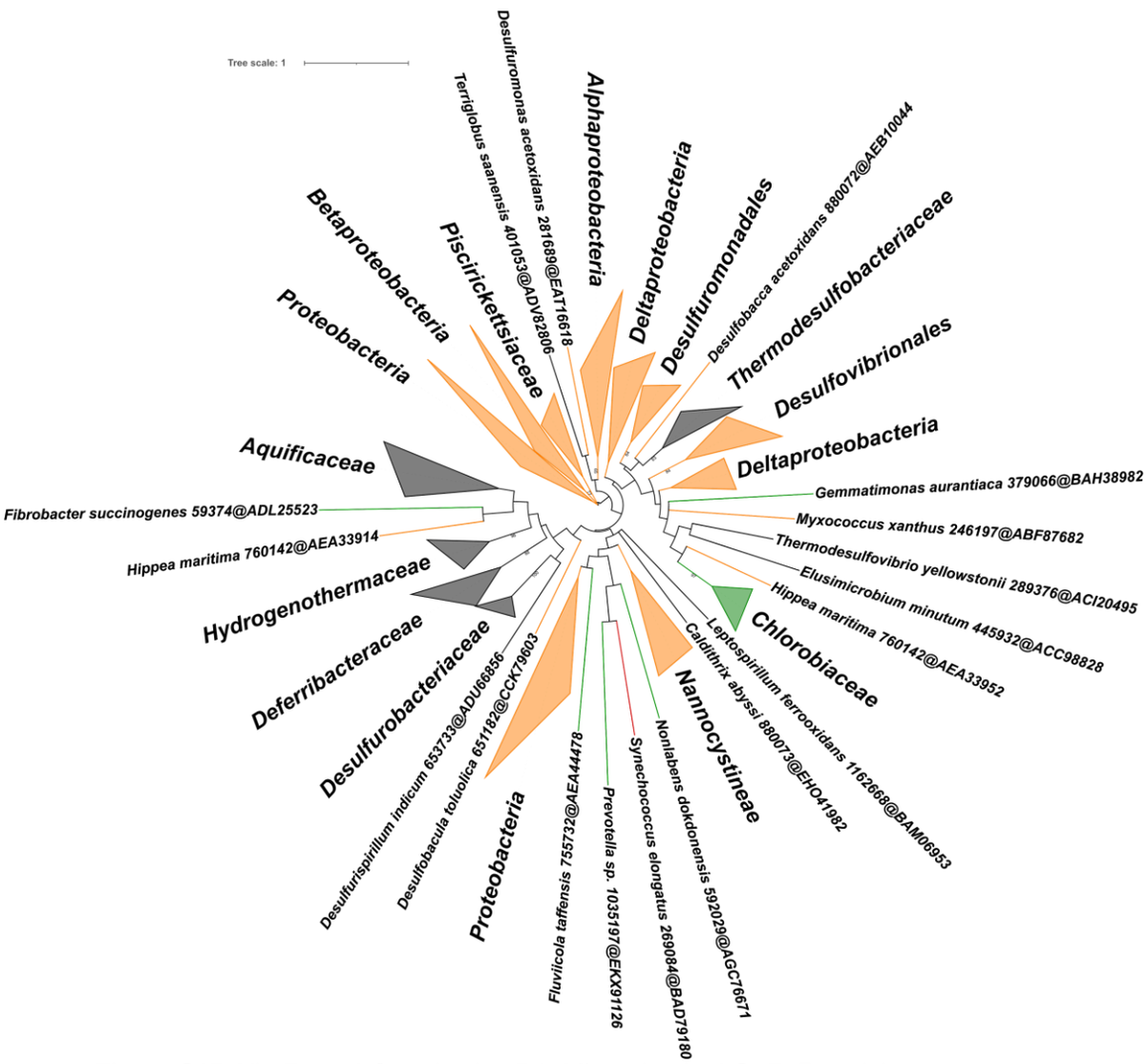


Figure S15: YbgF tree inferred using RAxML under the LG+F+Γ model.

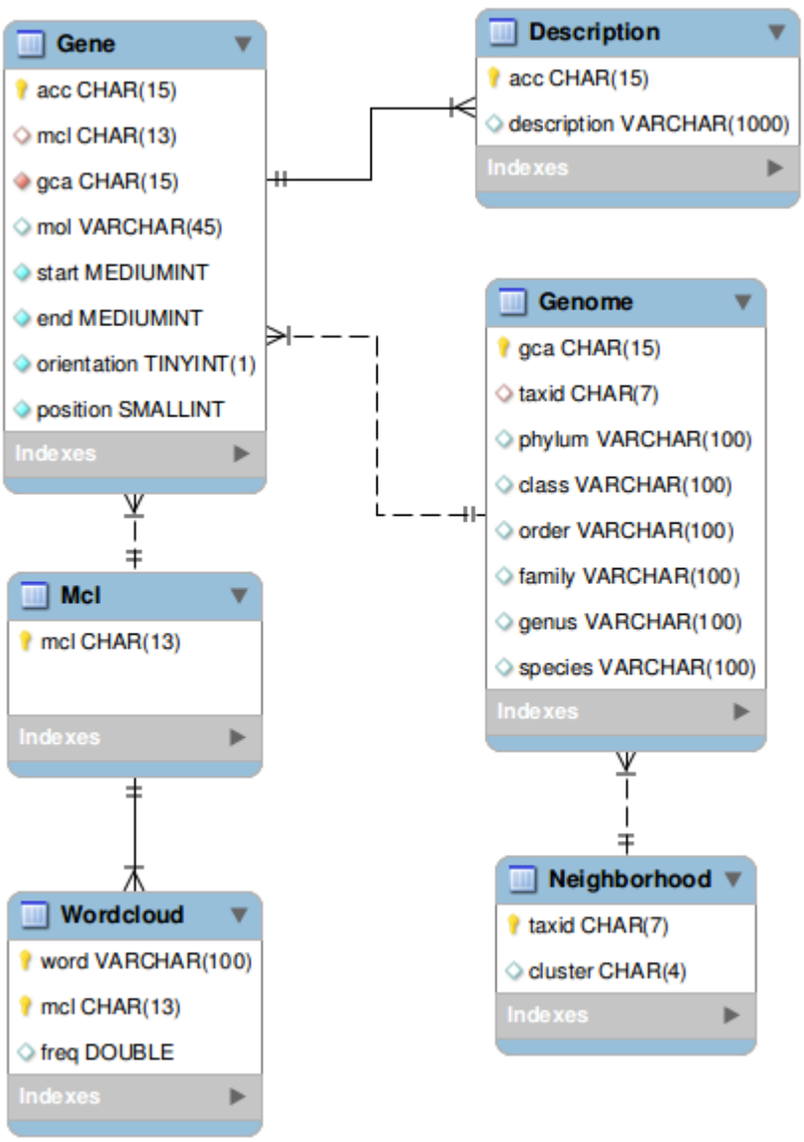


Figure S16: Schema of the MySQL database used by the synteny tool.

OG	# species	# AA	description
MCLdcw110100	100	246	L2
MCLdcw110104	101	175	S3
MCLdcw110105	100	138	S5
MCLdcw110107	100	126	L11
MCLdcw110109	98	111	hydrolase, TatD family
MCLdcw110112	99	109	L14
MCLdcw110114	100	76	S19
MCLdcw110116	101	451	translation initiation factor IF-2
MCLdcw110118	101	190	metalloendopeptidase, glycoprotease family

MCLdcw110124	100	116	S11
MCLdcw110125	96	246	cysteinyl-tRNA synthetase
MCLdcw110131	98	110	dimethyladenosine transferase
MCLdcw110132	100	200	DNA polymerase III, subunits gamma and tau
MCLdcw110139	96	383	GMP synthase
MCLdcw110140	97	124	tRNA pseudouridine synthase B
MCLdcw110159	98	124	S4
MCLdcw110162	98	109	tRNA dimethylallyltransferase
MCLdcw110169	94	292	Methionine adenosyltransferase
MCLdcw110172	98	372	aspartyl-tRNA synthetase
MCLdcw110178	100	188	DNA-directed RNA polymerase, alpha subunit
MCLdcw110179	98	446	CTP synthase
MCLdcw110188	99	118	S12
MCLdcw110189	100	153	L3
MCLdcw110190	99	92	L4/L1e
MCLdcw110192	99	189	S2
MCLdcw110195	101	126	L16
MCLdcw110198	100	276	UvrABC system protein C
MCLdcw110199	100	245	Peptide chain release factor 1
MCLdcw110202	101	69	L27
MCLdcw110204	100	120	L6
MCLdcw110205	100	87	L15
MCLdcw110206	99	90	L7/L12
MCLdcw110208	100	95	S9
MCLdcw110209	100	134	S7
MCLdcw110210	100	174	L5
MCLdcw110211	98	277	GTP-binding protein EngA
MCLdcw110214	100	155	DNA primase
MCLdcw110216	101	185	MraW
MCLdcw110217	100	108	L13
MCLdcw110218	100	79	L21
MCLdcw110219	101	265	GTP-binding protein Obg/CgtA

MCLdcw110222	100	564	Excinuclease ABC subunit B
MCLdcw110224	101	154	Ribosome-recycling factor
MCLdcw110225	100	95	S8
MCLdcw110226	101	116	S13
MCLdcw110228	97	129	Translation initiation factor IF-3
MCLdcw110230	100	85	L22
MCLdcw110233	100	128	SsrA-binding protein
MCLdcw110235	101	200	transcription elongation factor NusA
MCLdcw110239	101	112	L20
MCLdcw110242	100	194	L1
MCLdcw110243	100	161	tRNA-(guanine-N1)-methyltransferase
MCLdcw110244	99	79	S15
MCLdcw110246	100	40	L24
MCLdcw110247	100	90	L18
MCLdcw110248	101	258	preprotein translocase, SecY subunit
MCLdcw110249	98	76	L17
MCLdcw110253	93	234	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2, 6-diaminopimelate ligase
MCLdcw110255	100	84	L19
MCLdcw110257	100	105	NusG antitermination factor
MCLdcw110258	100	232	Phenylalanyl-tRNA synthetase alpha chain
MCLdcw110259	101	60	S16
MCLdcw110260	100	97	S10
MCLdcw110265	98	91	L9
MCLdcw110269	93	229	Chorismate synthase
MCLdcw110270	100	73	S17
MCLdcw110272	98	183	Methionyl-tRNA formyltransferase
MCLdcw110273	101	183	uridylate kinase
MCLdcw110277	99	296	Holliday junction ATP-dependent DNA helicase ruvB
MCLdcw110287	98	323	lysyl-tRNA synthetase (class II)
MCLdcw110294	100	130	Guanylate kinase
MCLdcw110295	95	129	Phospho-N-acetylmuramoyl-pentapeptide-transferase
MCLdcw110297	100	129	riboflavin biosynthesis protein RibF

MCLdcw110298	87	85	tRNA threonylcarbamoyladenosine biosynthesis protein RimN
MCLdcw110306	99	304	Phenylalanyl-tRNA synthetase beta chain
MCLdcw110309	92	100	N-acetylglucosamine transferase
MCLdcw110313	96	108	pantetheine-phosphate adenylyltransferase
MCLdcw110314	96	170	glycerol-3-phosphate dehydrogenase
MCLdcw110317	92	56	L25/L23
MCLdcw110318	99	494	Polyribonucleotide nucleotidyltransferase
MCLdcw110321	95	134	Recombination protein recR
MCLdcw110327	93	56	L35
MCLdcw110332	92	238	Peptide chain release factor 2
MCLdcw110342	93	73	Holliday junction ATP-dependent DNA helicase ruvA
MCLdcw110345	87	29	S6
MCLdcw110349	94	502	transcription-repair coupling factor
MCLdcw110352	98	136	oxygen-independent coproporphyrinogen III oxidase
MCLdcw110353	91	134	DNA protecting protein DprA
MCLdcw110358	96	34	Uncharacterized protein family UPF0079, ATPase
MCLdcw110365	91	67	tRNA(Ile)-lysidine synthase
MCLdcw110373	87	338	ATP-dependent DNA helicase RecG
MCLdcw110380	90	126	pyrroline-5-carboxylate reductase
MCLdcw110383	92	213	DNA repair protein RecN
MCLdcw110388	92	52	Dephospho-CoA kinase
MCLdcw110394	90	116	6,7-dimethyl-8-ribityllumazine synthase
MCLdcw110405	97	370	Glutamyl-tRNA(Gln) amidotransferase subunit A
MCLdcw110408	93	48	iojap-like protein
MCLdcw110409	93	254	primosomal protein NÔÇÖ
MCLdcw110416	98	321	tRNA(Asn/Gln) amidotransferase subunit B
MCLdcw110420	98	73	L10
MCLdcw110425	93	66	nicotinate-nucleotide adenylyltransferase
MCLdcw110435	88	297	Argininosuccinate synthase
MCLdcw110444	90	136	Cytidylate kinase
MCLdcw110449	89	30	trigger factor
MCLdcw110457	91	105	Riboflavin synthase alpha chain

MCLdcw110466	89	222	S-adenosylmethionine: tRNA ribosyltransferase-isomerase
MCLdcw110494	88	79	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase
MCLdcw110495	88	221	Imidazole glycerol phosphate synthase subunit hisF
MCLdcw110507	88	292	Porphobilinogen synthase
MCLdcw110513	89	338	chromosome segregation protein SMC
MCLdcw110524	87	96	Septum formation protein Maf
MCLdcw110525	90	72	crossover junction endodeoxyribonuclease RuvC
MCLdcw110556	93	224	1-deoxy-D-xylulose 5-phosphate reductoisomerase
MCLdcw110559	90	119	2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
MCLdcw110595	90	62	UPF0133 protein ybaB
MCLdcw110608	89	347	glutamate-1-semialdehyde-2,1-aminomutase
MCLdcw110617	92	194	fatty acid/phospholipid synthesis protein PlsX

Table S5: List of the 117 genes used for the phylogenomic tree of Figure 1. The genes are listed in their order of concatenation in the supermatrix. # species corresponds to the number of genomes (in the 101-species version; see Figure S2) for which a given gene was present in the orthologous group (OG), and thus included here. # AA is the number of unambiguously aligned amino-acid positions used for each gene.

6 Discussion & conclusion

6.1 Forewords

As explained in the objectives, the aim of this thesis was to produce scenarios for the evolution of the bacterial cell wall. My work was organized into two main parts, the genome dereplication (tool) and the production of the evolutionary scenarios. In this section, I will further discuss several points which have marked our interest during this thesis. As above, I will discuss separately the two main parts.

6.2 ToRQuEMaDA

ToRQuEMaDA (TQMD)[1] is a tool made for everyone working in the field of prokaryotic phylogenetics. Its capacity to **automatically dereplicate large datasets of genomes** while keeping representatives in every phylum allows users to reduce the time allocated to the selection of genomes. Without the help of automated tools, it is extremely time-consuming to dereplicate a large dataset correctly, for example in a **phylogenomic context**

In our article about scenarios for the evolution of the bacterial cell wall, we have hit a wall extremely soon. In the context of my Master's thesis, I tried to work with the 9000 prokaryotic (2013) genomes available in the Ensembl database but ended up stuck at the creation of orthologous groups of proteins due to the then already large number of genomes. Having less than a year and having no pre-existing tool for the job, I had to create a "quick and dirty" way to dereplicate genomes. This was the ancestor of TQMD, also based on ***k-mers*** but shorter ones (pentamers or hexamers). The program was relatively fast and the results were acceptable, but it was not an automated process. Instead, it required a lot of human input (which can lead to errors). I realized I needed not only to make the process **fully automatic** but also improve it (clustering quality, speed) and make it easily scalable to the ever-growing number of genomes.

In 2013, we had problems dealing with 9000 prokaryotic genomes, in 2017 we had to deal with around 80,000 genomes and in **January 2021** we had to deal with **211,001 genomes from NCBI RefSeq Prokaryotes**[2]. Public databases are also highly redundant: as of release 203 of GenBank[3], the totality of **GenBank** (Eukaryotic and Prokaryotic genomes) amounts to **939,798 genomes**. Amongst them, there are **624,750 Proteobacteria**, of which **105,081 *Escherichia coli*** genomes alone! In contrast, Firmicutes represent 149,410 genomes, of which 1245 *Bacillus subtilis* genomes. The problem is ever-increasing but, except for TQMD and, as of 2017, dRep[4], there is a lack of programs publicly available and published to do the work on a large scale (as of January 2021, Assembly-Dereplicator is not yet published).

6.2.1 Alternatives to ToRQuEMaDA

dRep can dereplicate and select representatives but is not optimized for aggressive dereplication like TQMD. At first, TQMD was not optimized for dereplication at the species level like dRep because I used only JELLYFISH[5] as the *k*-mer engine. However, the addition of the **capacity to switch the *k*-mer engine between JELLYFISH and Mash**[6] allowed TQMD to remain competitive when dereplicating on the species level.

There is an alternative to the use of dereplication tools: manually selecting a genome per taxonomic group/level of interest (e.g., in the **Genome Taxonomy Database** (GTDB)[7,8]). In the case of GDTB, a given collection of genomes is reduced to a single genome assembly, a “type strain”, based on an ANI threshold of 95%[8]. It could be argued that dereplication tools like TQMD are thus not really useful since such an easy alternative exists. Yet, relying ready to use genome selection also has its disadvantages.

First **disadvantage**, it promotes the **uniformization of research**, which could be a great danger in Science. Indeed, if genome mis-assemblies or contaminations are not detected by the automatic filters, all the publications relying on GDTB (or other publicly available genome selection) could potentially be affected. Our opinion is that it is in the interest of Science to propose alternatives to researchers, even if only for **corroboration**. Hence, in spite of the existence of GTDB, different tools are still currently used by genomic researchers, such as dRep[4] or pyani[9].

Second **disadvantage**, **pre-made genome selections are fixed**. Indeed, tools like TQMD are able to dereplicate while specifying a target number of genomes for a specific clade. It is indeed well accepted that undereplicated datasets cause problems in downstream analyses[10], notably for read mapping. Nevertheless, Evans and Denef 2020[10] have also shown that hard dereplication can be the cause of gene losses in genomic populations. TQMD is a useful alternative that allows researchers to select their preferred dereplication threshold and to specify a target number of representative organisms. This approach can for instance be useful in the context of **metapangenomics**. Moreover, TQMD supports “priority lists" that allow the user to ensure that specific genomes are chosen as representatives for their clusters, which can be extremely convenient in **comparative genomics** applications where some model genomes must appear in the dataset in spite of dereplication.

Last **disadvantage**, **users have to wait for an update** for the newest genomes to be implemented, while **private genomes** are obviously not available either. The part of the unknown is a well-known phenomenon in metagenomics, with around 20% of undescribed microbial sequences in a microbiome[11]. Therefore, enabling the use of private genomes (in-house) during dereplication is important, for example in the course of a **metagenomic** study of a novel environment leading to the identification of **rare genomes**, which would be interesting to include in a dereplication process before publication. Rare genomes recently uploaded on the NCBI servers, but not yet included in GTDB, could reveal essential to a given study. **TQMD supports the use of unpublished genomes during the dereplication, along with publicly available NCBI genomes**. To the best of our knowledge, such a degree of automatization and support for both public and private genomes is not currently available in other dereplication tools. These features will certainly be useful for future metagenomic projects.

While this may be a matter of taste, in the specific case of GDTB (but also tools like dRep and pyani), the pipeline only relies on ANI for clustering, whereas **TQMD can use multiple distance metrics (up to 30)** to find the best representative to work with for each cluster and allows a complete customisation of the metrics used.

6.2.2 Future of ToRQuEMaDA

TQMD can always be improved and we already have ideas. For now, we only worked with prokaryotic genomes but in theory we should also be able to work with (small-sized) **eukaryotic genomes** (e.g., fungi). However, first we have to answer a few questions. Can we work with the complete genomes and their repetitive parts or should we remove the repetitive parts? Do we keep introns? These are the types of tests we would need to conduct in order to verify the possibility of using TQMD with eukaryotic genomes (or to seek a way of improving TQMD for such cases).

For now, we use the Identical Genome Fraction, the Jaccard Index and an estimate of the Jaccard Index combined with a greedy clustering algorithm for clustering and selecting a representative but **other clustering (or distances) could be tried** (e.g., the K-Means, the Mean-Shift clustering or the agglomerative hierarchical clustering). These are only the "basic" and well-known ones. Implementing them would require extensive testing and may also require important modifications to TQMD structure. However, it could be worth the time and effort if it could alleviate some of the limitations of TQMD. For example, we use single linkage (stops at the first "good enough" comparison) to reduce the computing time instead of a "all-against-all" comparison, but it sometimes causes genomes to create a bridge between two clusters that should have remained separated. As proposed by one of TQMD's reviewers, we created a stricter option which limits the single-linkage comparison to the best genome of each cluster. Yet, it does not entirely prevent the issue since a problematic genome could still be the best genome of a cluster.

Another place with room for improvements in TQMD is our default selection of criteria for the selection of representatives. What new criteria should we use for the selection of the representatives or add to our list of metrics? What current criteria should we drop because they are not useful or they are redundant with others? We focused on **fast and simple to produce** (to reduce the computing time)**, and easy to understand criteria** (so that the user will not use a "blackbox"). These prerequisites for the criteria should be maintained (or at least try to). A possibility would be to identify the best characterised genomes by attributing a score based on the completeness of the description in knowledge databases (peptidoglycan thickness, number of membranes, etc).

6.2.3 ToRQuEMaDA and genome contamination

Contaminations (and chimerical genomes) influence TQMD in its clustering phase by allowing genomes which should not be grouped together at a given threshold with a specific algorithm to cluster. A single wrongly clustered genome can create a snowball effect during a single round but stops at that specific round of TQMD unless the contaminated/chimerical genome is used as the representative. If a problematic genome enables several snowball effects, we call it a **"black-hole" genome**. Caution is thus advised while using TQMD and the curation of a list of known contaminated genomes to exclude is a must-have. Tools exist for this purpose (or retooled for this purpose) and some are used and shown in Cornet et al. (2018)[12]. We also **included support for two tools which can estimate the contamination level of a genome**, **FortyTwo**[13,14] and **CheckM**[15]. FortyTwo is based on the comparison of the genome ribosomal proteins to the reference sequences of the RiboDB database[16] and CheckM is based on lineage-specific marker genes in addition to ribosomal proteins[15]. We also use RNAmmer and CD-HIT-EST[17,18], which

respectively retrieve the predicted SSU (16S) rRNA and cluster them, to identify potential contamination in input genomes. If a genome possesses at least one SSU rRNA which does not cluster with the other SSU rRNA predicted for this genome, then it is considered to be contaminated.

TQMD could also be used to **detect heavy contamination, chimeric genomes or taxonomic mislabeling**. First TQMD will need a (relatively) small list of high taxonomic level representatives bereft of contaminations as a basis. Using this list as a "database", TQMD will be launched with these curated genomes and the genomes of interest. If TQMD's results show the genomes of interest to be clustered in an unexpected way, a further and more in-depth check of these genomes will allow verifying if they are indeed contaminated, correspond to chimeras and/or are taxonomically mislabeled. The level of precision of TQMD would allow to detect genomes like *Bacillus subtilis BEST 7613*[19] (which includes a full cyanobacterial genome) but will not be able to be as efficient as **Physeter**, another tool of our lab, the purpose of which is to **specifically detect contaminated regions in whole genomes** (available at: https://metacpan.org/dist/Bio-MUST-Apps-Physeter).

6.3 Cell-wall architecture

At first, we thought the diderm-LPS, or at least the diderm, architecture to be the cell-wall architecture of the **LBCA**. We made this educated guess because diderm bacteria are present at several places and in an overwhelming proportion. Our actual results were thus counter-intuitive because it is more parsimonious to create only once an outer membrane due to the difficulty of such creation than the multiple creation of a second membrane. This view, "diderm-first", is also shared by Cavalier-Smith (2020)[20], Taib et al. (2020)[21] or Coleman et al. (2021)[22]. However, as seen in our paper, the probabilistic (model-based) reconstruction of the ancestral state showed that the LBCA was **more likely monoderm** (Figure 1) and thus the cell-wall evolution of the bacteria is **not following a parsimonious path**.

6.3.1 Conundrum with the root

We considered what we call the **true diderms-LPS** (TDL) to arise from a single ancestor and thus rejected the rootings within this group. Indeed, once we give sufficient phylogenetic information, the corresponding phyla tends to regroup like in our results or in studies published by other teams (see Table 1 and Figure 9 in Introduction)[7,20,23–31]. Additional clues suggest a **monophyly of the TDL** like Gupta's study of a 20-23 AA insertion in the Hsp70 protein specific to the TDL[32,33].

In order to strengthen the conclusions stemming from our reconstruction, we **checked genes that should be unique to the diderm architecture**. As described in our manuscript, we ended with four different patterns that confirmed our intuition to not root within the TDL and separate the diderm architecture in, at least, two different groups (AD and TDL) with potentially a shared origin, which still needs to be confirmed.

Rooting at the base of the TDL or rooting at the base of the **Terrabacteria** group is thus the same but since Terrabacteria is an "official" denomination in the taxonomy and the true diderms-LPS

(TDL) is just a “non-official” way we use to designate a series of organisms, we systematically call this specific rooting, the Terrabacteria rooting. For the TDL, as mentioned above, there are sufficient clues to consider the monophyly of the group but for Terrabacteria there is still a lack of evidence for the group monophyly. Indeed, in their case we do not have elegant clues like the Hsp70 insertion. This is the reason why we did not consider rooting within the TDL group but did it within Terrabacteria.

Ideally, we should have used Archaea as an outgroup for the rooting but using Archaea with Bacteria creates artifacts that change the position of basal groups of Bacteria, hence diminishing their interest[34]. The effect is particularly important when including the Thermotogae, which are supposed to have inherited genes adapted to high temperature from Archaea[35]. The other genomes (Aquificae) with adaptations to high temperature were then attracted by the Thermotogae. This is the reason we removed both Thermotogae and Aquificae from our trees. The larger the distance between the outgroup and the organisms of interest, the more sensitive the models become[34] to the **long branch attraction** (LBA) effect and in the case of the Bacteria and the Archaea, the distance is almost as large as it could be. A way to try to offset these problems would be to soften the gap by adding the **Candidate Phyla Radiation** (CPR)[36]. In most studies (see Table 1 and Figure 9 from the Introduction), the CPR is located at the base of the tree and would close, at least slightly, the gap and reduce the LBA effect caused by the Archaea. However, in the study of Coleman et al. 2021[22], the CPR is located within the Terrabacteria group, so it might be inefficient to rely on the CPR to break the long branch leading to Archaea and reduce the artefacts.

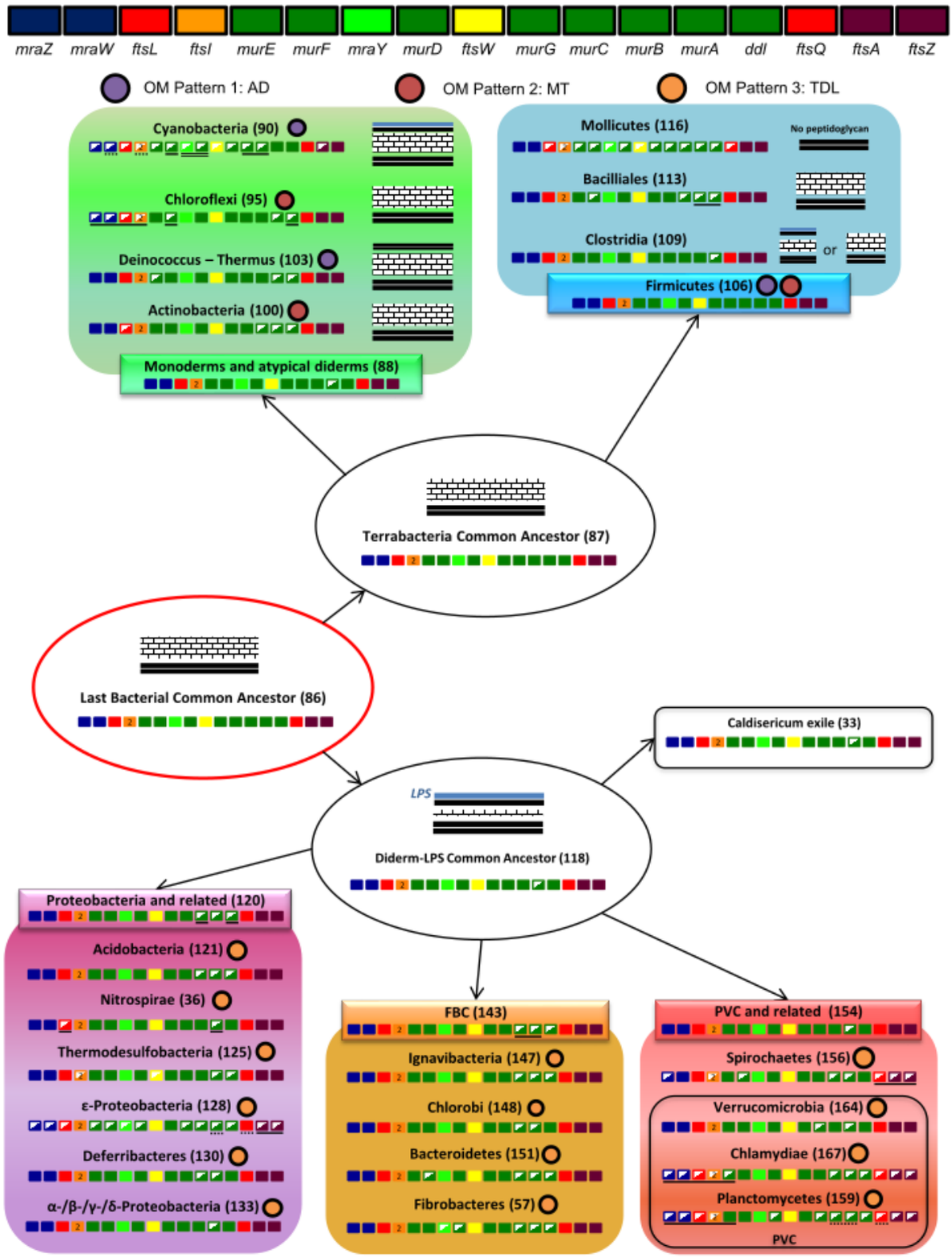


Figure 1: Schematic representation of the evolution of the bacterial *dcw* gene cluster based on the ancestral cluster *dcw* reconstruction, the cell-wall reconstruction and the study of the OM presence. Numbers in parentheses correspond to the node number in our phylogenomic tree (tree annotated with node numbers available at https://figshare.com/s/fd6a7e5cd11070e63b3d, folder ProCARs).

Half-colored boxes represent genes present in the genome but outside the cluster, lines below boxes denote genes united in a cluster different from the main one. Colored circles correspond to the Outer-membrane (OM) pattern of Figure 3 (B) from “Was the last common bacterial ancestor a monoderm?”. LPS = Lipopolysaccharide, AD = Atypical Diderms, MT = Monoderm Terrabacteria, TDL = True Diderms-LPS.

6.3.2 Limitations of our approach

Between the moment we finished the computations for our trees and now, new genomes have become available, and their number has been continuously expanding. For our purpose, developing possible scenarios for bacterial evolution, adding new groups of genomes to the trees and reconstructing the ancestral traits of the cell wall would have been interesting. Among these genomes, many of which are MAGs, including those of the interesting CPR organisms.

With the prototype of TQMD, dereplicating them would have been difficult and time consuming enough to justify not using them. In contrast, with TQMD, the dereplication step is not time consuming anymore. There is also a better alternative to OrthoMCL[37] for finding orthologs while our protocol to find the “best” genes to construct the AA supermatrix is still valid. **The phylogenetic inference would still have required a lot of computational power and would have taken probably between three and six months for a chain using 96 CPUs per chain**. An enormous problem still persists: these new genomes are “only” genomes. The corresponding organisms are not cultivated nor described.

Why is that a problem for phylogenomics? Simply because phylogenomics is only the “skeleton” of the scenario, the “flesh” being the reconstruction of the ancestral cell wall. Our objective is not to produce a new phylogeny of the Bacteria but to propose possible scenarios for the bacterial evolution based on cell-wall reconstruction. To reconstruct the ancestral cell wall, we need information on the cell wall of the currently existing bacteria. We have this information **only for the genomes of the organisms that can be cultivated in a laboratory**.

By definition, every metagenome is a genome that we cannot (yet) cultivate on a Petri dish and the CPR are also mostly uncultivable (only one genome has its cell wall described[38]) and thus the genomes belonging to these two groups/categories are not described. The programs for the reconstruction of ancestral traits depends on the quality of the tree (and its root) and also the **quality of the information about the current traits**. Using a tree with these new genomes in the absence of data pertaining to their cell wall would only have sent us askew and the results would have been unusable (for our objectives).

Even with the genomes which are described, the reconstruction was difficult due to the **uncertainty of the descriptions**. Too many times we could only find the information at the phylum level and not at the species level. So many genomes are sequenced nowadays but so few are described that, for our method, the “old” genomes are the only ones usable. Consequently, our results may only concern the LBCA of the cultured bacteria instead of the LBCA of all currently sequenced bacteria, and our selection could be called obsolete, but they are the only organisms where we can have all the needed information. We dream that our work would spark the interest for a **standardized way of describing bacterial cell wall** instead of just “Gram negative” without any certitude as of the presence of an OM or LPS, and maybe slow down the sequencing frenzy in order to redirect a part of the effort to the description of the bacteria. If you

look at Table S3 from “Was the last common bacterial ancestor a monoderm?”, you realise that we had to simplify our input due to the lack of confidence in our information about the cell-wall architecture. Thus, having a standardized way to describe is as important as having more descriptions available.

Another consequence of our need for correctly described genomes is the amount of work required to compile these informations. To be convinced,have a look at our Table S4 from “Was the last common bacterial ancestor a monoderm?”, you will see that we needed 94 references just for 85 genomes. Thus, the more organisms are represented, the more time consuming this step will become. Ideally, the number of organisms represented should be more important, in order to account for the possibility that using the MultiStates models from BayesTraits[39–41] with only 85 organisms could cause the models to be unable to correctly estimate the rates, due to the limited number of transitions between the different states. Indeed, this is a possible technical explanation for our results, a monoderm LBCA, compared to the other model-based study[22], which concludes that the LBCA is a diderm. Nevertheless Coleman et al. reached this conclusion by using an indirect prediction, since it is logically inferred from the results of a gene reconciliation model (predictions of genes present in the LBCA) instead of the direct result of the model of trait evolution. Moreover, the genes used for the prediction of the cell-wall architecture in Coleman et al. 2021 are genes involved with the LPS precursors synthesis and the flagellar subunits of the type IV pili, which are not exclusive to the diderm bacteria[42]. For the type IV pili, its presence even in monoderm genomes implies that its inference in the LBCA is quite uninformative when trying to infer if the LBCA had an OM or not. In the case of the LPS genes, our results show that some are also found in **Atypical diderms** (AD) or even in monoderms. Thus, solely **relying on the presence of genes to predict the cell-wall architecture might not be entirely reliable**, and it might explain the different conclusions reached by us on one side and Coleman et al. on the other side.

6.4 Plans evolve

Initially, we planned to create TQMD and then use it for a selection of prokaryotes for devising new scenarios about the evolution of the bacterial cell wall from a “clean slate”. We were a bit too optimistic and soon realized that our plan did not fit into one thesis but in two. The second part of the thesis, the devising of scenarios for the evolution of the cell wall, is being revised as its own thesis on a similar subject: the study of the **cell-wall biosynthesis** of the **Archaea with a pseudo-murein cell wall** instead of the bacterial cell wall. My colleague Valérian Lupo is currently using a similar approach as described in “Was the last bacterial common ancestor a monoderm after all?”.

Five Archaea with pseudo-murein and five Archaea without pseudo-murein were chosen as the base of his study, then OrthoFinder[43,44] was used to create orthologous groups (OGs) of proteins. The OGs were filtered to find those groups exclusive to pseudo-murein Archaea or groups with clues of a specific paralogue to pseudo-murein Archaea. GeneSpy[45], a tool similar to my tool for visualizing the synteny (but published), was then used on the OGs of interest to identify conserved and/or syntenic regions. Several rounds of enrichment (using FortyTwo) followed by trimming of sequences which are too short or too long (using HMMER and ompa-pa) were then used to complete the OGs with new sequences (Archaea and Bacteria). In total, 49 OGs of interest were identified.

The number of remaining OGs was then reduced by the use of several criteria (i.e., synteny or function). Five OGs were given priority over the 49: four homologous to murC, murD, murE and murF and one homologous to ddlB. The mur OGs were regrouped into a single OG then phylogenetic analyses were performed on the two OGs. The remaining 44 OGs will not be phylogenetically studied but my colleague will try to place them on a biosynthesis scheme for the pseudo-murein. The hope is then to get a better picture of both bacterial and archaeal cell-wall evolution at the end of both our PhD works.

6.5 References

1. Léonard, R. R. *et al.* ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies. *PeerJ* **9**, e11348 (2021).

2. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

3. Sayers, E. W. *et al.* GenBank. *Nucleic Acids Res.* **48**, D84–D86 (2020).

4. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

5. Marcais, G. & Kingsford, C. Jellyfish: A fast k-mer counter. *Tutorialis E Manuais* 1–8 (2012).

6. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

7. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

8. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).

9. Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G. & Toth, I. K. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal. Methods* **8**, 12–24 (2016).

10. Evans, J. T. & Denef, V. J. To Dereplicate or Not To Dereplicate? **5**, 7 (2020).

11. Thomas, A. M. & Segata, N. Multiple levels of the unknown in microbiome research. *BMC*

Biol. **17**, 48 (2019).

12. Cornet, L. *et al.* Consensus assessment of the contamination level of publicly available cyanobacterial genomes. 1–26 (2018).

13. Irisarri, I. *et al.* Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat. Ecol. Evol.* **1**, 1370–1378 (2017).

14. Simion, P. *et al.* A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr. Biol.* **27**, 958–967 (2017).

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

16. Jauffrit, F. *et al.* RiboDB database: a comprehensive resource for prokaryotic systematics. *Mol. Biol. Evol.* **33**, 2170–2172 (2016).

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

18. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

19. Watanabe, S., Shiwa, Y., Itaya, M. & Yoshikawa, H. Complete Sequence of the First Chimera Genome Constructed by Cloning the Whole Genome of Synechocystis Strain PCC6803 into the Bacillus subtilis 168 Genome. *J. Bacteriol.* **194**, 7007 (2012).

20. Cavalier-Smith, T., Ema, E. & Chao, Y. Multidomain ribosomal protein trees and the planctobacterial origin of neomura (eukaryotes, archaebacteria). *Protoplasma* 1–133 (2020).

21. Taib, N. *et al.* Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat. Ecol. Evol.* **4**, 1661–1672 (2020).

22. Coleman, G. A. *et al.* A rooted phylogeny resolves early bacterial evolution. *Science* **372**, (2021).

23. Battistuzzi, F. U. & Hedges, S. B. A major clade of prokaryotes with ancient adaptations to life on land. *Mol. Biol. Evol.* **26**, 335–343 (2009).

24. Yutin, N., Puigbo, P., Koonin, E. V. & Wolf, Y. I. Phylogenomics of Prokaryotic Ribosomal

Proteins. *Curr. Sci.* **101**, 1435–1439 (2012).

25. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

26. Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci.* 201420858 (2015) doi:10.1073/pnas.1420858112.

27. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).

28. Castelle, C. J. & Banfield, J. F. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* **172**, 1181–1197 (2018).

29. Boussau, B., Guéguen, L. & Gouy, M. Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. **18**, 1–18 (2008).

30. Lasek-nesselquist, E. & Gogarten, J. P. The effects of model choice and mitigating bias on the ribosomal tree of life. *Mol. Phylogenet. Evol.* **69**, 17–38 (2013).

31. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* (2019) doi:10.1038/s41467-019-13443-4.

32. Lake, J. A., Herbold, C. W., Rivera, M. C., Servin, J. A. & Skophammer, R. G. Rooting the Tree of Life Using Nonubiquitous Genes. **24**, 130–136 (2007).

33. Gupta, R. S. 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 Int. J. Gen. Mol. Microbiol.* **100**, 171–182 (2011).

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

35. Zhaxybayeva, O. *et al.* On the chimeric nature , thermophilic origin , and phylogenetic placement of the Thermotogales. **106**, 5865–5870 (2009).

36. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

37. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of ortholog groups for

eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).

38. Soro, V. *et al.* Axenic culture of a candidate division TM7 bacterium from the human oral cavity and biofilm interactions with other oral bacteria. *Appl. Environ. Microbiol.* **80**, 6480–6489 (2014).

39. Pagel, M., Meade, A. & Barker, D. Bayesian Estimation of Ancestral Character States on Phylogenies. *Syst. Biol.* **53**, 673–684 (2004).

40. Pagel, M. & Meade, A. Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. *Am. Nat.* (2015) doi:10.1086/503444.

41. Meade, A. & Pagel, M. *BayesTraits V3. 0.1*. (2017).

42. Melville, S. & Craig, L. Type IV pili in Gram-positive bacteria. *Microbiol. Mol. Biol. Rev.* **77**, 323–341 (2013).

43. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).

44. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 1–14 (2015).

45. Garcia, P. S., Jauffrit, F., Grangeasse, C. & Brochier-Armanet, C. GeneSpy, a user-friendly and flexible genomic context visualizer. *Bioinformatics* **35**, 329–331 (2019).

7 Appendices

7.1 Acknowledgements

Le moment des remerciements est arrivé. Grands dieux que c'est compliqué de savoir quoi mettre ! J'ai passé 30 minutes en face d'une page blanche avant d'écrire ces trois phrases.

Je pense que je devrais d'abord remercier trois personnes: Fred, Denis et Eric. Pourquoi ? Parce qu'ils sont mes promoteurs. Ce sont eux qui m'ont soutenu pour le passage d'étudiant à chercheur (si je réussis ma thèse, ne vendons pas la peau de l'ours avant de l'avoir tué). Ou en d'autres termes, ce sont eux qui m'ont aidé à passer d'une petite main qui ne fait que répondre à des demandes et réfléchir au côté technique à un esprit plus critique qui voit plus loin que le bout de son nez (et là l'ours est déjà tué donc je peux vendre sa peau). Donc, merci.

Qui d'autres ? Mmmmmh. Les amitiés qui ont été forgées au labo me semblent un bon point non ? Loïc, Catherine, Amandine et Valérian. Désolé pour les autres qui liraient ces lignes et seraient déçus de ne pas être cités mais vous êtes "juste" des potes. Le travail ne fait pas tout, aimer ce que l'on fait ne suffit pas à rester motivé et à éviter les jours sans. C'est là que les collègues interviennent, car lors des jours sans, ce sont eux qui peuvent rendre le travail supportable, qui peuvent donner envie de se lever. Et vu que je suis devenu pote (ou mieux) avec tout le monde, j'adore toujours venir au labo, même les jours sans. Alors à mes amis et à mes potes, merci.

Mes parents et mon chat aussi sont à remercier, ce sont eux qui me soutenaient les jours sans une fois rentrés à la maison. A ceux et celles qui se souviendraient des remerciements de mon mémoire, oui j'ai récidivé et ai, à nouveau, remercié mon chat. Oui, c'est toujours le même, Percy. Papa, maman, Percy, merci.

J'ai mis du temps à défendre cette thèse mais bon je ne regrette rien, ça m'a permis de passer plus de temps avec vous.

A une prochaine fois,

Raphaël Léonard