A journey into prokaryotic cell-wall evolution: the two half-brothers Murein and Pseudomurein

Valérian LUPO

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in Science

Academic year

2022-2023

Jury:

Prof. Jean-Denis DOCQUIER (President)

Dr. Johan MICHAUX (Secretary)

Prof. Denis BAURAIN (Promoter)

Dr. Frédéric KERFF (Copromoter)

Dr. Leonor PALMEIRA

Dr. Bogdan I. IORGA

Prof. Xavier DE BOLLE

A la mémoire de mon zio Mommo.
Pour toutes ces heures qu'on a passé
à parler de sciences, je te dédie cette thèse.

J'aimerais dédier ces quelques lignes aux personnes qui m'ont aidé et soutenu durant ce parcours pour leur adresser mes remerciements.

Hormis le FNRS qui a financé ces quatre années de doctorat, je voudrais d'abord remercier mes deux promoteurs, Denis et Fred, sans qui je ne serai jamais arrivé au bout de cette thèse. Malgré un emploi du temps très chargé (surtout Denis), ils ont toujours pu m'accorder du temps pour mener à bien tous mes projets.

Ensuite, j'aimerais remercier tous mes (anciens) collègues et amis du laboratoire de Phylogénomique des Eucaryotes : Amandine, Marie, Catherine, Rosa, Raph, Loïc, Mick, Luc, Richard et Damien. Merci de m'avoir aidé dans ma formation de bioinformaticien. Malgré le peu de moments passés ensemble (merci l'incendie et la pandémie), je me suis quand même bien amusé. Et comme dirait Raph "We are pack". Oh… j'allais oublier mes étudiants ! Célyne, Edmée, Charlotte, Loïc et Olivier, merci pour votre travail et votre sympathie.

Merci à tous mes amis. Je ne vous cite pas car vous êtes nombreux, mais sachez que vous m'avez permis de me vider l'esprit à chaque fois que mon travail devenait trop prenant.

Bien entendu, je remercie ma famille. Il est toujours plus facile de terminer un travail de cette ampleur lorsqu'on est entouré de personnes qui nous soutiennent. Merci à ma belle-soeur, Mélanie. Si tu ne m'avais pas accompagné pour défendre mon projet à Bruxelles, je ne serais peut-être pas en train d'écrire ces mots.

Et enfin, un remerciement tout particulier à toi, Roxane, qui partage ma vie depuis toutes ces années. Malgré mon entêtement et ma mauvaise humeur (parfois), tu m'as toujours soutenu et encouragé, peu importe mes projets et mes décisions. Je n'ai pas assez de mots pour t'exprimer ma gratitude, donc je te dis juste MERCI !

Pour toutes les personnes ayant des difficultés avec l'anglais, profitez de ces dernières phrases rédigées en français car le reste du manuscrit sera très hermétique pour vous… For the others, I hope you will enjoy reading this thesis.

A bientôt,
See you soon,

Valérian Lupo

I. Résumé

Les procaryotes, organismes unicellulaires ne présentant pas de noyau, sont actuellement divisés en deux domaines : Bactéries et Archées. L’une des différences majeures entre les deux domaines réside dans leur paroi cellulaire. En effet, bien que les bactéries ont majoritairement du peptidoglycane (aussi appelé muréine) dans leur paroi, la plupart des archées ont une paroi composée d’une couche protéique assemblée en un réseau cristallin, que l’on nomme couche S. Cependant, il existe deux ordres d’Euryarchaeota, les Methanopyrales et Methanobacteriales, qui possèdent dans leur paroi un polymère structurellement analogue au peptidoglycane. Par conséquent, ce polymère a été nommé pseudomuréine.

L’objectif de cette thèse était d’étudier l’évolution de différentes familles de gènes impliquées dans la synthèse du peptidoglycane et de la pseudomuréine, afin de déterminer si les deux polymères partagent des déterminants génétiques communs.

Pour conduire ces analyses, nous avons exploité plus de 80 000 génomes bactériens et plus de 800 génomes archéens provenant tous de la base de données RefSeq du NCBI. Or, au début de notre travail, un faisceau d’indices laissait penser que RefSeq, en dépit de sa curation extensive, présente des problèmes de contamination des génomes pouvant fausser l’interprétation des résultats phylogénétiques. Dans un premier temps, nous avons donc développé un programme de détection des contaminations baptisé *Physeter*. Celui-ci a ensuite été utilisé pour détecter les potentielles contaminations génomiques présentes dans les génomes procaryotes. Par cette étude, nous avons montré qu’environ 0.9% des génomes bactériens de RefSeq ont un taux de contamination d’au moins 5%.

Par ailleurs, si RefSeq offre une bonne couverture de la diversité procaryotique, elle souffre de biais d’échantillonnage. Dans le but de concevoir et tester des stratégies bioinformatiques pour améliorer l’informativité des phylogénies en réduisant les redondances dues à l’inclusion de nombreuses souches très apparentées, nous avons choisi de prototyper nos méthodes sur la famille des bêta-lactamases de classe D. Ces dernières sont des enzymes produites par les bactéries pour lutter contre les antibiotiques à noyau bêta-lactame, une famille d'antibiotiques qui ciblent

la synthèse du peptidoglycane et provoquent la lyse de la cellule. Nous avons conduit une étude phylogénétique et bioinformatique complète de cette famille. A la suite de ces résultats, nous avons exprimé dans *Escherichia coli* dix séquences de protéines nouvellement identifiées et montré que les bactéries environnementales (même non-exposées aux antibiotiques d'origine anthropique) constituent un grand réservoir de gènes de résistance contre les agents antimicrobiens.

Enfin, fort d'une version décontaminée de RefSeq et des méthodes bioinformatiques permettant d'en optimiser l'exploitation, nous avons identifié différentes familles de gènes potentiellement impliquées dans la synthèse de la pseudomuréine archéenne. Certains des gènes identifiés sont homologues à ceux impliqués dans la synthèse du peptidoglycane, comme des Mur ligases ou la protéine transmembranaire MraY. Nous avons montré que ces gènes sont regroupés dans deux régions synténiques dans les génomes de Methanopyrales et Methanobacteriales. De plus, nos analyses phylogénétiques suggèrent que les Mur ligases archéennes sont le résultat de transferts de gènes horizontaux depuis une ou plusieurs anciennes lignées bactériennes.

En combinant tous les résultats obtenus, nous avons proposé l'hypothèse à vérifier que c'est l'acquisition de gènes bactériens par un ancêtre commun des Methanopyrales et des Methanobacteriales qui a entraîné l'apparition de la pseudomuréine archéenne.

II. Abstract

Prokaryotes (i.e., single-celled organisms without a nucleus) are currently divided into two domains: Bacteria and Archaea. One of the major differences between the two domains lies in their cell wall. Indeed, although bacteria have mostly peptidoglycan (also known as murein) in their cell wall, most archaea have a cell wall composed of a protein layer assembled into a crystalline network named S-layer (Surface layer). However, there exist two orders of Euryarchaeota, the Methanopyrales and Methanobacteriales, which possess in their wall a polymer structurally analogous to peptidoglycan. Therefore, this polymer was called pseudomurein.

The objective of this thesis was to study the evolution of different gene families involved in the biosynthesis of peptidoglycan and pseudomurein, in order to determine if these two polymers share common genetic determinants.

To conduct our analyses, we exploited more than 80,000 bacterial genomes and more than 800 archaeal genomes, all collected from the NCBI RefSeq database. However, at the beginning of our work, there were indications that RefSeq, in spite of its extensive curation, presents problems of genomic contamination that could bias the interpretation of phylogenetic results. As a first step, we developed a contamination detection software called Physeter. This software was then used to detect potential genomic contamination in prokaryotic genomes from RefSeq. Through this study, we have shown that about 0.9% of the bacterial genomes in RefSeq have a contamination rate of at least 5%.

Although RefSeq provides a good coverage of prokaryotic diversity, it suffers from sampling biases. In order to design and test bioinformatics strategies to improve the informativeness of phylogenies by reducing redundancies due to the inclusion of many closely related strains, we chose to prototype our methods on the class D beta-lactamase protein family. These are enzymes produced by bacteria to resist beta-lactam antibiotics, a family of antibiotics that target peptidoglycan synthesis and lead to cell lysis. Here, we conducted a comprehensive phylogenetic and bioinformatic study of this protein family. Following these results, we expressed in

Escherichia coli ten newly identified protein sequences and thus showed that environmental bacteria (including those never exposed to human-made antibiotics) constitute a large reservoir of resistance genes against antimicrobial agents.

Finally, using a decontaminated version of RefSeq and bioinformatics methods to optimize its exploitation, we identified different gene families potentially involved in archaeal pseudomurein biosynthesis, on which we applied a bioinformatic pipeline similar to the one implemented with class D beta-lactamases. Some of the identified genes are homologous to those involved in peptidoglycan biosynthesis, such as Mur ligases or the transmembrane protein MraY. We have shown that these genes are clustered in two syntenic regions in the genomes of Methanopyrales and Methanobacteriales. Furthermore, our phylogenetic analyses suggest that the archaeal Mur ligases result from horizontal gene transfers from one or more ancient bacterial lineages.

Based on all these results, we proposed that the hypothesis that the acquisition of bacterial genes in a common ancestor of the Methanopyrales and Methanobacteriales has led to the origin of the archaeal pseudomurein.

III. List of Abbreviations

AA = amino acid

AMR = antimicrobial resistance

ARMAN = archaeal Richmond Mine acidophilic nanoorganisms

ASTRAL = accurate species tree algorithm

BLDB = betalactamase database

CIIM = class II methanogens

CIM = class I methanogens

CM = cytoplasmic membrane

CPS = carbamoyl phosphate synthetase

CTD = C-terminal domain

CTX-M = cefotaximase from Munich

dcw = cell-wall synthesis

DD-TPases = DD-transpeptidases

DNA = deoxyribonucleic acid

EMBL-EBI = European Molecular Biology Laboratory-European Bioinformatics Institute

ENA = European Nucleotide Archive

ESBLs = extended-spectrum beta-lactamases

FPGS = folylpolyglutamate synthase

GlcNAc = N-acetylglucosamine

GTase = glycosyltransferases

Gy = billion years

HGT = horizontal gene transfer

HMM = hidden Markov model

ICNP = International Code of Nomenclature of Prokaryotes

IMP = imipenemase

INSDC = International Nucleotide Sequence Database Collaboration

KPC = *K. pneumoniae* carbapenemase

LACA = last archaeal common ancestor

LBCA = last bacterial common ancestor

LCA = last (or lowest) common ancestor

LECA = last eukaryotic common ancestor

LPS = lipopolysaccharide

LTA = lipoteichoic acid

LUCA = last universal common ancestor

MAGs = metagenome-assembled genomes

MBLs = metallo-beta-lactamases

MRSA = methicillin-resistant *S. aureus*

MurNAc = N-acetylmuramic acid

NAT = N-acetyl-L-talosaminuronic acid

NCBI = National Center for Biotechnology Information

NDM = New Delhi metallo-β-lactamase

NTD = N-terminal domain

OM = outer membrane

OXA = oxacillinase

PBP = penicillin-binding protein

PG = peptidoglycan

PGAP = Prokaryotic Genome Annotation Pipeline

PM = pseudomurein

RNA = ribonucleic acid

SBLs = serine-beta-lactamases

SHV = sulfhydryl variant

SRA = sequence read archive

SSU rRNA = small subunit ribosomal ribonucleic acid

TEM = Temoniera

TPase = transpeptidases

VIM = Verona imipenemase

VRSA = vancomycin-resistant *S. aureus*

WTA = wall teichoic acids

IV. Table of contents

1. Introduction

1.1. From two to three… or maybe two… domains of life

In 1925, the French zoologist Edouard P.L. Chatton was the first scientist to introduce the concepts of *prokaryote* and *eukaryote* to classify living organisms according to their cellular organization (Chatton 1925). The two terms derive from Greek roots, which mean "before the nucleus" (pro- = 'before', -karyon = 'kernel') and "true nucleus" (eu- = 'good', -karyon = 'kernel'), respectively. Those terms were then reintroduced in 1962 by Roger Stanier and Cornelis B. van Niel in their article entitled "The Concept of a Bacterium", where they separated all living organisms in **two domains**: bacteria (prokaryotes) and the others (eukaryotes). "Prokaryote" was defined as an unicellular organism without a nucleus or organelles (e.g., mitochondria or chloroplast), which mostly divides by binary fission. In contrast, "Eukaryote" refers to uni- or multicellular organisms, where the genomic DNA is enclosed within a membrane-bound nucleus, which do possess organelles and divide by mitosis (Stanier and Van Niel 1962).

A decade later, Carl Woese and George Fox used the RNA of the small subunit of the ribosome (SSU rRNA = rRNA 16S) to study phylogenetic relationships among prokaryotes. Their analyses showed that methanogenic bacteria were clearly distinct from the other bacteria. Consequently, they proposed to classify the methanogenic bacteria as archaebacteria and the "typical" bacteria as eubacteria (Woese and Fox 1977). In the next few years, further phylogenetic analyses of the SSU rRNA tended to confirm the dichotomy between eubacteria and archaebacteria (Fox et al. 1980). In 1990, Woese, Otto Kandler and Mark Wheelis showed with molecular comparison of SSU rRNA that all living organisms are actually divided into **three domains** (Fig. 1A): Archaea (formally archaebacteria), Bacteria (formally eubacteria) and Eucarya, which will later be renamed to Eukaryota (Woese et al. 1990). Furthermore, phylogenetic reconstructions taking advantage of universal paralogous genes provided strong evidence that archaea and eukaryotes are sister groups (Cavalier-Smith 1987; Gogarten et al. 1989; Iwabe et al. 1989; Woese et al. 1990).

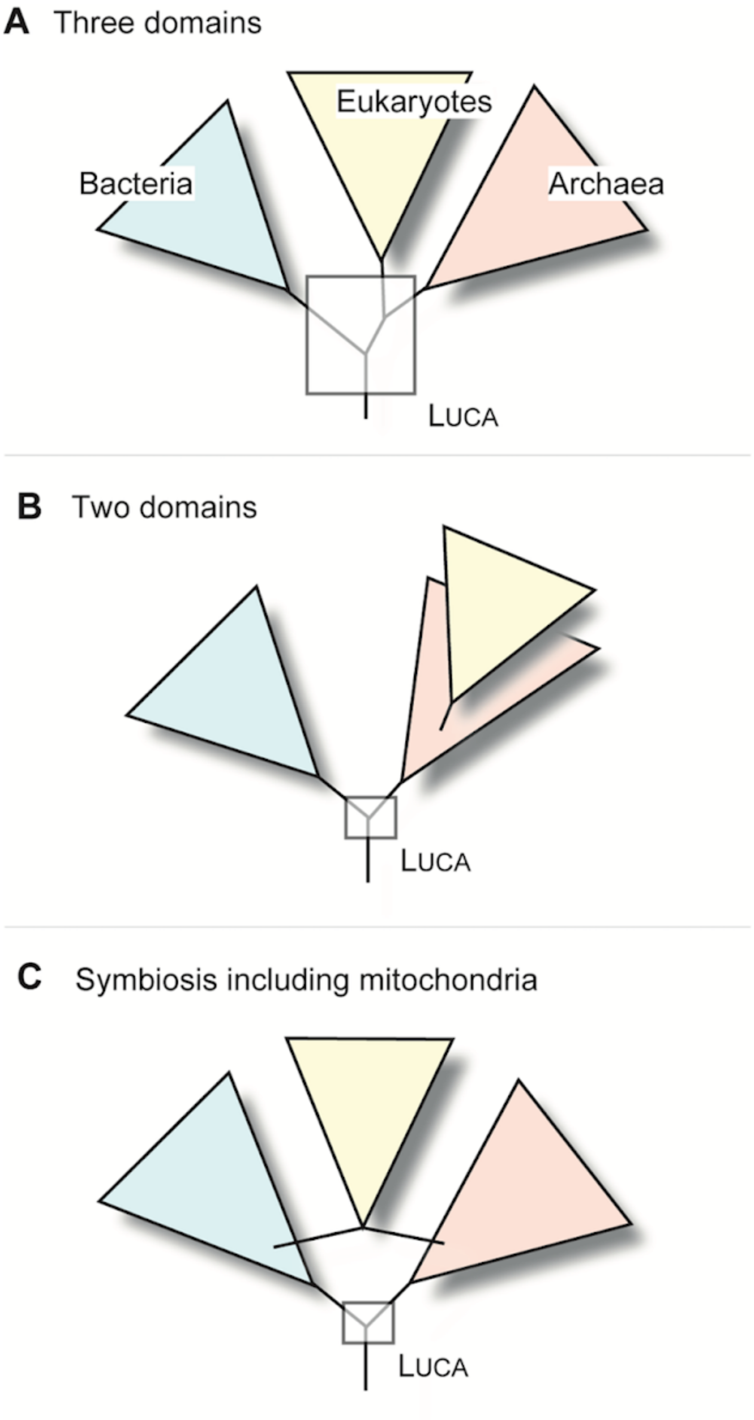


Figure 1. Schematic illustration of the different views for the relationships among all living organisms (adapted from Weiss et al. 2018). (**A**) The three-domain tree with Eukaryotes as a sister group of Archaea. (**B**) The two-domain tree where Eukaryotes emerged from Archaea. (**C**) The two-domain tree including the endosymbiosis between an Alphaproteobacteria-like bacteria and an archaeal host cell. LUCA = Last Universal Common Ancestor.

Since the 2000s, the cost of DNA sequencing has dramatically decreased, especially following the introduction of highly parallel sequencing techniques (van Dijk et al. 2014; Heather and Chain 2016) . Subsequently, this has led to an exponential growth of sequenced organisms (Sayers et al. 2022), which enables researchers to better investigate the relationships between the three major branches of the tree of life. With this huge amount of molecular data, it has been suggested that eukaryotes are not the sister group of archaea but rather branch within the archaeal domain (Guy and Ettema 2011; Kelly et al. 2011; Lasek-Nesselquist and Gogarten 2013; Williams et al. 2013; Williams and Embley 2014; Raymann et al. 2015). In 2015, archaeal organisms belonging to a new lineage (named Asgard) were isolated from marine sediments near the Loki's castle (hydrothermal vents), located between Greenland and Norway. Phylogenetic analyses revealed that eukaryotes are closely related to this newly identified archaeal lineage (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017). Following this discovery, some scientists now consider the tree of life as a **two-domain** system (Fig. 1B) instead of a **three-domain** tree, with eukaryotes emerging from archaea (Raymann et al. 2015; Eme et al. 2017; Williams et al. 2020; Nobs et al. 2022). Moreover, it is widely accepted that the origin of eukaryotes results from an endosymbiosis between an archaeal host cell and an Alphaproteobacteria-like bacteria (Fig. 1C), where the latter has evolved into present-day mitochondria (Margulis 1970; Lang et al. 1999).

In this section, I have briefly presented the “recent” classification for all living organisms (which do not include viruses) into Archaea, Bacteria and Eukaryotes. However, although phylogenetically very distant, Archaea and Bacteria do present a similar cellular organization. Archaeal and bacterial cells are even so similar that they were both considered as members of the same domain before the development of molecular phylogenetics. Despite this shared *prokaryotic* cellular organization, one of the main structural differences between Archeae and Bacteria lies in their cell wall.

1.2. Prokaryotic cell walls

In Prokaryotes, the cell wall is a structure that surrounds the cell, right above the cytoplasmic membrane (CM). It constitutes a protective layer against different types of external aggression, which can be either biotic (e.g., viruses) or abiotic (e.g., heat or acidity). It also helps the cell to preserve its shape by maintaining internal turgor pressure (Vollmer et al. 2008; Silhavy et al. 2010; Klingl et al. 2019; Pazos and Peters 2019; Meyer and Albers 2020).

1.2.1. Bacterial cell walls

1.2.1.1. Monoderm vs Diderm

In 1884, Hans Christian Gram published a staining method to observe bacteria under a microscope (Gram 1884). Following this method, bacteria were classified into **Gram positive** or **Gram negative** on whether the cell is coloured by the crystal violet stain or not. This coloration results from the properties of the bacterial cell wall. Indeed, almost all bacteria do possess peptidoglycan (PG), also called murein, in their cell wall, which is a mesh-like polymer consisting of sugars and amino acids (AAs) (Vollmer et al. 2008). Typically, **Gram positive** bacteria possess a thick layer of PG that is stained by the crystal violet, while **Gram negative** have a thin layer of PG that does not retain the stain. In addition to the PG, **Gram negative** bacteria also possess a second (outer) membrane (OM) outside the PG layer (Baurain et al. 2016; Sperandeo et al. 2019). Many bacteriologists are still using those two terms to classify bacteria because Gram staining is cheap and fast to set up. However, Gram classification does not reflect the real diversity of bacterial cell walls. Thus, we now privilege the more descriptive terms **monoderm** and **diderm** instead of **Gram positive** and **Gram negative** (Gupta 1998).

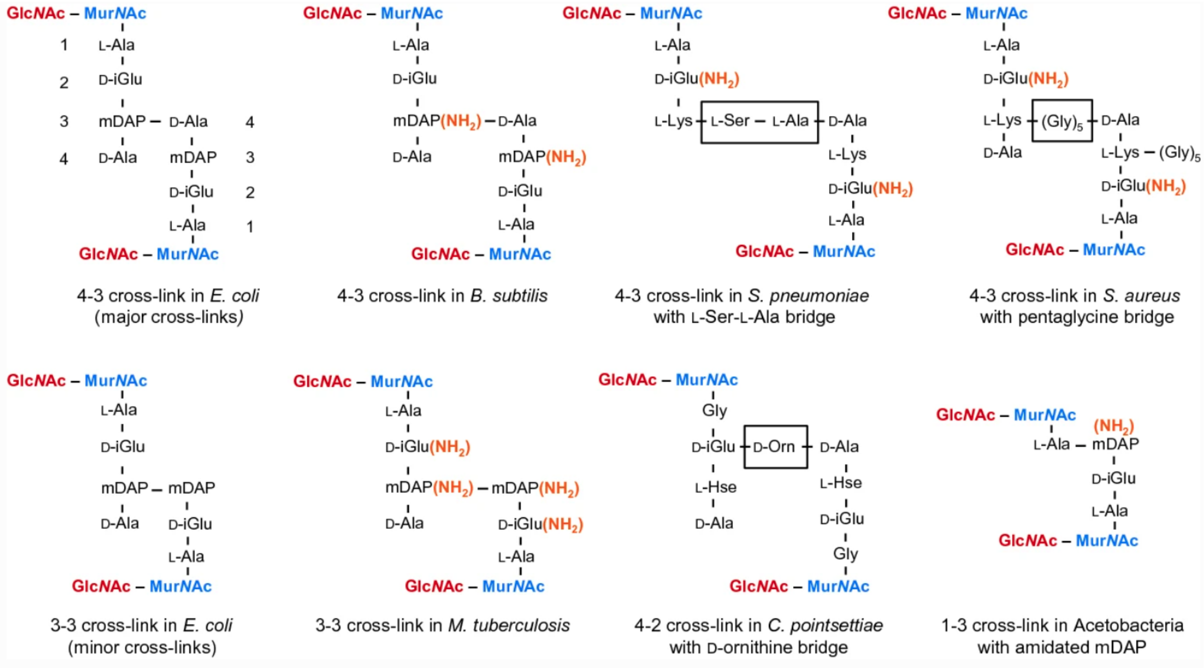
1.2.1.2. Cell-wall architectures

A cell wall is defined as monoderm or diderm depending on whether there is one or two membranes surrounding the cell. There exists a wide variation in the diderm cell-wall architecture in bacteria. The archetype of monoderm bacteria is *Bacillus subtilis*. Its cell wall is mostly composed of a PG layer about 30 nm thick (Matias and Beveridge 2005; Beeby et al. 2013), and also contains anionic polymers, which are anchored to the PG (wall teichoic acids; WTA) or to the CM (lipoteichoic acid; LTA) (Angeles and Scheffers 2021). On the other hand, the archetype of diderm bacteria is *Escherichia coli*, where the cell wall is composed of a thin layer of PG (between 3 and 6 nm) (Yao et al. 1999; Matias et al. 2003) and an asymmetric OM composed of lipopolysaccharide (LPS) in its outer leaflet (Sperandeo et al. 2019). In the *Thermotoga* lineage, the LPS OM is replaced by a loose sheath-like structure named *toga* (Rachel et al. 1988; Rachel et al. 1990; Ranjit and Noll 2016). The cell wall of Cyanobacteria, which stains Gram negative, is indeed similar to typical diderms with the exception of the PG layer, of which the thickness ranges between 10 and 35 nm, and can even reach up to 700 nm in *Oscillatoria princeps* (Hoiczyk and Hansel 2000). The most complex diderm cell wall is found in Corynebacteriales, an order of the Actinobacteria phylum. In this lineage, the PG is covalently linked to an arabinogalactan layer, which is covalently linked to a mycolic acid-based outer membrane (mycomembrane). The three latter layers are surrounded by a capsular matrix composed of exopolysaccharide and various proteins (Burkovski 2013; Rahlwes et al. 2019). In contrast, there exist some bacterial species that completely lack a cell wall, like those from the Mollicutes lineage (e.g., *Mycoplasma* sp.) (Trachtenberg 1998).

1.2.1.3. The peptidoglycan

1.2.1.3.1. Composition and structure

Despite their sometimes different architectures, almost all bacterial cell walls bear a similar component, the PG, which is composed of long glycosidic chains linked by short peptides, forming an overall net-like structure. In *E. coli*, the glycosidic chains are made out of alternating N-acetylglucosamine (GlcNAc) and N-acetylmuramic acid (MurNAc) subunits linked by a β-(1→4) bond. To the lactic acid residue in position C3 of the MurNAc is attached a stem peptide composed of five AAs: L-alanine (L-Ala), D-glutamic acid (D-Glu), meso-diaminopimelic acid (meso-DAP) and two D-alanine (D-Ala). Cross-linking of two adjacent pentapeptides frequently occurs between the carboxyl group of the D-Ala in position four of one stem peptide and the ε-amino group of the meso-DAP (in position 3) of the second one (4-3 cross-link). During cross-linking, the D-Ala in position five is released (Vollmer et al. 2008; Pazos and Peters 2019). Depending on the species, PG can show variation in stem peptide composition, cross-links (Fig. 2) or modifications of the glycosidic chain. Here, I will only describe the two first types of variation.



4-3 cross-link in *E. coli* (major cross-links)

4-3 cross-link in *B. subtilis*

4-3 cross-link in *S. pneumoniae* with L-Ser-L-Ala bridge

4-3 cross-link in *S. aureus* with pentaglycine bridge

3-3 cross-link in *E. coli* (minor cross-links)

3-3 cross-link in *M. tuberculosis*

4-2 cross-link in *C. pointsettiae* with D-ornithine bridge

1-3 cross-link in Acetobacteria with amidated mDAP

Figure 2. Examples of peptide and cross-linking variations in the peptidoglycan of different species (from Pazos and Peters 2019). Amidation of residues is shown in orange and interpeptide bridges are framed with a black square.

1.2.1.3.2. Variation in peptide composition

The variation in the stem peptide can be due to the specificity of the enzymes that synthesize the peptide or by post-synthesis modification (Vollmer et al. 2008). The most common AA in the first position is L-Ala. However, it can be replaced by a glycine (Gly) in *Mycobacterium leprae* (Mahapatra et al. 2000) or by a L-Ser in other species (Hesse et al. 2003; Vollmer et al. 2008). The second AA to be added in the stem peptide is always a D-Glu (Vollmer et al. 2008; Pazos and Peters 2019). However, it has been described in *Staphylococcus aureus, Streptococcus pneumoniae* and *Mycobacterium tuberculosis* that most of the D-Glu residues have the α-carboxyl group amidated by the complex MurT/GadT to form D-glutamine (D-Gln) (Münch et al. 2012; Morlot et al. 2018; Nöldeke et al. 2018; Maitra et al. 2021). The third position of the stem peptide shows the greatest variation. However, meso-DAP and L-lysine (L-Lys) are the two most commonly encountered AAs. The first one is found in most of diderm bacteria with a LPS OM (diderm-LPS), in some Bacilli (e.g., *B. subtilis*) and Mycobacteriales (e.g., *M. tuberculosis*) (Vollmer et al. 2008; Pazos and Peters 2019). In addition, the meso-DAP of *B. subtilis* and *M.*

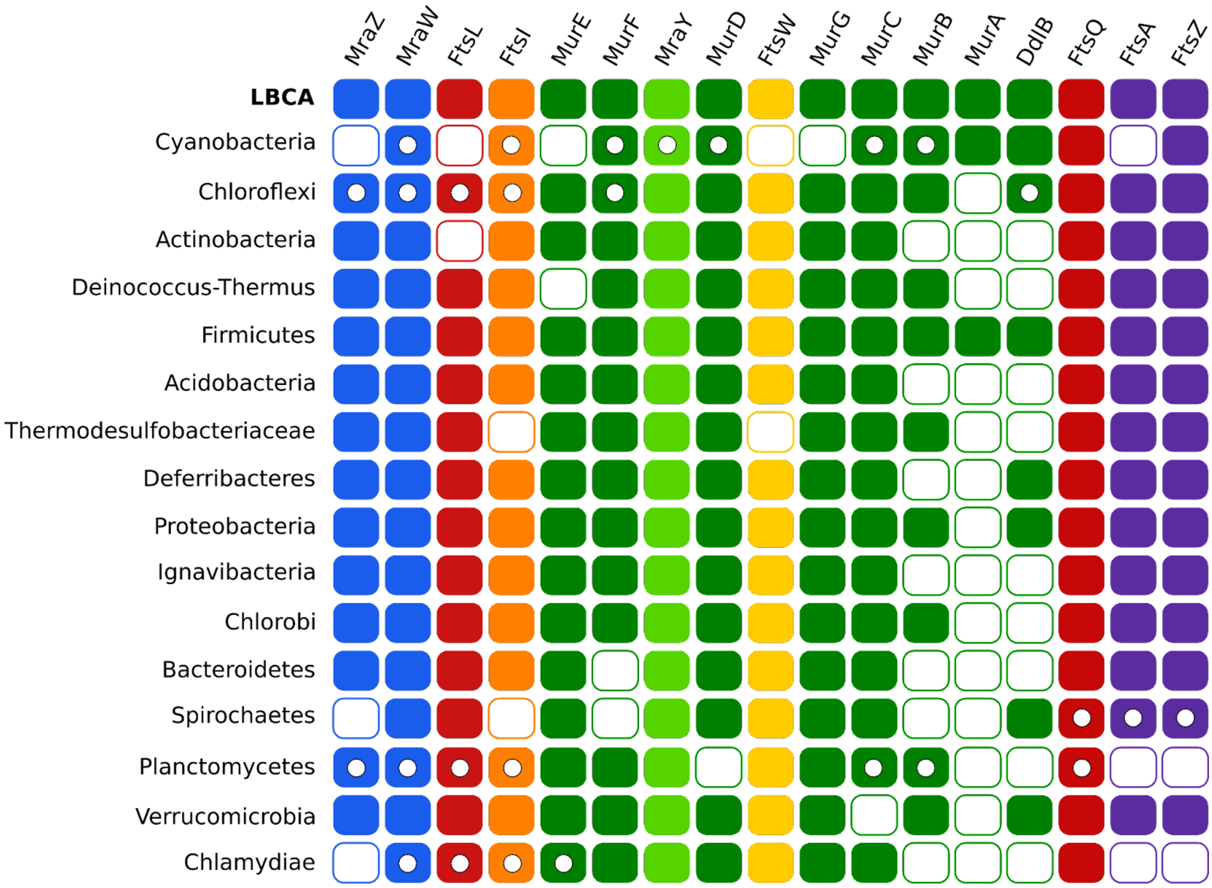
tuberculosis is amidated by the AsnB amidotransferase (Atrih et al. 1999; Dajkovic et al. 2017; Ngadjeua et al. 2018). Spirochetes have L-ornithine (L-Orn) (Schleifer and Kandler 1972; Yanagihara et al. 1984), while other species have diamino acids, such as meso-lanthionine or L-2,4-Diaminobutyric acid (L-DABA), or monoamino acids like L-homoserine (L-Hse), L-Ala or L-Glu (Vollmer et al. 2008; Pazos and Peters 2019). The dipeptide D-Ala-D-Ala in position 4 and 5 is found in almost all bacteria. However, D-Ala in position 5 is replaced by a D-serine (D-Ser) or a D-lactate (D-Lac) in species that have acquired the *vanA, vanB or vanC* genes, which confer a resistance toward the vancomycin antibiotic (Healy et al. 2000).

1.2.1.3.3. Variation in cross-links

The most abundant cross-link found in PG is the 4-3 cross-link (Fig. 2), which connects the D-Ala to the meso-DAP (or L-Lys). This linkage, performed by DD-transpeptidases (DD-TPases) (Sauvage et al. 2008), can either be direct, like in *E. coli* or *B. subtilis*, or through an interpeptide bridge composed of five Gly in *S. aureus* or L-Ala-L-Ser in *S. pneumoniae*. There also exist minor 3-3 cross-links in *E. coli* or *M. tuberculosis* for instance (Vollmer et al. 2008; Pazos and Peters 2019). This kind of linkage is made by LD-TPases (Magnet et al. 2008). In *Corynebacterium pointsettiae*, the third AA of the stem peptide is a L-Hse, which can not be involved in cross-links. Therefore, the D-Ala in position 1 of a peptide is connected via a D-Orn to the D-Glu in position 2 of the second peptide, forming a 4-2 cross-link (Schleifer and Kandler 1972). In 2016, an unusual 1-3 cross-link has been described in Acetobacteria, where the L-Ala is connected to an amidated meso-DAP (Espaillat et al. 2016).

1.2.1.3.4. The *dcw* gene cluster

Many genes involved in PG biosynthesis lie in the division and cell-wall synthesis (dcw) cluster. Even if some species lack specific genes (Pilhofer et al. 2008; Martínez-Torró et al. 2021), the composition of this cluster and its gene order are well conserved across bacterial lineages (Tamames 2001; Mingorance and Tamames 2004; Real and Henriques 2006). The most complete version of the cluster includes 17 genes (Fig. 3), among which six (*ftsA*, *ftsI*, *ftsL*, *ftsQ*, *ftsW*, *ftsZ*) are involved in division, whereas nine (*ddlB*, *murA*, *murB*, *murC*, *murD*, *murE*, *murF*, *murG*, *mraY*) are involved in synthesis of PG precursors. The last two genes (*mraW*, *mraZ*) are neither involved in cell division nor PG biosynthesis. Instead, both genes code for regulatory proteins (Kimura and Suzuki 2010; Eraso et al. 2014).



	MraZ	MraW	FtsL	FtsI	MurE	MurF	MraY	MurD	FtsW	MurG	MurC	MurB	MurA	DdlB	FtsQ	FtsA	FtsZ
LBCA	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Cyanobacteria	□	◙	□	◙	□	◙	◙	◙	□	□	◙	◙	■	■	■	□	■
Chloroflexi	◙	◙	◙	◙	■	◙	■	■	■	■	■	■	□	◙	■	■	■
Actinobacteria	■	■	□	■	■	■	■	■	■	■	■	□	□	□	■	■	■
Deinococcus-Thermus	■	■	■	■	□	■	■	■	■	■	■	■	□	□	■	■	■
Firmicutes	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Acidobacteria	■	■	■	■	■	■	■	■	■	■	■	□	□	□	■	■	■
Thermodesulfobacteriaceae	■	■	■	□	■	■	■	■	□	■	■	■	□	□	■	■	■
Deferribacteres	■	■	■	■	■	■	■	■	■	■	■	□	□	■	■	■	■
Proteobacteria	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■
Ignavibacteria	■	■	■	■	■	■	■	■	■	■	■	□	□	□	■	■	■
Chlorobi	■	■	■	■	■	■	■	■	■	■	■	■	□	□	■	■	■
Bacteroidetes	■	■	■	■	■	□	■	■	■	■	■	□	□	□	■	■	■
Spirochaetes	□	■	■	□	■	□	■	■	■	■	■	□	□	■	◙	◙	◙
Planctomycetes	◙	◙	◙	◙	■	■	■	□	■	■	◙	◙	□	□	◙	□	□
Verrucomicrobia	■	■	■	■	■	■	■	■	■	■	□	■	□	■	■	■	■
Chlamydiae	□	◙	◙	◙	◙	■	■	■	■	■	■	□	□	□	■	□	□

Figure 3. Ancestral state of the *dcw* cluster in the last bacterial common ancestor (LBCA) and in the ancestors of various bacterial phyla (adapted from Léonard et al. 2022). Full rectangle = gene present and in the main cluster; empty circle in rectangle = gene present but in a sub-cluster; empty rectangle = gene present but outside of any cluster.

1.2.1.3.5. Synthesis of the peptidoglycan

The synthesis of PG has been well characterized in *E. coli* and occurs in two main stages, as reviewed in Pazos and Peters 2019 and Egan et al. 2020 (Fig. 4). First, the PG precursor is synthesized in the cytoplasm. Then, the precursor is exported to the periplasm (i.e., the space between the inner CM and the outer LPS membrane), where it is assembled into the nascent PG molecule. The first steps start with the formation of the two glycosidic units. Three enzymes (GlmS, GlmM, GlmU) synthesize the UDP-GlcNAc from fructose 6-phosphate. A fraction of UDP-GlcNAc is converted into UDP-MurNAc by the action of MurA and MurB. Then, L-Ala, D-Glu, meso-DAP and the D-Ala-D-Ala dipeptide are sequentially added to the UDP-MurNAc by four muramyl ligases: MurC, MurD, MurE and MurF. Prior to addition to the UDP-MurNAc, racemases catalyze the formation of D-AAs from their corresponding L- enantiomers, whereas the DdlA and DdlB ligases link two D-Ala to form the D-Ala-D-Ala dipeptide. The variation in stem peptide, which we previously discussed in this chapter, depends on the respective affinities of the four muramyl ligases. MurD and MurF are the most specific enzymes and always add the D-Glu and the dipeptide, respectively (Vollmer et al. 2008). It has been shown that MurC enzymes of *M. tuberculosis*, *M. leprae* and *Chlamydia trachomatis* have the same *in vitro* specificity toward L-Ala, Gly and L-Ser (only for *C. trachomatis*) (Mahapatra et al. 2000; Hesse et al. 2003). Although MurE shows a high specificity toward meso-DAP in *E. coli* and L-Lys in *S. aureus* (Vollmer et al. 2008), it shows a lower specificity in other species, such as *Thermotoga maritima*, where *in vitro* it can add L-Lys, D-Lys and meso-DAP with same efficiency (Boniface et al. 2006). The UDP-MurNAc-pentapeptide is transferred to the lipid carrier undecaprenyl phosphate by MraY, located in the inner leaflet of the CM, to form the lipid I. Then, MurG transfers the UDP-GlcNAc to the lipid I to form lipid II. The latter is flipped from the inner leaflet of the CM to the outer leaflet in the periplasmic side by a flippase. The identity of this flippase is controversial and RodA, FtsW and MurJ have all been proposed as potential candidates (Höltje 1998; Ruiz 2008; Mohammadi et al. 2011; Mohammadi et al. 2014). However, recent studies have shown that RodA and FtsW are more likely to be glycosyltransferases (GTases) (Cho et al. 2016; Meeske et al. 2016; Taguchi et al. 2019). In the periplasmic space, glycan strands are polymerized by GTases and linked to pre-existing strands by transpeptidases (TPases) that cross-link two adjacent pentapeptides. These functions are performed by penicillin-binding proteins (PBPs), which are divided into three groups: 1) the class A PBPs, which are bifunctional enzymes with both GTase and TPase activity, 2) the class B PBPs, which have only a TPase activity, and 3) the class C PBPs, which have both TPase and carboxypeptidase activity. In *E*. coli, there exist twelve different PBPs, of which three class A and two class B enzymes (Goffin and Ghuysen 1998; Sauvage et al. 2008)].

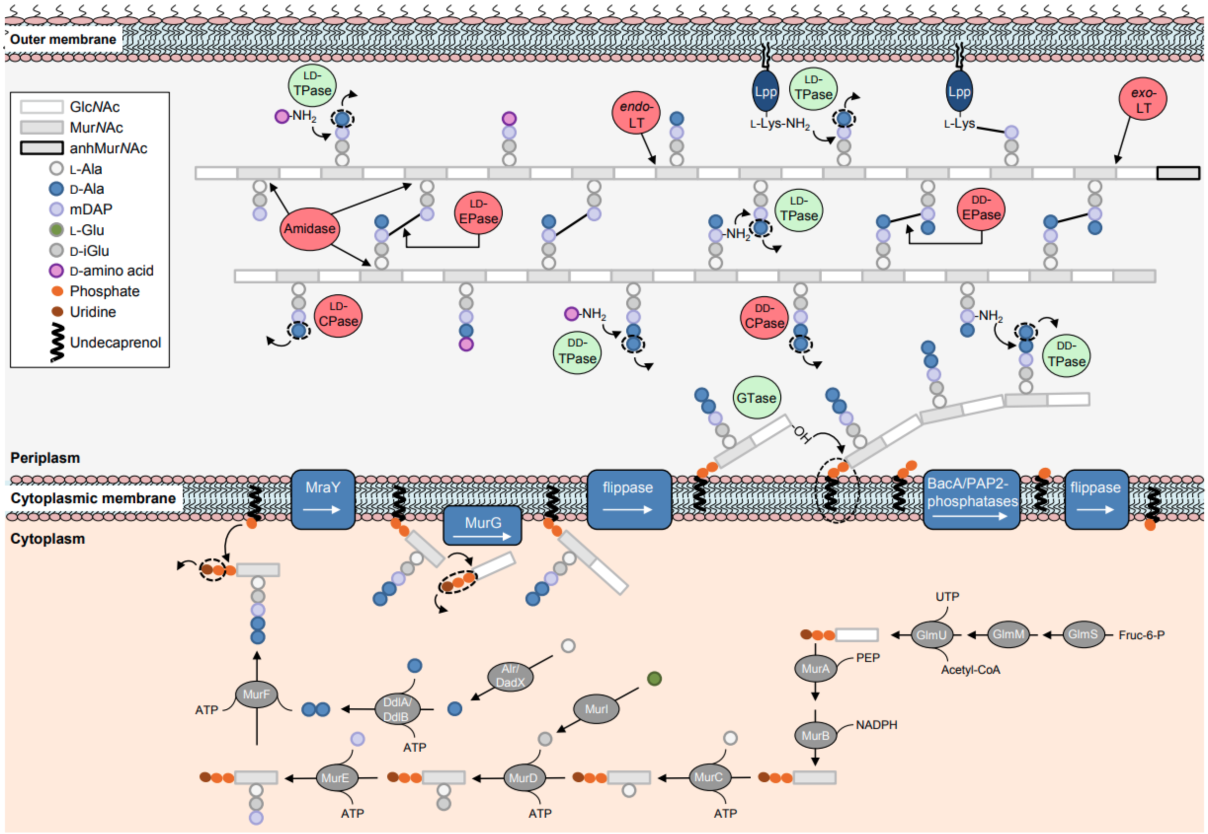


Figure 4. Schematic view of peptidoglycan biosynthesis in *E. coli* (from Pazos and Peters 2019).

1.2.1.4. Cell-wall targeting antibiotics

1.2.1.4.1. Antibiotics overview

After returning from his vacation in September 1928, the Scottish microbiologist Alexander Fleming observed that one old Petri plate was contaminated by a blue-green mold. Interestingly, this mold had created a halo zone around a *Staphylococcus* colony, corresponding to a zone where bacterial cells had undergone lysis. Actually, this mold named *Penicillium notatum* synthesizes a molecule that kills bacteria, which was named penicillin (Fleming 1929; Bennett and Chung 2001). Penicillin was isolated by Ernst Chain and Howard Florey in the early 1940s (Gaynes 2017) and widely used at the end of World War II to heal wounded soldiers (Ventola 2015). Since then, numerous bacteria and fungi that naturally produce antimicrobial compounds have been identified, notably *Actinomycetes* species, which are known to produce different classes of antibiotics with different modes of action (Hutchings et al. 2019; De Simeis and Serra 2021). In addition to natural antibiotics, the pharmaceutical industry also develops synthetic and semi-synthetic antibiotics (Hutchings et al. 2019; Leisner 2020; Miethke et al. 2021). The release of these antibiotics has saved many lives. However, their intensive use in both human health and industrial farming has led to antimicrobial resistance (AMR) in bacteria, which makes the control of pathogenic bacteria more complicated. These AMR genes are often found on plasmids, which is a major factor of AMR dissemination between bacterial species through horizontal gene transfer (HGT) (Feng et al. 2022).

1.2.1.4.2. Classes of antibiotics

The different antibiotics can be classified according to their mode of action. Sulfonamides and derivatives (e.g., sulfamates and sulfamides) are synthetic compounds, which were the first class of antibiotics used against pathogenic bacteria, before the introduction of penicillin. They are structural analogues of the para-aminobenzoic acid (PABA) and inhibit nucleic acid metabolism by interfering with dihydropteroate synthase and dihydrofolate reductase enzymes of the folic acid pathway (Bhattacharjee 2016a; Kapoor et al. 2017; Supuran 2017). Like sulfonamides, most of the protein synthesis inhibitors are bacteriostatic (i.e., they stop cell multiplication without killing), except for aminoglycosides, which are bactericidal (i.e., they kill bacteria) (Bhattacharjee 2016b; Kapoor et al. 2017). Many antibiotics can inhibit protein synthesis by targeting different elements of the ribosome. Tetracyclines and aminoglycosides bind to the 30S subunit, while chloramphenicol, macrolides, lincosamides, oxazolidinones and streptogramins bind to the 50S subunit of the ribosome. Instead of binding to the ribosome, mupirocin binds to isoleucyl tRNA synthetase. There also exist antibiotics that target different stages of PG synthesis. At the cytoplasmic stage (see Synthesis of the peptidoglycan), fosfomycin inhibits the MurA enzyme, while D-cycloserine is a structural analogue of D-Ala and acts as a competitive inhibitor of racemases and Ddl enzymes. Glycopeptides (e.g., vancomycin) and β-lactams (e.g., penicillin) block the formation of the PG in the periplasm, by respectively inhibiting the GTase and the TPase activity of the PBPs (see Synthesis of the peptidoglycan). Those kinds of antibiotics are bactericidal. Indeed, inhibiting PG synthesis leads to cell lysis (Bhattacharjee 2016c; Kapoor et al. 2017).

1.2.1.4.3. Beta-lactam antibiotics

The β-lactams are the most widely used class of antibiotics so far. It has been estimated that 55% of the antibiotics in use belong to this class (Bhattacharjee 2016c). The core structure of these antibiotics is the β-lactam ring (Fig. 5), which is a four-membered cyclic amide (Aoki and Okuhara 1980). This β-lactam ring is a structural analogue of the D-Ala-D-Ala dipeptide from the PG precursor. Therefore, β-lactam antibiotics bind to PBP and block the active site by suicide inhibition. Indeed, during the acylation step, the antibiotic is covalently linked to the serine found in the active site of all PBPs. However, the resulting acyl enzyme cannot be hydrolyzed efficiently (Fig. 6). Consequently, the TPase activity of PBPs is inhibited, which prevents cross-linking between glycosidic chains and leads to cell lysis (Bush and Bradford 2016). From natural β-lactam antibiotics, the pharmaceutical industry develops semi-synthetic antibiotics by chemical modification of substituents attached to the β-lactam ring (Elander 2003). The β-lactam antibiotics are divided into four families: penicillins, cephalosporins, carbapenems and monobactams.

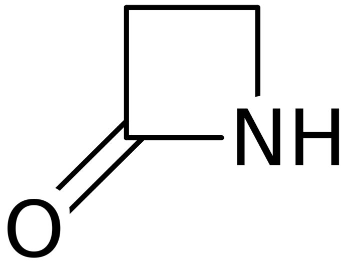


Figure 5. Beta-lactam ring.

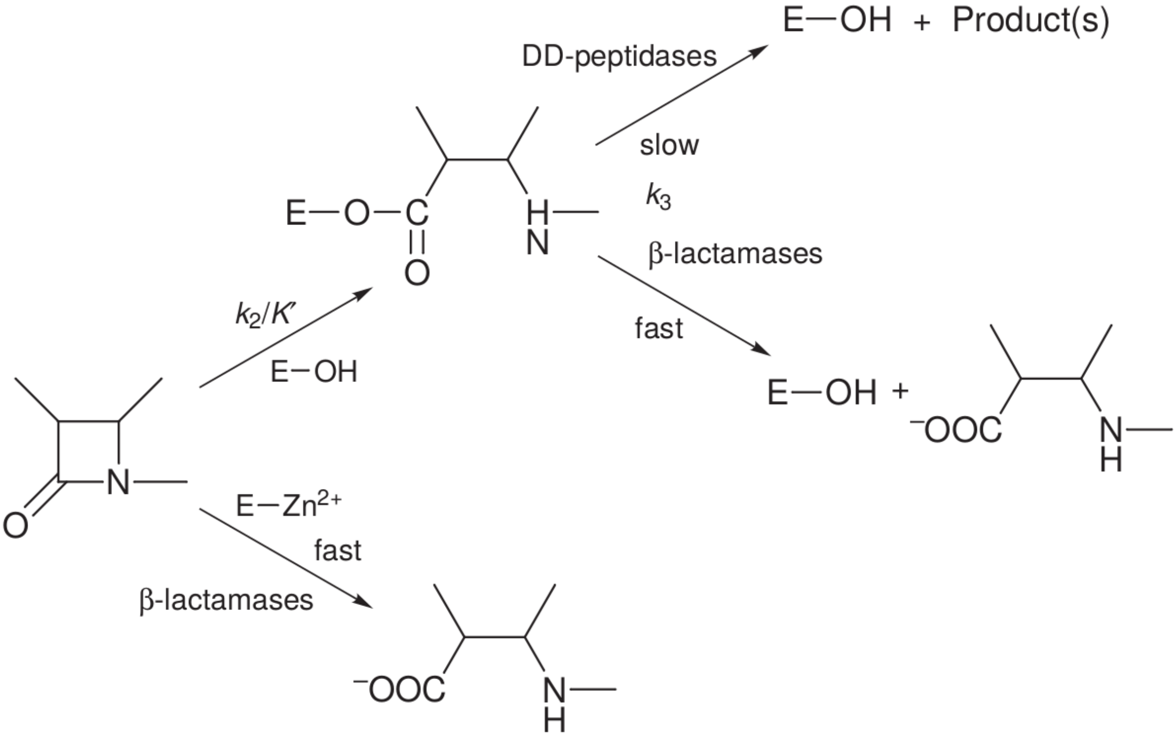


Figure 6. Chemical reaction between PBPs (DD-peptidases)/beta-lactamases and beta-lactam antibiotics (from Matagne et al. 1999). E-OH represents active-site serine enzymes (i.e., PBPs, class A, C and D beta-lactamases), while E-Zn^{2+} represents metallo-beta-lactamases. E-OH and beta-lactam antibiotics form a covalent intermediate (acyl-enzyme). Beta-lactamases can hydrolyze the substrate, whereas PBPs hydrolyze the substrate poorly or not at all, and remain stuck as an inactive acyl-enzyme. E-Zn^{2+} does not form a covalent intermediate and hydrolyzes the beta-lactam ring directly by a water molecule activated by coordination to the zinc ion(s).

Penicillins gather β-lactam antibiotics having a 6-aminopenicillanic acid (6-APA) nucleus (Fig. 7) like in penicillin G (or benzylpenicillin), which was the first antibiotic ever used at large-scale (Ball et al. 1978). Penicillins are further classified into four generations (Fig. 8), according to their side-chain residues, which confer different activity spectra to the antibiotic. Put simply, the activity spectrum of an antibiotic is the range of microorganisms it can inhibit or kill. The first generation groups natural penicillin, such as penicillin G and penicillin V, whereas the second (e.g., oxacillin, methicillin), the third (e.g., amoxicillin, ampicillin) and the fourth generation (e.g., carboxypenicillin, ureidopenicillin) are semi-synthetic β-lactams (Lobanovska and Pilla 2017).

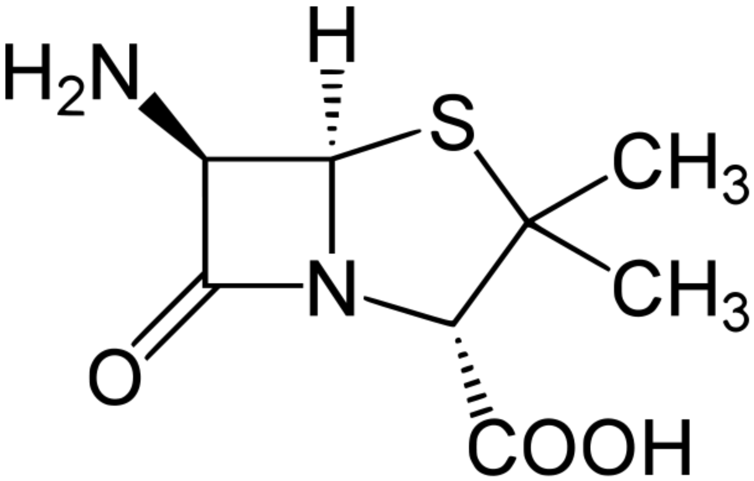


Figure 7. 6-aminopenicillanic acid ring (6-APA).

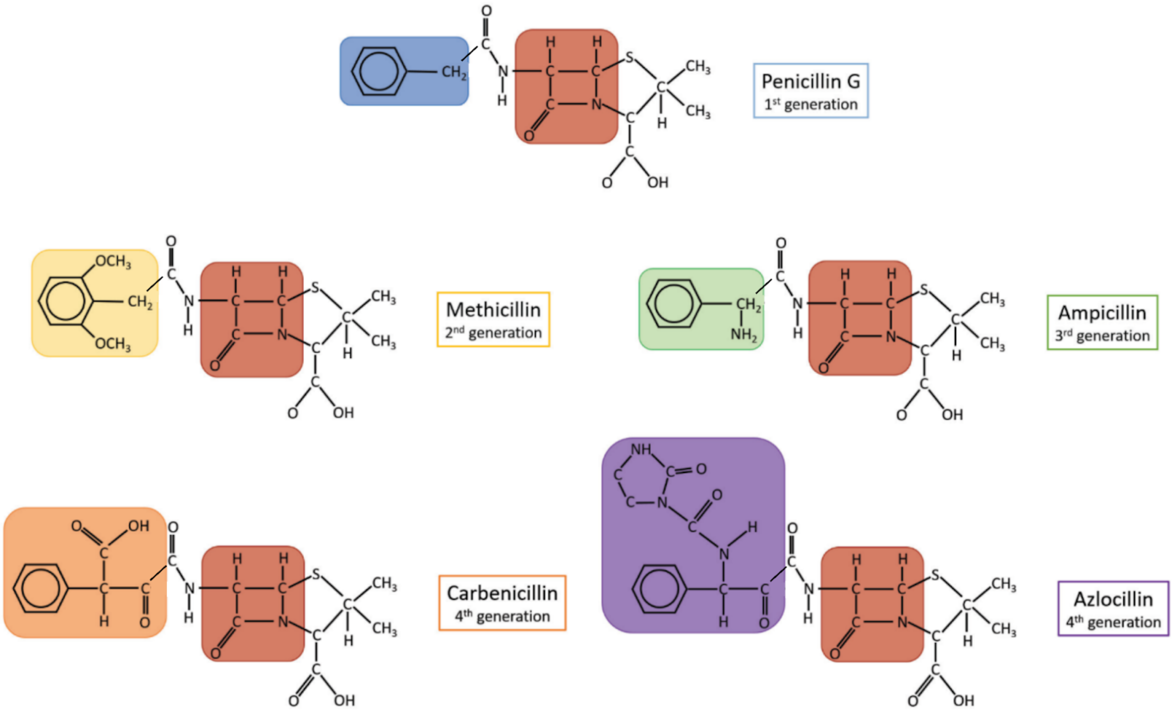


Figure 8. Molecular structure of the four generations of penicillins (adapted from Lobanovska and Pilla 2017). The β-lactam ring is framed in red and the different side-chains follow a color code. Penicillin G (first generation) is shown in blue, methicillin (second generation) in yellow, ampicillin (third generation) in green, carbenicillin (fourth generation) in orange, and in purple azlocillin (fourth generation).

The first cephalosporin, the cephalosporin C, was extracted from the fungi *Acremonium chrysogenum* (previously *Cephalosporium acremonium*) in 1953 by Newton and Abraham (Newton and Abraham 1955). In contrast to penicillins, cephalosporins have a 7-aminocephalosporanic acid (7-ACA) nucleus (Fig. 9) (Jago and Heatley 1961). They are classified into five generations, according to their activity spectrum (Page 2012).

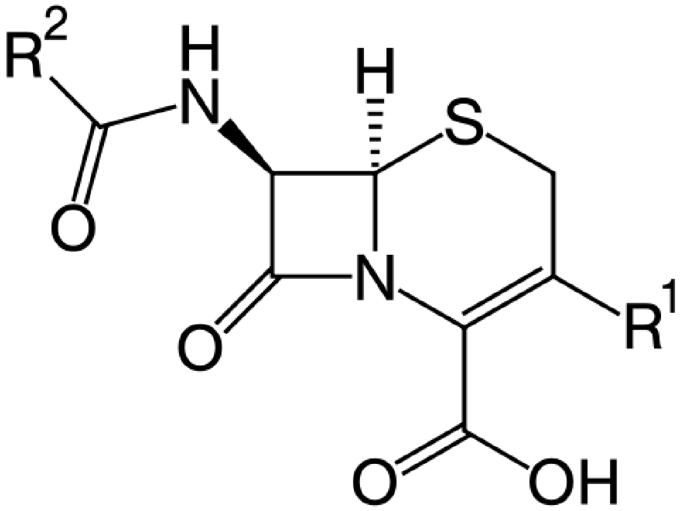


Figure 9. 7-aminocephalosporanic acid (7-ACA).

Carbapenems are β-lactam antibiotics that were first isolated from *Streptomyces* species. They have a similar structure to penicillin but differ in the double bond between carbon C-2 and C-3, and the sulfur atom at position C-1 is replaced by a carbon (Fig. 10). Carbapenems have high clinical relevance, as they are used as last resort antibiotics against multidrug-resistant bacteria (Papp-Wallace et al. 2011).

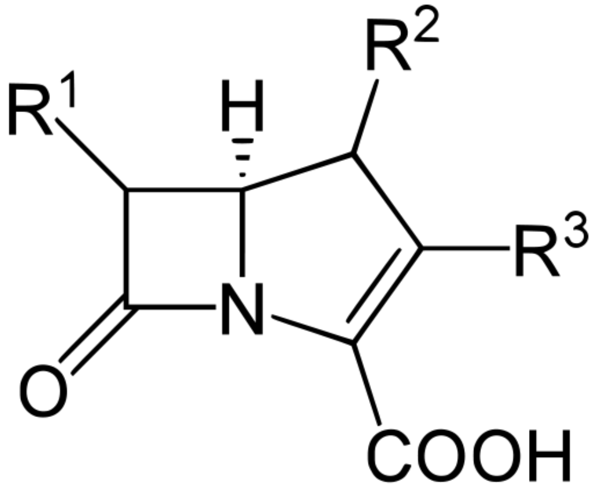


Figure 10. Carbapenem ring.

The fourth family of β-lactam antibiotics is the monobactams, which are characterized by the monocycle β-lactam ring (Fig. 11). Aztreonam is a synthetic antibiotic and the only monobactam used in human health (Page 2012; Fernandes et al. 2013; Ramsey and MacGowan 2016).

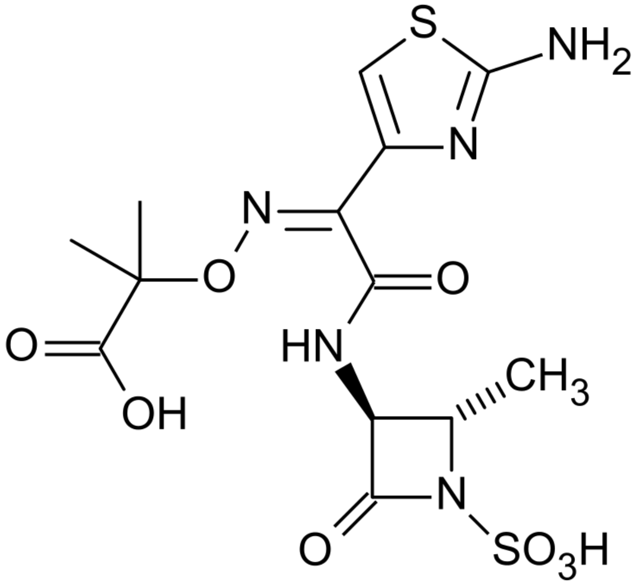


Figure 11. Aztreonam.

1.2.1.5. Antibiotic-resistant bacteria

The extensive use of antimicrobial therapy since the middle of the 20th century has led to the rise of resistant bacteria. A bacterium is considered as resistant when antibiotics can not inhibit growth efficiently (Zaman et al. 2017). The first resistance in bacteria was reported in the late 1930s, shortly after the introduction of sulfonamides in 1937. Even before the therapeutic use of penicillin in the late 1940s, Abraham and Chain described 1940 an enzyme able to hydrolyze penicillin (Davies and Davies 2010; Zaman et al. 2017). Ever since, numerous new resistances have been identified, which follow the deployment of specific antibiotics (Fig. 12) (Ventola 2015). AMR genes have the ability to spread out bacterial species via plasmidic vectors and can lead to multi-resistant bacteria, such as the methicillin-resistant *S. aureus* (MRSA), which is resistant to almost all classes of antibiotics, at the exception of glycopeptides (including vancomycin) targeting PG synthesis. Nonetheless, some vancomycin-resistant *S. aureus* (VRSA) have also been reported (Haaber et al. 2017).

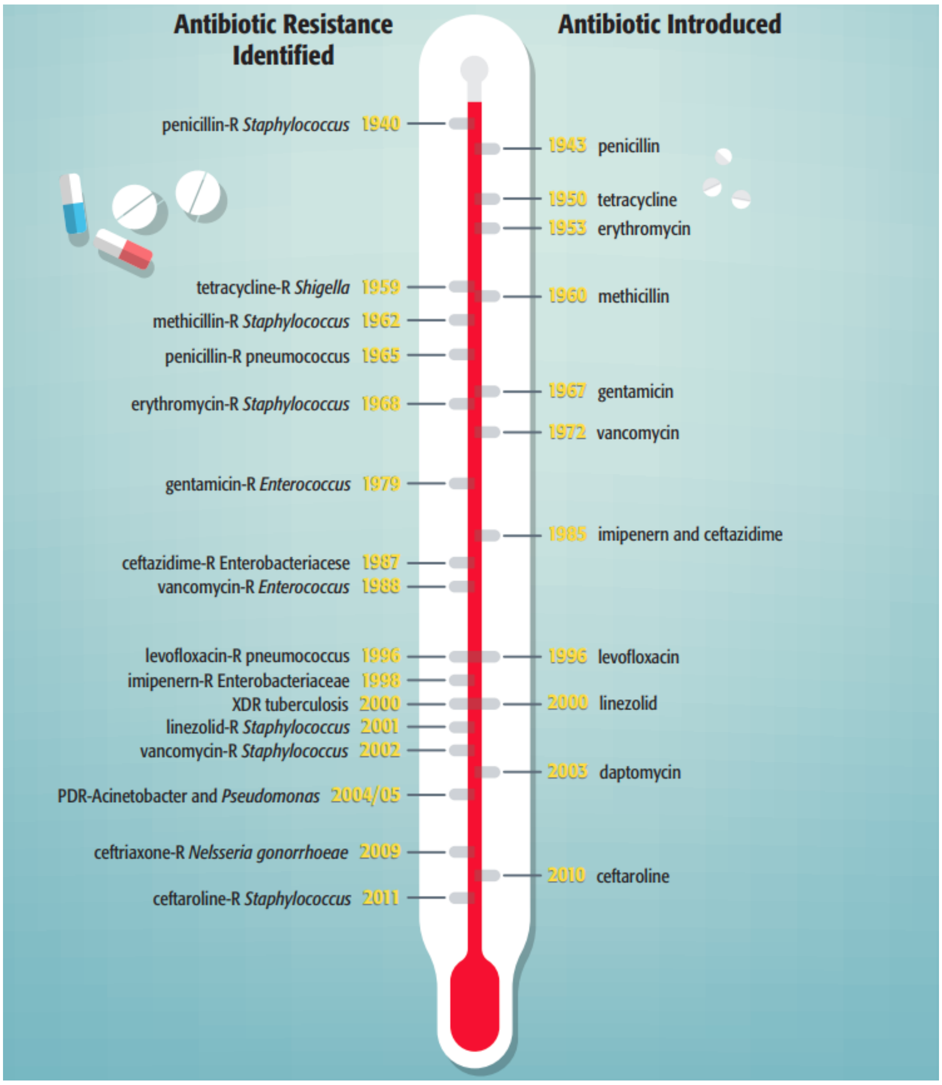

Antibiotic Resistance Identified
Antibiotic Introduced
penicillin-R *Staphylococcus* 1940
1943 penicillin
1950 tetracycline
1953 erythromycin
tetracycline-R *Shigella* 1959
1960 methicillin
methicillin-R *Staphylococcus* 1962
penicillin-R pneumococcus 1965
1967 gentamicin
erythromycin-R *Staphylococcus* 1968
1972 vancomycin
gentamicin-R *Enterococcus* 1979
1985 imipenern and ceftazidime
ceftazidime-R Enterobacteriacese 1987
vancomycin-R *Enterococcus* 1988
levofloxacin-R pneumococcus 1996
1996 levofloxacin
imipenern-R Enterobacteriaceae 1998
XDR tuberculosis 2000
2000 linezolid
linezolid-R *Staphylococcus* 2001
vancomycin-R *Staphylococcus* 2002
2003 daptomycin
PDR-Acinetobacter and *Pseudomonas* 2004/05
ceftriaxone-R *Nelsseria gonorrhoeae* 2009
2010 ceftaroline
ceftaroline-R *Staphylococcus* 2011

Figure 12. Timeline of key antibiotic resistance events (from www.biomerieux-usa.com).

1.2.1.5.1. Resistance to β-lactam antibiotics

Bacteria have developed different strategies In order to protect themselves against β-lactam antibiotics. The two major strategies are the synthesis of PBPs exhibiting a decreased affinity for β-lactam antibiotics or the production of specific hydrolytic enzymes, named β-lactamases (Frère 1995). Moreover, diderm-LPS bacteria have the ability to decrease the concentration of antibiotics in their periplasmic space by reducing membrane permeability and/or by activating efflux pumps to extrude antibiotics (Munita and Arias 2016).

1.2.1.5.1.1. Beta-lactamases

β-lactamases are enzymes able to hydrolyze the amide bond of the β-lactam ring, which makes the antibiotic inefficient (Fig. 6). There exist two systems to classify those enzymes: the Bush–Jacoby–Medeiros system, based on the activity spectrum (Bush et al. 1995), and the Ambler system, based on the AA sequence (Ambler 1980). The latter is the most widely used classification for β-lactamases (Hall and Barlow 2005), and it is the one that will be used in this thesis. According to the Ambler system, β-lactamases are divided into classes A, B, C and D. Classes A, C and D group active-site serine β-lactamases (SBLs), while all metallo-β-lactamases (MBLs) are included in class B (Babic et al. 2006; Palzkill 2013). β-lactamase genes can be either chromosome- or plasmid-encoded. Owing to the structural and mechanistic similarities between PBPs and SBLs, it was proposed that SBLs evolved from PBPs, probably due to competition with β-lactam-producing microorganisms. Therefore, the origin of β-lactamases dates back to before the selection pressure generated by the extensive use of antibiotics (Poole 2004; Bush 2018). Acquisition of point mutations in β-lactamase sequence can directly affect the activity spectrum of the enzyme. Consequently, this phenomenon has generated extended-spectrum beta-lactamases (ESBLs), which possess an activity against all penicillins, as well as cephalosporins from the first to the third generation and monobactams. ESBLs are mainly found in class A, while some class C and class D enzymes are also characterized as ESBLs (Paterson and Bonomo 2005; Rawat and Nair 2010; Tooke et al. 2019; Sawa et al. 2020; Castanheira et al. 2021). Moreover, carbapenemases are β-lactamases with a very broad spectrum of action that are able to hydrolyze carbapenems, penicillins, cephalosporins and aztreonam. Due to their ability to inactivate last resort antibiotics (i.e., carbapenems), the spread of carbapenemases during the last 20 years is a real burden for public health. The carbapenemase activity has been reported in classes A,B and D enzymes (Queenan and Bush 2007; Hammoudi Halat and Ayoub Moubareck 2020; Sawa et al. 2020). However, as reviewed in (Philippon et al. 2022), class C enzymes exhibiting carbapenem resistance are found only in association with porin impairment or efflux pump overexpression. On the other hand, the action of some β-lactamases can be blocked by β-lactam-containing (e.g., clavulanic acid, sulbactam, tazobactam) or by non-β-lactam (e.g., avibactam, relebactam, vaborbactam, zedibactam, nacubactam) β-lactamase inhibitors (Fig. 13) (Bush and Bradford 2016; Tooke et al. 2019; Carcione et al. 2021). As for August 2022, 7537 β-lactamases have been recorded in the Beta-Lactamase DataBase (BLDB) (Naas et al. 2017).

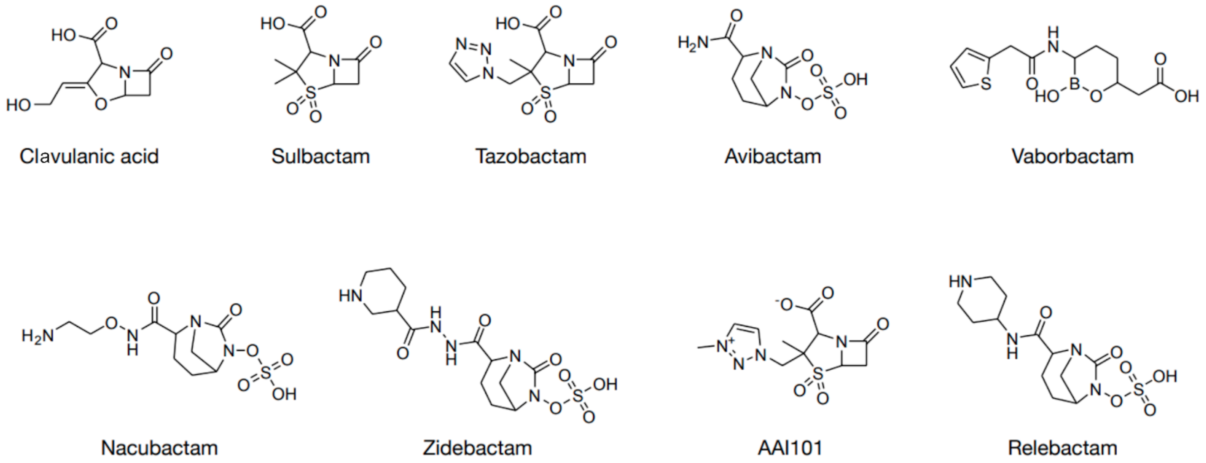


Figure 13. Chemical structures of beta-lactamase inhibitors (from Eiamphungporn et al. 2018).

1.2.1.5.1.1.1. Class A

The class A is the most studied class of β-lactamases and contains the most diverse set of enzymes in terms of number of families. TEM, SHV, CTX-M and KPC are the four most widespread class A families. The TEM family contains plasmid-encoded β-lactamases. The name is derived from **Tem**oniera, the Greek patient infected by an *E. coli* strain containing the TEM-1 gene, which was the first plasmid-borne β-lactamase ever isolated, in 1963. Since then, numerous ESBL variants of TEM-1 have been described. They differ from TEM-1 by one to five substitutions in their AA sequence (Salverda et al. 2010; Tooke et al. 2019; Sawa et al. 2020; Castanheira et al. 2021). In the 1970s, the first plasmid-encoded SHV-1 (**s**ulf**h**ydryl **v**ariant) β-lactamase was isolated in *E. coli* (Liakopoulos et al. 2016). However, a study of 1997 revealed that the genome of *Klebsiella pneumoniae* encodes an SHV-1-like gene. Thus, it was proposed that the latter gene reflects the ancestral state of SHV enzymes (Haeggman et al. 1997). As for TEM, there exist numerous ESBL and non-ESBL variants of SHV-1 (Liakopoulos et al. 2016; Tooke et al. 2019; Sawa et al. 2020; Castanheira et al. 2021). CTX-M (**c**efo**t**a**x**imase from **M**unich) β-lactamases were first reported in the late 1980s. All CTX-M are plasmid-encoded and characterized as ESBLs. In contrast to TEM and SHV families, the different CTX-M enzymes have much more AA sequence diversity (Castanheira et al. 2021). Since the early 2000s, we have been facing an important dissemination of CTX-M genes among bacterial pathogens, which makes now the CTX-M family the most widespread ESBL group (Cantón et al. 2012; Tooke et al. 2019; Castanheira et al. 2021). Regarding the KPC (*K. pneumoniae* carbapenemase) family, it is the most well known example of class-A carbapenemase. KPC genes are encoded on plasmids, which are mainly found in *Enterobacteriaceae* (Queenan and Bush 2007; Tooke et al. 2019; Sawa et al. 2020).

1.2.1.5.1.1.2. Class B

MBLs are structurally and mechanistically different from the three classes of SBLs. Indeed, during β-lactam hydrolysis, MBLs do not form a covalent acyl enzyme intermediate (Fig. 6). Instead, β-lactam antibiotics are directly hydrolyzed by nucleophilic attack of the hydroxyde (OH^-), which is stabilized by 1 or 2 Zn^{2+} ion(s) present in the active site (Palzkill 2013; Bonomo 2017). Actually, MBLs are not related to SBLs and PBPs but instead are members of the metallohydrolase superfamily (Tooke et al. 2019). MBLs are active against all families of β-lactam antibiotics except monobactam (Bebrone 2007). Moreover, they are not inactivated by β-lactamase inhibitors (Mojica et al. 2022). Based on the AA sequence, MBLs are divided into three subclasses: B1, B2 and B3 (Galleni et al. 2001). The AA sequence identity between the different subclasses is really low (<20%). Subclasses B1 and B3 have two Zn^{2+} atoms in their active site while subclass B2 only has one Zn^{2+}. The subclass B1 includes enzymes with high clinical relevance, such as the plasmid-encoded IMP (imipenemase), NDM (New Delhi metallo-β-lactamase) or VIM (Verona imipenemase) families (Bebrone 2007; Palzkill 2013; Sawa et al. 2020).

1.2.1.5.1.1.3. Class C

Class C β-lactamases (also known as AmpC) are inducible cephalosporinases that are encoded by the chromosome of numerous bacteria, particularly Proteobacteria (Philippon et al. 2022). AmpC enzymes also have a low ability to hydrolyze monobactams and are poorly inhibited by β-lactamases inhibitors like clavulanic acid, sulbactam and tazobactam (Jacoby 2009; Philippon et al. 2022). Furthermore, some AmpC can exhibit a weak carbapenemase activity (Hammoudi Halat and Ayoub Moubareck 2020; Philippon et al. 2022). It has been shown that an overproduction of AmpC can increase the hydrolysis of β-lactam antibiotics for enzymes exhibiting a poor substrate sensitivity (Lakaye et al. 1999). Although AmpC are mainly chromosome-encoded, many plasmid-encoded enzymes have been described in bacterial species without a chromosomal *ampC* gene (Beceiro and Bou 2004; Doi and Paterson 2007; Philippon et al. 2022).

1.2.1.5.1.1.4. Class D

Historically, class D β-lactamases were distinguished from the other SBLs by the ability of the first two enzymes (i.e., OXA-1 and OXA-2) to hydrolyze oxacilline at a higher rate than penicillin G. Consequently, the term OXA (for oxacillinase) was used to designate class D enzymes. In addition, an increasing number is assigned to each OXA sequence, which merely follows chronological order of identification (Poirel et al. 2010; Leonard et al. 2013). Yet, class D is a very heterogeneous family, where sequence identity can be as low as 17% (Antunes and Fisher 2014). The first described OXA β-lactamases were part of a transposon and carried on plasmids of diverse clinical pathogenic diderm-LPS bacteria (e.g., Enterobacteriaceae) (Poirel et al. 2010; Antunes and Fisher 2014; Evans and Amyes 2014). In 1994, an OXA β-lactamase, termed OXA-12, was described in *Aeromonas sobria* as the first chromosomally encoded OXA (Rasmussen et al. 1994). Later, OXA genes were identified in the chromosome of Enterobacteriaceae, and notably in numerous *Acinetobacter* species (Bou et al. 2000; Bonnet et al. 2002; Poirel et al. 2010; Evans and Amyes 2014; Yoon and Jeong 2021). Although OXAs have for long been described in diderm-LPS bacteria, recent studies showed that the chromosome of Bacilli (Toth et al. 2016) and *Clostridioides difficile* (formerly *Clostridium difficile*) (Toth et al. 2018) also encodes OXAs. One of those chromosome-encoded OXAs, BSD-1 (Toth et al. 2016), is actually the YbxI protein, a class D enzyme exhibiting a low β-lactamases activity, which was identified in *B. subtilis* in the early 2000s (Colombo et al. 2004). The first discovered OXAs exhibited a narrow spectrum of hydrolysis toward β-lactam antibiotics, inactivating only penicillins and first-generation cephalosporins (Poirel et al. 2010; Antunes and Fisher 2014). However, point mutations in some OXAs have led to ESBL enzymes. For instance, OXA-11, OXA-13, OXA-14, OXA-16, OXA-17, OXA-19 and OXA-28 are ESBL variants of OXA-10, which have been first detected in isolates of *Pseudomonas aeruginosa* (Evans and Amyes 2014; Castanheira et al. 2021). Furthermore, OXAs with a carbapenemase activity have been reported, especially in *Acinetobacter* species, where most OXAs are chromosome-encoded (Walther-Rasmussen and Høiby 2006; Evans and Amyes 2014).

Despite a low sequence identity, OXA sequences display highly conserved AA residues that define three motifs (Fig. 14): SxxK, SxV and KTG (Szarecka et al. 2011; Leonard et al. 2013; Antunes and Fisher 2014). The 3-dimensional (3D) structure of OXAs reveals that those motifs are located in the active site of the enzyme. In OXA-10 sequence, the motif S^{115}AV (Serine-Alanine-Valine) is located on the loop between the α-helices α4 and α5, just in front of the motif K^{205}TG (Lysine-Threonine-Glycine) present on the β-sheet β5. At the active site entry, the α-helix α7 from the ω loop contains a W^{154} (Tryptophane), which is also conserved in all studied OXAs. The motif S^{67}TFK (Serine-Threonine-Phenylalanine-Lysine) is found at the N-terminal part of the α-helix α3. The latter S^{67} is actually the active serine (Leonard et al. 2013). In contrast to other SBLs, OXA β-lactamases uniquely have the K^{70} carboxylated, which is essential for their activity (Golemi et al. 2001; Leonard et al. 2013). Finally, an additional motif, Y^{141}GN (Tyrosine-Glycine-Asparagine) at the end of the α-helix α6, is relatively well conserved across OXA sequences, although this motif is outside the active site (Afzal-Shah et al. 2001; Alfredson and Korolik 2005; Antonelli et al. 2015; Toth et al. 2016; Toth et al. 2018).

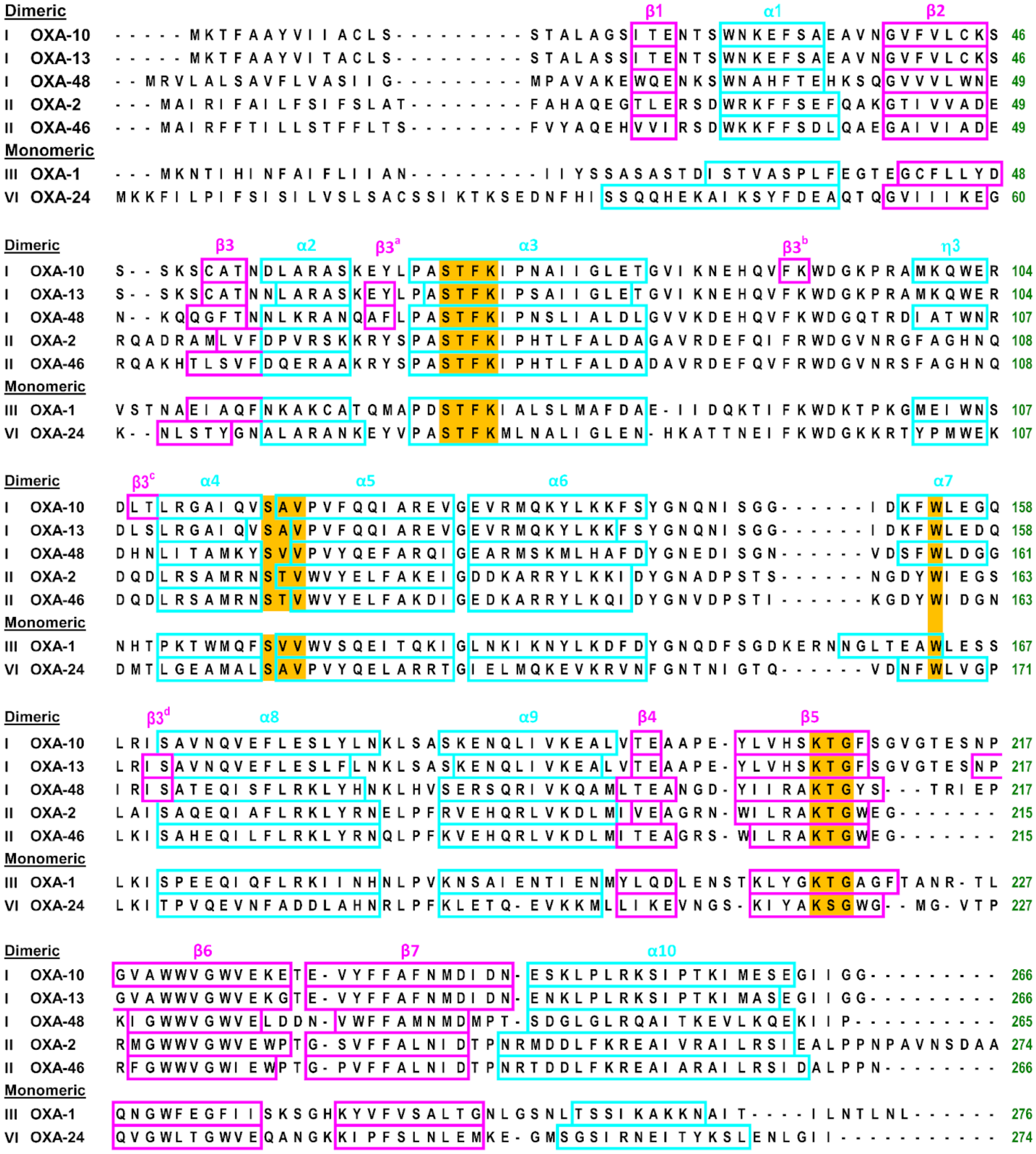

Dimeric
β1 α1 β2
I OXA-10 -----MKTFAAYVIIACLS---------STALAGSITENTSWNKEFSAEAVNGVFVLCKS 46
I OXA-13 -----MKTFAAYVITACLS---------STALASSITENTSWNKEFSAEAVNGVFVLCKS 46
I OXA-48 --MRVLALSAVFLVASIIG---------MPAVAKEWQENKSWNAHFTEHKSQGVVVLWNE 49
II OXA-2 ---MAIRIFAILFSIFSLAT--------FAHAQEGTLERSDWRKFFSEFQAKGTIVVADE 49
II OXA-46 ---MAIRFFTILLSTFFLTS--------FVYAQEHVVIRSDWKKFFSDLQAEGAIVIADE 49
Monomeric
III OXA-1 ---MKNTIHINFAIFLIIAN---------IIYSSASASTDISTVASPLFEGTEGCFLLYD 48
VI OXA-24 MKKFILPIFSISILVSLSACSSIKTKSEDNFHISSQQHEKAIKSYFDEAQTQGVIIIKEG 60
Dimeric
β3 α2 β3a α3 β3b η3
I OXA-10 S--SKSCATNDLARASKEYLPASTFKIPNAIIGLETGVIKNEHQVFKWDGKPRAMKQWER 104
I OXA-13 S--SKSCATNNLARASKEYLPASTFKIPSAIIGLETGVIKNEHQVFKWDGKPRAMKQWER 104
I OXA-48 N--KQQGFTNNLKRANQAFLPASTFKIPNSLIALDLGVVKDEHQVFKWDGQTRDIATWNR 107
II OXA-2 RQADRAMLVFDPVRSKKRYSPASTFKIPHTLFALDAGAVRDEFQIFRWDGVNRGFAGHNQ 108
II OXA-46 RQAKHTLSVFDQERAAKRYSPASTFKIPHTLFALDADAVRDEFQVFRWDGVNRSFAGHNQ 108
Monomeric
III OXA-1 VSTNAEIAQFNKAKCATQMAPDSTFKIALSLMAFDAE-IIDQKTIFKWDKTPKGMEIWNS 107
VI OXA-24 K--NLSTYGNALARANKEYVPASTFKMLNALIGLEN-HKATTNEIFKWDGKKRTYPMWEK 107
Dimeric
β3c α4 α5 α6 α7
I OXA-10 DLTLRGAIQVSAVPVFQQIAREVGEVRMQKYLKKFSYGNQNISGG------IDKFWLEGQ 158
I OXA-13 DLSLRGAIQVSAVPVFQQIAREVGEVRMQKYLKKFSYGNQNISGG------IDKFWLEDQ 158
I OXA-48 DHNLITAMKYSVVPVYQEFARQIGEARMSKMLHAFDYGNEDISGN------VDSFWLDGG 161
II OXA-2 DQDLRSAMRNSTVWVYELFAKEIGDDKARRYLKKIDYGNADPSTS------NGDYWIEGS 163
II OXA-46 DQDLRSAMRNSTVWVYELFAKDIGEDKARRYLKQIDYGNVDPSTI------KGDYWIDGN 163
Monomeric
III OXA-1 NHTPKTWMQFSVVWVSQEITQKIGLNKIKNYLKDFDYGNQDFSGDKERNNGLTEAWLESS 167
VI OXA-24 DMTLGEAMALSAVPVYQELARRTGIELMQKEVKRVNFGNTNIGTQ------VDNFWLVGP 171
Dimeric
β3d α8 α9 β4 β5
I OXA-10 LRISAVNQVEFLESLYLNKLSASKENQLIVKEALVTEAAPE-YLVHSKTGFSGVGTESNP 217
I OXA-13 LRISAVNQVEFLESLFLNKLSASKENQLIVKEALVTEAAPE-YLVHSKTGFSGVGTESNP 217
I OXA-48 IRISATEQISFLRKLYHNKLHVSERSQRIVKQAMLTEANGD-YIIRAKTGYS---TRIEP 217
II OXA-2 LAISAQEQIAFLRKLYRNELPFRVEHQRLVKDLMIVEAGRN-WILRAKTGWEG------- 215
II OXA-46 LKISAHEQILFLRKLYRNQLPFKVEHQRLVKDLMITEAGRS-WILRAKTGWEG------- 215
Monomeric
III OXA-1 LKISPEEQIQFLRKIINHNLPVKNSAIENTIENMYLQDLENSTKLYGKTGAGFTANR-TL 227
VI OXA-24 LKITPVQEVNFADDLAHNRLPFKLETQ-EVKKMLLIKEVNGS-KIYAKSGWG--MG-VTP 227
Dimeric
β6 β7 α10
I OXA-10 GVAWWVGWVEKETE-VYFFAFNMDIDN-ESKLPLRKSIPTKIMESEGIIGG--------- 266
I OXA-13 GVAWWVGWVEKGTE-VYFFAFNMDIDN-ENKLPLRKSIPTKIMASEGIIGG--------- 266
I OXA-48 KIGWWVGWVELDDN-VWFFAMNMDMPT-SDGLGLRQAITKEVLKQEKIIP---------- 265
II OXA-2 RMGWWVGWVEWPTG-SVFFALNIDTPNRMDDLFKREAIVRAILRSIEALPPNPAVNSDAA 274
II OXA-46 RFGWWVGWIEWPTG-PVFFALNIDTPNRTDDLFKREAIARAILRSIDALPPN-------- 266
Monomeric
III OXA-1 QNGWFEGFIISKSGHKYVFVSALTGNLGSNLTSSIKAKKNAIT----ILNTLNL------ 276
VI OXA-24 QVGWLTGWVEQANGKKIPFSLNLEMKE-GMSGSIRNEITYKSLENLGII----------- 274

Figure 14. Multiple alignment of a few OXA beta-lactamases (from F. Kerff, personal communication). α-helices are framed in cyan, while β-sheets are framed in magenta. Conserved AA residues located in the active site are shown in yellow.

Interestingly, OXA sequences exhibit an homology with the C-terminal domain (CTD) of the membrane receptors BlaR found in *B. licheniformis* and *S. aureus*. This homology is more obvious when comparing the 3D structures (Fig. 15) (Leonard et al. 2013). In presence of β-lactam antibiotics, BlaR induces the production of the BlaP (for *B. licheniformis*), or BlaZ (for *S. aureus*), class A β-lactamase (Hardt et al.

1997; Golemi-Kotra et al. 2003). In *B. licheniformis*, BlaR is a 601-AA long membrane protein that is composed of two domains (Fig. 16A): 1) the N-terminal domain (NTD), which consists of four transmembrane segments and three loops (L1 to L3) and, 2) the CTD located outside the CM, which is devoid of β-lactamase activity (Hardt et al. 1997; Joris and Dusart 2012). The BlaR-CTD plays a role of sensor, which detects β-lactam antibiotics, while BlaR-NTD acts as a transducer. The acylation of the active serine of BlaR-CTD by the β-lactam molecule leads to conformational change in the receptor (Joris and Dusart 2012), which activates the autolysis of the BlaR-NTP L3 Zn metallo-protease located in the cytoplasm (Fig. 16B) (Berzigotti et al. 2012; López-Pelegrín et al. 2013). The lysis of the L3 eventually triggers the synthesis of the β-lactamase from the gene of the same operon (Llarrull et al. 2011; Joris and Dusart 2012).

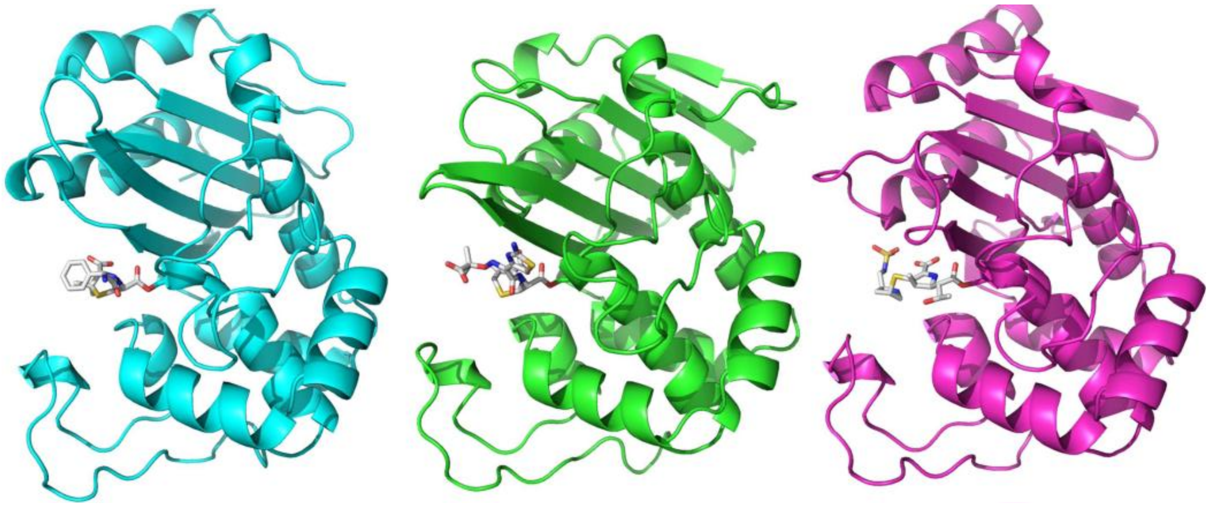


Figure 15. Comparison of 3D structures of OXA-10, OXA-24 and BlaR-CTD (from Leonard et al. 2013). In cyan, OXA-10 forming an acyl-enzyme with ampicillin. In green, OXA-24 forming an acyl-enzyme with a carbapenem. In magenta, BlaR-CTD forming an acyl-enzyme with a cephalosporin.

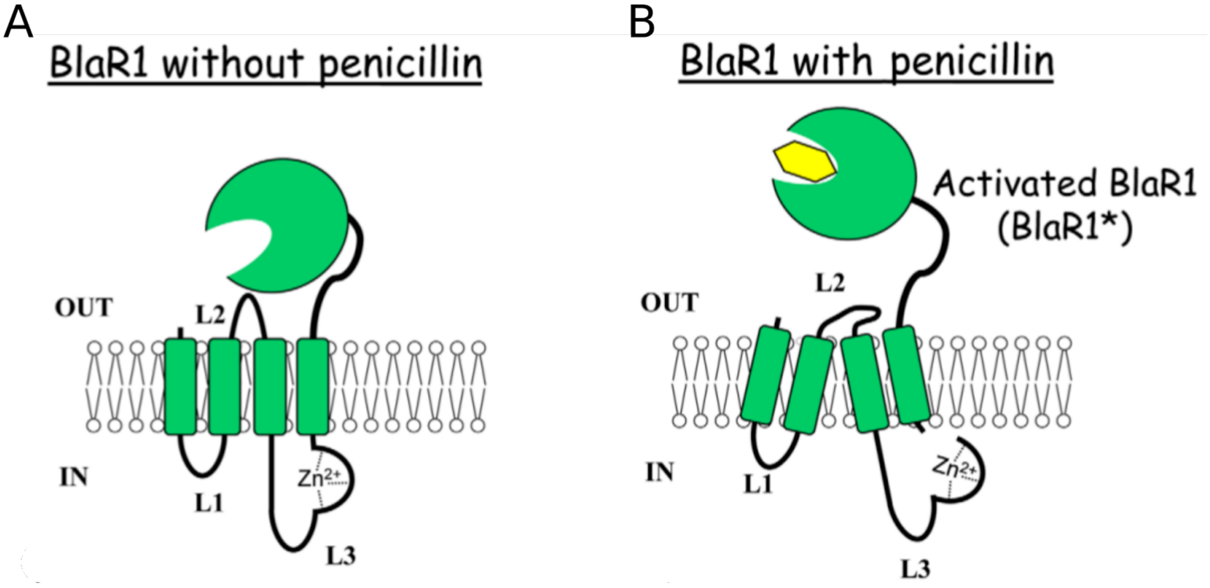


Figure 16. Schematic structure of BlaR from *Bacillus licheniformis* (from Joris and Dusart 2012). (**A**) BlaR in absence of beta-lactam antibiotics. (**B**) BlaR in presence of beta-lactam antibiotics.

1.2.2. Archaeal cell walls

The PG is a universal feature present in the cell wall of almost all bacterial species. However, this polymer is completely absent from archaea. In contrast, the archaeal domain is characterized by the large diversity of cell walls it harbors. The most encountered cell wall is a paracrystalline protein surface layer (S-layer). Moreover, the cell wall of some euryarchaeal species contains a rigid polymer mainly composed of polysaccharides, like glutaminylglycan or methanochondroitin. Interestingly, some methanogenic archaea do possess a polymer structurally similar to PG, which was named pseudomurein (PM). Although archaea are mostly monoderm (i.e., are surrounded by only one membrane), some species exhibit a diderm (i.e., two membranes) cell wall (Albers and Meyer 2011; Klingl et al. 2019; Meyer and Albers 2020).

1.2.2.1. Cell wall-less archaea

In prokaryotes, the cell wall acts as a protective layer against the external environment. However, *Thermoplasma*, a class of Euryarchaeota, are thermoacidophilic organisms that lack a cell wall (Golyshina and Timmis 2005). In order to survive in extreme environments (i.e., 60°C, pH 1-2) without a cell wall, those species have adapted their cytoplasmic membrane (Klingl et al. 2019; Meyer and Albers 2020). Hence, in the cytoplasmic membrane of *Thermoplasma acidophilum* are anchored glycoproteins and lipoglycan mainly built of mannose residues. This protective coat is named glycocalyx (Langworthy et al. 1972; Yang and Haug 1979; Klingl et al. 2019; Meyer and Albers 2020). In addition, it has been reported that *Ferroplasma acidarmanus* has a monolayer CM formed by tetraether lipids (Fig. 17b) instead of a bilayer. Hence, the high resistance to acid hydrolysis of the monolayer CM enables *Thermoplasma* cells to live in acidic environments (Macalady et al. 2004; Klingl et al. 2019).

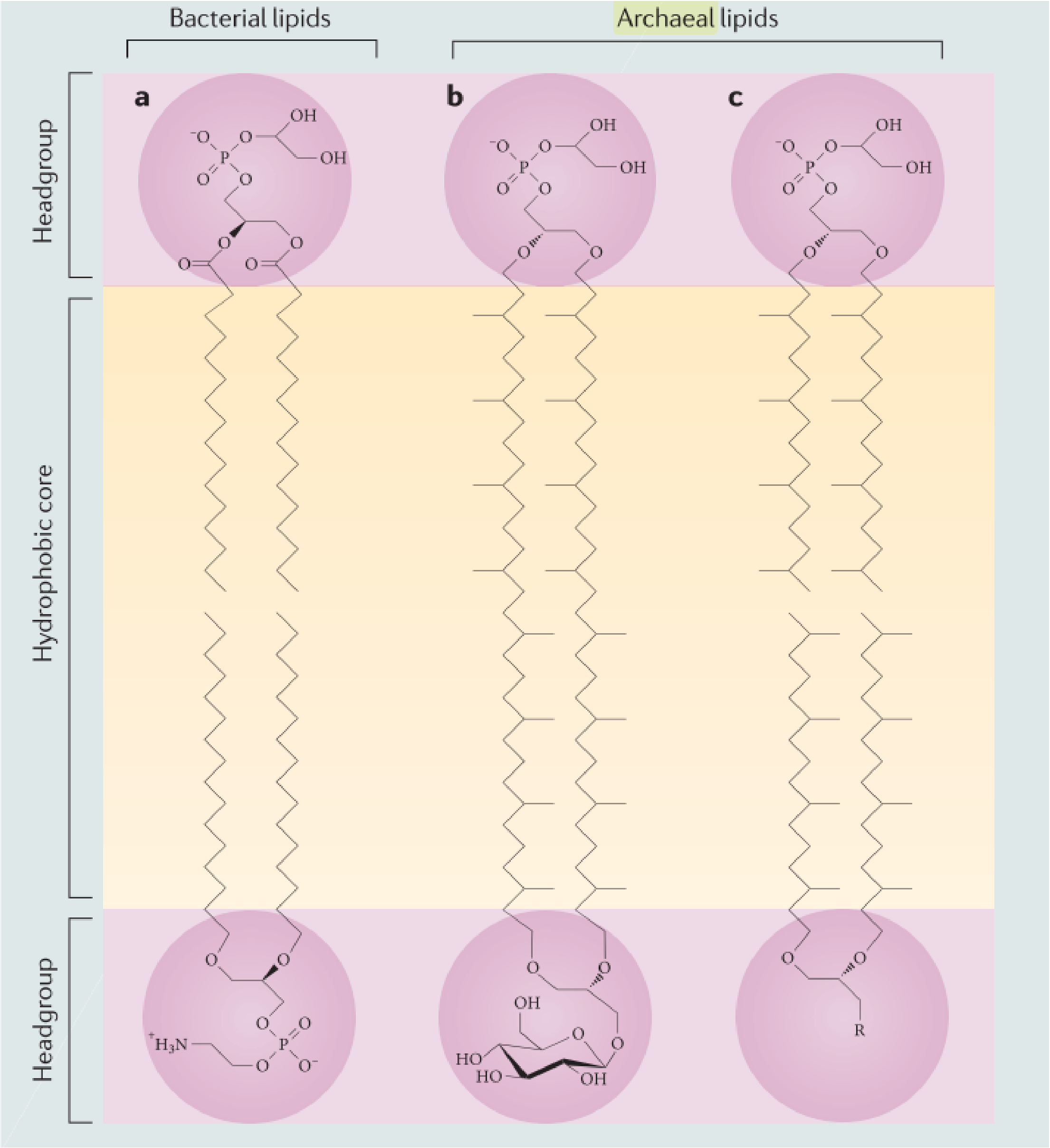


Figure 17. Phospholipids composing the cytoplasmic membrane (CM) of Bacteria and Archaea (from Albers and Meyer 2011). The phospholipids composing the CM of Bacteria and Archaea are fundamentally different. In Bacteria, fatty acids are linked to the glycerol-3-phosphate via an ester bond, whereas in Archaea, isoprenoids are linked to the glycerol-1-phosphate through an ether bond. (**a**) The common bilayer-forming lipids in Bacteria are phosphatidylglycerol (upper lipid) and phosphatidylethanolamine (lower lipid). (**b**) The monolayer-forming tetraether lipids of *T. acidophilum*. (**c**) Representation of the bilayered-forming diether lipids found in Archaea.

1.2.2.2. Diderm archaea

In early 2000s, the crenarchaeon *Ignicoccus hospitalis* was the first archaea with a second (i.e., outer) membrane to be discovered (Rachel et al. 2002). Since then, many other diderm archaea have been identified in different phyla, such as ARMAN archaea (**a**rchaeal **R**ichmond **M**ine **a**cidophilic **n**anoorganisms) (Baker et al. 2006; Comolli et al. 2009) or *Methanomassiliicoccus luminyensis*, isolated from human feces (Dridi et al. 2012). In *I. hospitalis*, the space between the two membranes is called pseudo-periplasm, and can compose up to 40% of the cell volume. Moreover, the distance between the two membranes can be up to 500 nm (Heimerl et al. 2017). No cell wall polymer has been detected so far in double membraned archaea, in contrast to diderm bacteria, in which a PG layer is sandwiched between the two membranes (Klingl et al. 2019; Meyer and Albers 2020).

1.2.2.3. S-layer

The S-layer is the most simple and widespread type of archaeal cell wall (Albers and Meyer 2011; Klingl et al. 2019; Meyer and Albers 2020). It is usually composed of one or, sometimes two, (glyco-)proteins, which self-assemble into a 2-dimensional paracrystalline layer. Depending on the species, the lattice unit can have an oblique (p1 or p2), square (p4) or hexagonal (p3 or p6) symmetry. Therefore, these units are composed of one to six identical proteins, which leave regularly spaced pores identical in shape and size (Fig. 18). S-layer proteins can also undergo either N-glycosylation or O-glycosylation, usually on Asp, Ser or Thr residues (Sleytr et al. 2014; Rodrigues-Oliveira et al. 2017). Interestingly, it has been shown that many

Thermococcales species are surrounded by two S-layers (Rodrigues-Oliveira et al. 2017; Klingl et al. 2019; Meyer and Albers 2020). Although more patchily distributed, S-layers have also been described and characterized in some bacterial species (Fagan and Fairweather 2014).

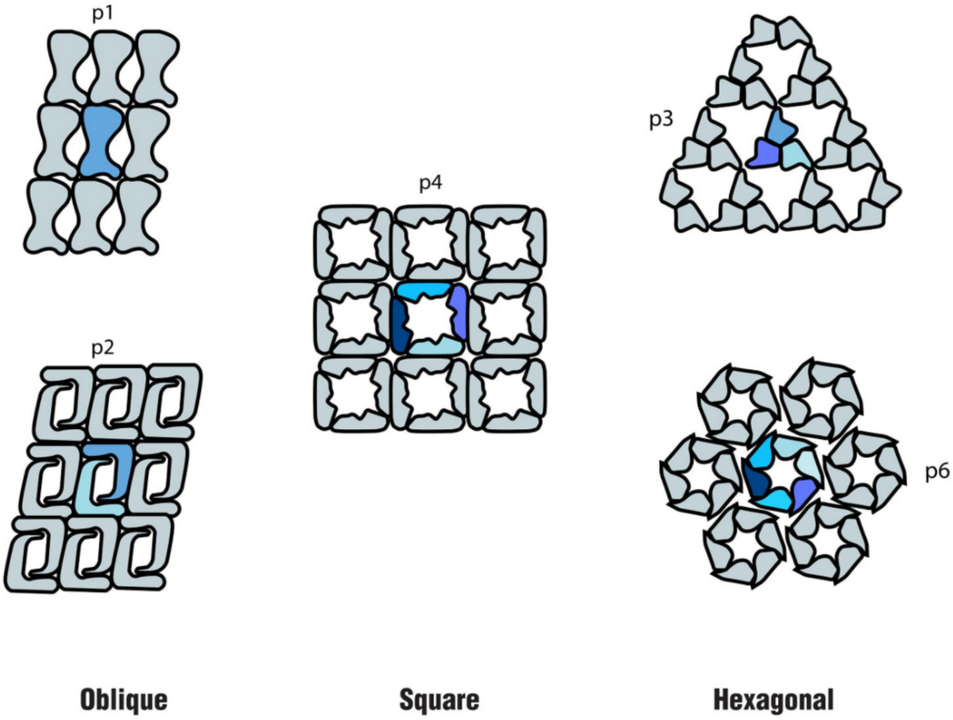


Figure 18. Schematic representation of different S-layer lattice units (from Rodrigues-Oliveira et al. 2017). The oblique (p1, p2), square (p4) and hexagonal (p3, p6) symmetries.

1.2.2.4. Halomucin

Haloquadratum walsbyi is an unusual square-shaped halophilic euryarchaeon, which possesses a double S-layer cell wall. According to genomic data, it was proposed that *H. walsbyi* cells are additionally surrounded by a poly-γ-glutamate capsule. Indeed, the genome of *H. walsbyi* codes for homologous proteins of the CapBCA complex, which synthesizes the poly-γ-glutamate capsule in some Bacilli species (Hsueh et al. 2017). In addition, the cells are surrounded by a very large glycoprotein (more than 1000 KDa) called halomucin, owing to its similarity to mammalian mucin. It has been shown that halomucin does not entirely surround cells, but rather is loosely associated with them (Zenke et al. 2015; Klingl et al. 2019; Meyer and Albers 2020).

1.2.2.5. Cell-wall polymers

In contrast to bacteria, no known archaea does possess PG. Moreover, archaeal species do not share a universal polymer in their cell wall, like PG for bacteria. However, different cell-wall polymers, mainly composed of glycan and AAs, are actually found in specific Euryarchaeota lineages. Hence, methanochondroitin is found in Methanosarcina, glutaminylglycan and sulfated heteropolysaccharides are found in Halobacteria, while PM is found in two different classes or Euryarchaeota, Methanopyri and Methanobacteria. In some species, these polymers can be additionally surrounded by an S-layer (Albers and Meyer 2011; Klingl et al. 2019; Meyer and Albers 2020).

1.2.2.5.1. Methanochondroitin

Methanosarcina is a class of Euryarchaeota. Along with the Methanomicrobia, they belong to a monophyletic group named class II methanogens (CIIM) (Bapteste et al. 2005). Methanosarcina cells often form a cubic aggregate of four cells named sarcina. This aggregate is surrounded by a fibrillar polymer that maintains the structure. This polymer was named methanochondroitin due to its similarity with eukaryotic chondroitin sulfate. However, methanochondroitin lacks sulfate residues (Klingl et al. 2019; Meyer and Albers 2020). The methanochondroitin is formed by repeated trisaccharide units composed of one glucuronic acid and two N-acetylgalactosamines (Kreisl and Kandler 1986). In addition, some Methanosarcina species possess a S-layer between the CM and the methanochondroitin layer (Francoleon et al. 2009; Arbing et al. 2012).

1.2.2.5.2. Glutaminylglycan

The glutaminylglycan polymer has been described from the cell wall of *Natronococcus occultus*. It is similar to the poly-γ-glutamate capsule found in some Bacilli. In the archaeal version of this polymer, the poly-γ-glutamate backbone contains only L-Glu (instead of the mix of L- and D-Glu in bacteria), which are moreover glycosylated. The glycosylation consists of two oligosaccharides composed of about 60 monomers, which are linked via the γ–carboxylic group of the L-Glu residues. The first oligosaccharide is composed of GlcNAc and galacturonic acid, while the second is composed of N-acetyl-d-galactosamine and glucose (Niemetz et al. 1997; Klingl et al. 2019; Meyer and Albers 2020).

1.2.2.5.3. Heteropolysaccharides

The cell wall of *Halococcus morrhuae* and *Halococcus salifodinae* contains highly sulfated heteropolysaccharides composed of glucose, mannose, galactose, glucuronic acid, galacturonic acid, glucosamine, and galosamunronic acid with different molar ratios depending on the species. In addition, there are also N-acetylated amino sugars (Klingl et al. 2019; Meyer and Albers 2020). Moreover, it has been suggested that glucosamine units are linked to uronic residues through glycine bridges (Steber and Schleifer 1979; Klingl et al. 2019; Meyer and Albers 2020).

1.2.2.5.4. Pseudomurein

Methanopyri and Methanobacteria are two other classes of methanogenic euryarchaeota, which form with Methanococci the monophyletic group of class I methanogens (CIM) (Bapteste et al. 2005; Williams et al. 2020). Cells of Methanopyri and Methanobacteria are surrounded by a polymer that shows an architecture similar to the bacterial PG (murein), hence its name of pseudomurein (PM) (Albers and Meyer 2011; Klingl et al. 2019; Meyer and Albers 2020). In contrast to PG, the PM disaccharide unit is composed of N-acetyl-L-talosaminuronic acid (NAT) linked to GlcNAc through a β-(1→3) bond instead of the MurNAc-β-(1→3)-GlcNAc (Fig. 19A). Furthermore, the archaeal stem peptide attached to the carboxyl group of NAT contains only L-AAs. In most cases, this stem peptide is composed of two L-Glu, two L-Ala and one L-Lys (Fig. 19B) (Formanek 1985).

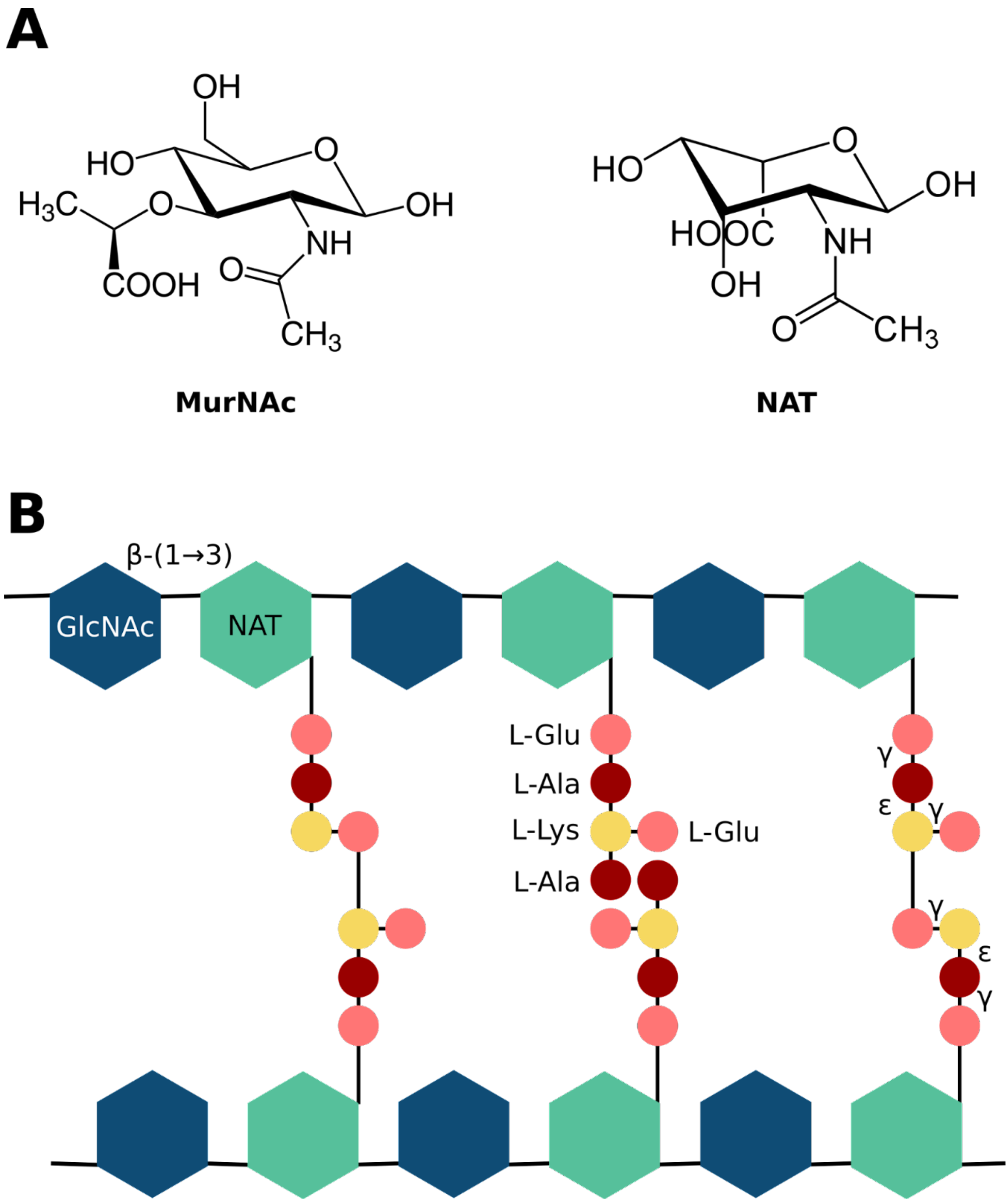


Figure 19. Structure of the archaeal pseudomurein. (**A**) Comparison of the structures of the N-acetylmuramic acid (MurNAc) and N-acetyl-L-talosaminuronic acid (NAT). (**B**) The glycosidic chain of pseudomurein is composed of alternating NAT and N-acetylglucosamine (GlcNAc) units linked by a β-(1→3) bond. To NAT is attached a pentapeptide composed of L-Glu, L-Ala and L-Lys rich in ε- and γ-peptide bonds. In contrast to its bacterial counterpart, pseudomurein has only D-AAs.

The different steps of PM synthesis are not completely resolved. In addition, the genes involved in this synthesis have not been isolated. However, during the 1990s, a pathway for PM biosynthesis (Fig. 20) was proposed by Evamarie Hartmann, Helmut König and Uwe Kärcher on the basis of precursors isolated from cell extracts. The synthesis starts in the cytoplasm with a three-step reaction, which converts the L-Glu into N^{α}-UDP-glutamyl-γ-phosphate (N^{α}-UDP-Glu^{γ}-P). Then, ATP-dependent successive reactions add to the N^{α}-UDP-Glu^{γ}-P, L-Ala, L-Lys and a second L-Ala, to yield N^{α}-UDP-Glu^{γ}-Ala-$^{\varepsilon}$Lys-Ala. A second L-Glu is linked to the L-Lys through a γ bond. In parallel, the UDP-N-acetyl-D-galactosamine is converted into NAT via epimerization and oxidation, and then linked to GlcNAc through a β-(1→3) bond. Finally, the activated pentapeptide is linked to the disaccharide (Hartmann and König 1990; König et al. 1993; Hartmann and König 1994). Similarly to PG, some species show variation in PM composition. For instance, the presence of Asp, Thr, Ser and Orn has been reported (Kandler and Konig 1993; König et al. 1993). Although none of the genes involved in PM synthesis are yet characterized, two PM endopeptidases (PeiW and PeiP) were isolated from a prophage. These two enzymes cleave the ε peptide bond between the L-Ala in position two and the L-Lys in position three (Luo et al. 2001; Luo et al. 2002; Visweswaran et al. 2010; Schofield et al. 2015).

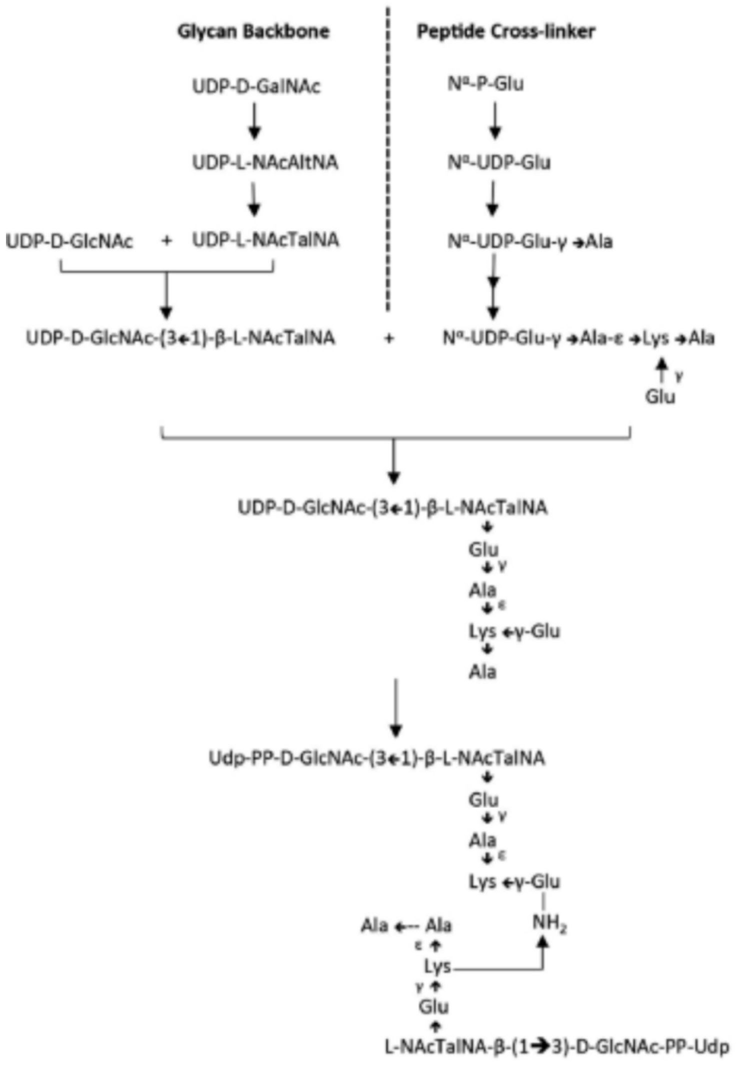

Glycan Backbone
Peptide Cross-linker
UDP-D-GalNAc
UDP-L-NAcAltNA
UDP-D-GlcNAc + UDP-L-NAcTalNA
Nα-P-Glu
Nα-UDP-Glu
Nα-UDP-Glu-γ →Ala
UDP-D-GlcNAc-(3←1)-β-L-NAcTalNA +
Nα-UDP-Glu-γ →Ala-ε →Lys →Ala
γ
Glu
UDP-D-GlcNAc-(3←1)-β-L-NAcTalNA
Glu
γ
Ala
ε
Lys ←γ-Glu
Ala
Udp-PP-D-GlcNAc-(3←1)-β-L-NAcTalNA
Glu
γ
Ala
ε
Lys ←γ-Glu
NH2
Ala ←- Ala
ε
Lys
γ
Glu
L-NAcTalNA-β-(1→3)-D-GlcNAc-PP-Udp

Figure 20. Pathway for pseudomurein biosynthesis proposed by Evamarie Hartmann, Helmut König and Uwe Kärche (adapted from Klingl et al. 2019). In this figure are depicted the three steps proposed for pseudomurein biosynthesis. The disaccharide and the pentapeptide are synthesized in parallel. Then, the pentapeptide is linked to the disaccharide unit. The last step corresponds to the polymerisation of the polymer outside the cytoplasmic membrane. L-NAcTalNA = N-acetyl-L-talosaminuronic acid, GlcNAc = N-acetylglucosamine.

Due to the difference between PG and PM biosynthesis pathways, it was concluded that the origins of the two polymers were unrelated and that their similarity is only due to convergent evolution (König et al. 1993; Scheffers and Pinho 2005; Albers and Meyer 2011). However, recent genomic data have highlighted homologues of genes involved in PG synthesis, such as muramyl ligases, in Methanopyri and

Methanobacteria (Smith et al. 1997; Slesarev et al. 2002; Samuel et al. 2007; Leahy et al. 2010). Therefore, some scientists now suggest that PM could have evolved from HGT of PG genes from Bacteria (Graham and Huse 2008; Subedi et al. 2021; Ithurbide et al. 2022).

1.3. Bioinformatic databases

Deciphering the evolution of prokaryotes, including how interdomain HGT has impacted their evolution, requires tremendous amounts of genomic data. As aforementioned, sequences have been accumulating at an exponential rate for decades in public repositories. Owing to this trend, the sampling of prokaryotic genomes is now so broad and deep that many important evolutionary questions can be tackled by bioinformatic mining of genomic sequence data, especially those making use of phylogenetic inference.

1.3.1. Open scientific data

Scientific data can be defined as a collection of information, e.g., observations, facts or results, which are used as evidence of phenomena for the purposes of research or scholarship (Leonelli 2015; Pasquetto et al. 2017). The term “open data” refers to data that is freely accessible and that anyone can (re-)use, modify or share without any restrictions from copyright or patents (Murray-Rust 2008). The ability to access and use open data is particularly crucial for progress in science. Indeed, this allows researchers to combine data from different sources to address new questions and make new advancements that benefit mankind (Murray-Rust 2008; Y. Demchenko et al. 2012; Leonelli 2015; Pasquetto et al. 2017). Moreover, the reuse of data is essential to validate and confirm studies for the sake of reproducibility. To this end, it is recommended that the format of released data meets standard requirements established by the scientific community (Y. Demchenko et al. 2012; Pasquetto et al. 2017). In order to be accessible to the community, scientific data is stored in databases from public repositories, which are often managed by national or international entities.

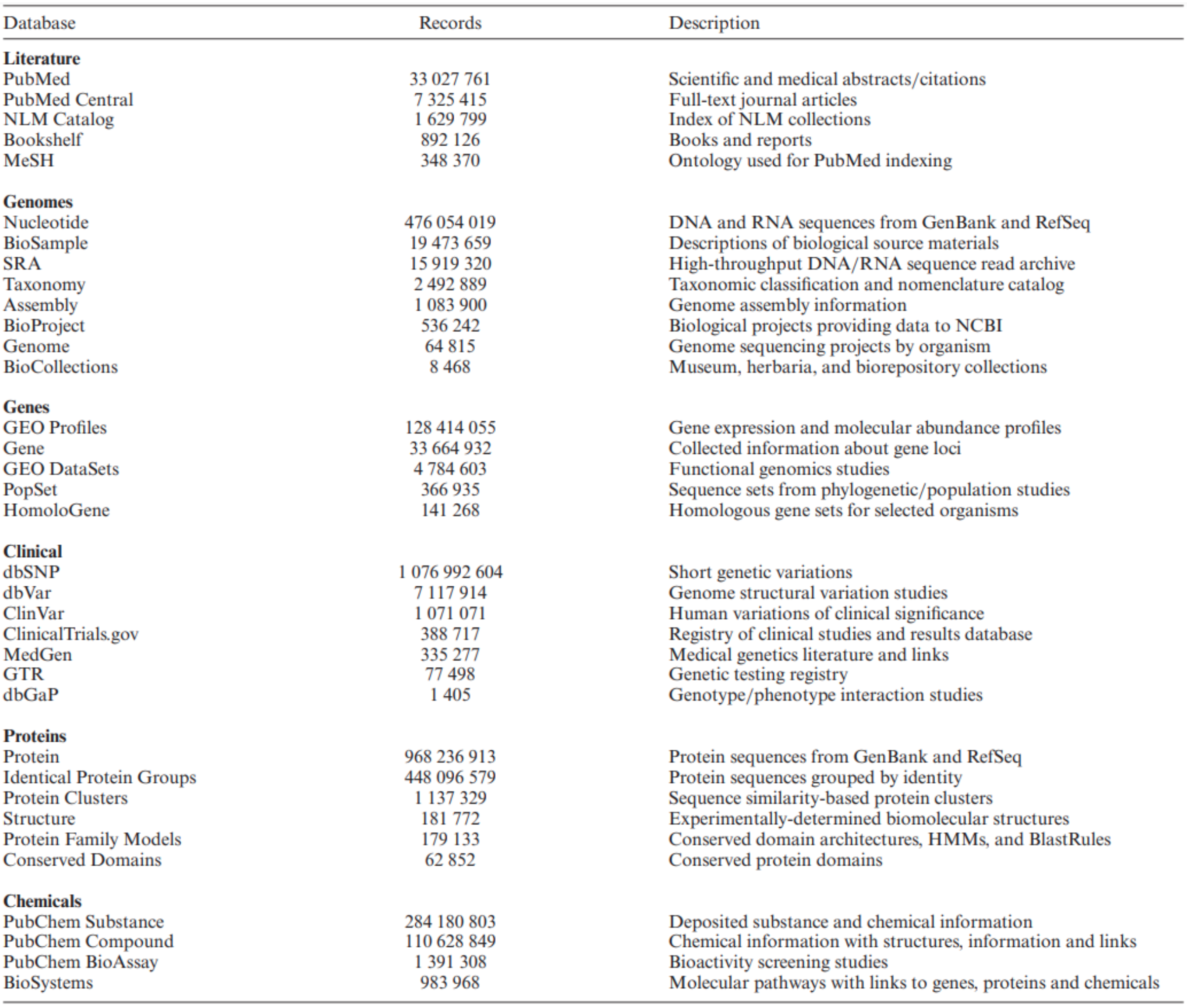
1.3.2. Multi-database infrastructures

The two major and well-known multi-database resources in life science are the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EMBL-EBI). The purpose of these entities is to centralize and store knowledge from medical, molecular biology, biochemistry, and genetics in databases that are freely accessible to scientific communities and the general public. They also provide bioinformatic services, such as the development and distribution of software tools for analyzing molecular and genomic data (https://www.ncbi.nlm.nih.gov/home/about/mission/ Accessed 30 November 2022; https://www.ebi.ac.uk/about Accessed 30 November 2022).

1.3.2.1. The NCBI

The NCBI is a division of the National Library of Medicine (NLM) at the U.S National Institutes of Health (NIH), based in Bethesda, Maryland. The initiative to create the NCBI began between 1984 and 1986, when groups of scientists convened meetings with legislators at Capitol Hill, Washington, US, to advocate and promote the financing of genomic research. The NCBI was finally created on November 4th, 1988. Since then, it has become one of the leading institutions for research in computational biology. The NCBI is notably involved in the development of the famous BLAST algorithm used for database search based on sequence (i.e., AA, DNA, RNA) comparisons (Sayers et al. 2011). In September 2021, the NCBI manages and maintains 35 databases (Table 1) containing a total of 3.6 billion records, which are divided into six categories: literature, genomes, genes, clinical, proteins and chemicals (Sayers et al. 2022).

Table 1. NCBI databases as of 4 September 2021 (from Sayers et al. 2022).



Database	Records	Description
Literature		
PubMed	33 027 761	Scientific and medical abstracts/citations
PubMed Central	7 325 415	Full-text journal articles
NLM Catalog	1 629 799	Index of NLM collections
Bookshelf	892 126	Books and reports
MeSH	348 370	Ontology used for PubMed indexing
Genomes		
Nucleotide	476 054 019	DNA and RNA sequences from GenBank and RefSeq
BioSample	19 473 659	Descriptions of biological source materials
SRA	15 919 320	High-throughput DNA/RNA sequence read archive
Taxonomy	2 492 889	Taxonomic classification and nomenclature catalog
Assembly	1 083 900	Genome assembly information
BioProject	536 242	Biological projects providing data to NCBI
Genome	64 815	Genome sequencing projects by organism
BioCollections	8 468	Museum, herbaria, and biorepository collections
Genes		
GEO Profiles	128 414 055	Gene expression and molecular abundance profiles
Gene	33 664 932	Collected information about gene loci
GEO DataSets	4 784 603	Functional genomics studies
PopSet	366 935	Sequence sets from phylogenetic/population studies
HomoloGene	141 268	Homologous gene sets for selected organisms
Clinical		
dbSNP	1 076 992 604	Short genetic variations
dbVar	7 117 914	Genome structural variation studies
ClinVar	1 071 071	Human variations of clinical significance
ClinicalTrials.gov	388 717	Registry of clinical studies and results database
MedGen	335 277	Medical genetics literature and links
GTR	77 498	Genetic testing registry
dbGaP	1 405	Genotype/phenotype interaction studies
Proteins		
Protein	968 236 913	Protein sequences from GenBank and RefSeq
Identical Protein Groups	448 096 579	Protein sequences grouped by identity
Protein Clusters	1 137 329	Sequence similarity-based protein clusters
Structure	181 772	Experimentally-determined biomolecular structures
Protein Family Models	179 133	Conserved domain architectures, HMMs, and BlastRules
Conserved Domains	62 852	Conserved protein domains
Chemicals		
PubChem Substance	284 180 803	Deposited substance and chemical information
PubChem Compound	110 628 849	Chemical information with structures, information and links
PubChem BioAssay	1 391 308	Bioactivity screening studies
BioSystems	983 968	Molecular pathways with links to genes, proteins and chemicals

1.3.2.2. The EMBL-EBI

The EMBL-EBI is a part of the European Molecular Biology Laboratory (EMBL), a non-profit intergovernmental organization supported by 27 countries (https://www.embl.org/about/history/ Accessed 30 November 2022). Historically, the EMBL established (in 1980) the Nucleotide Sequence Data Library database (now called the European Nucleotide Archive (ENA) (Harrison et al. 2021; Cummins et al. 2022) in Heidelberg, Germany, to centralize DNA sequences that were formerly only submitted to scientific journals. The need to maintain and manage the exponentially growing sequence database led to the creation of the EMBL-EBI in 1992, which was established in Hinxton, UK. The ENA and the UniProt (The UniProt Consortium 2021) were the two first databases of EMBL-EBI (https://www.ebi.ac.uk/history Accessed 30 November 2022). In 2020, the EMBL-EBI manages data from over 40 resources covering different branches of molecular biology (Fig. 21) (Cantelli et al.

2021; Cantelli et al. 2022). The EMBL-EBI also offers web tools such as InterProScan (Jones et al. 2014) or the HMMER web server (Potter et al. 2018) in collaboration with the algorithm developers (Mistry et al. 2013). HMMER is a software package written by Sean R. Eddy, which is used for sequence homology searches based on hidden Markov models (Eddy 2009), whereas InterProScan is the scanning algorithm of the InterPro database, which combines 13 protein signature databases: CATH-Gene3D, the Conserved Domains Database (CDD), HAMAP, PANTHER, Pfam, PIRSF, PRINTS, PROSITE Patterns, PROSITE Profiles, SMART, the Structure–Function Linkage Database (SFLD), SUPERFAMILY and TIGRFAMs (Paysan-Lafosse et al. 2022). InterProScan relies on multiple algorithms, such as HMMER (Mistry et al. 2013) or BLAST (Camacho et al. 2009), to predict the presence of functional domains and sites in an uncharacterized protein sequence (Jones et al. 2014).

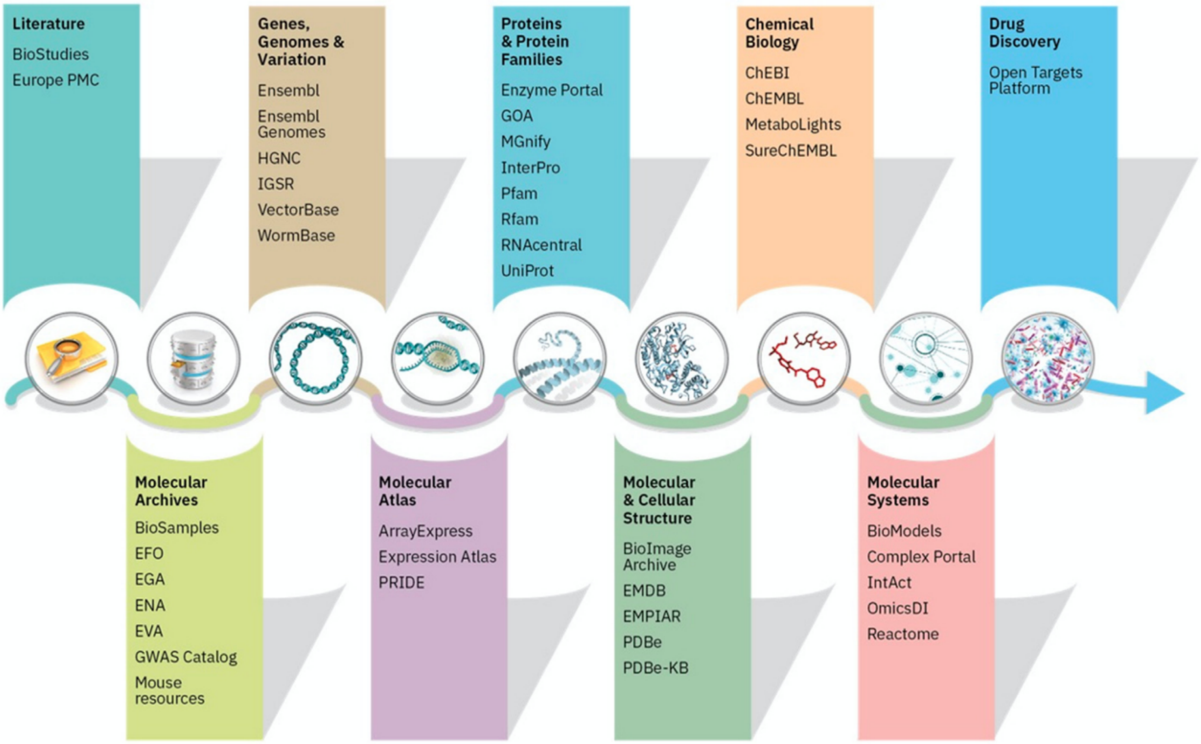


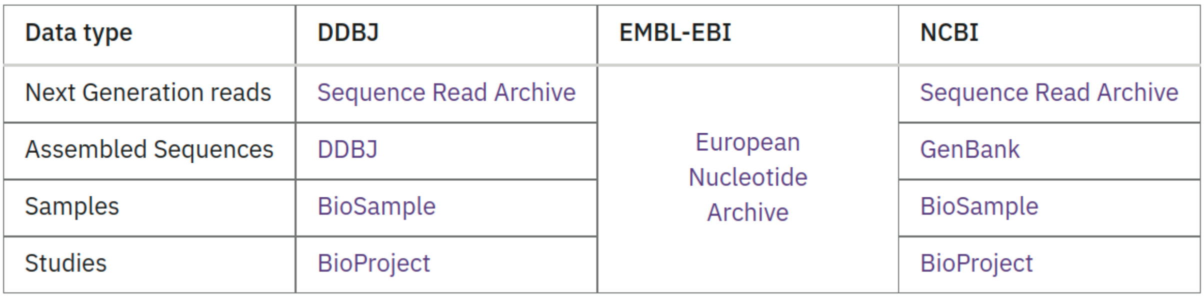
Figure 21. Summary of all EMBL-EBI Data Resources as of September 2020 (from Cantelli et al. 2021).

1.3.3. Data submission and collaboration

Primary biological data can be directly submitted by researchers through the submission portal of NCBI (https://www.ncbi.nlm.nih.gov/home/submit.shtml) or EMBL-EBI (https://www.ebi.ac.uk/submission/). To simplify the submission, both entities have implemented submission wizards to help to select the right archive to deposit new data. NCBI and EMBL-EBI servers can also receive data from national and international collaborations or research consortia (Arita et al. 2021; Cantelli et al. 2022; Sayers et al. 2022).

Although NCBI and EMBL-EBI were initially founded for different purposes (https://www.ncbi.nlm.nih.gov/books/NBK148949/ Accessed 30 November 2022; https://www.ebi.ac.uk/history Accessed 30 November 2022), one of their common objectives is to store and manage data, particularly nucleotide and protein sequences. Actually, sequence data is the most abundant kind of data in NCBI and EMBL-EBI repositories (Cantelli et al. 2022; Sayers et al. 2022). Different types of nucleotide data can be submitted to these repositories, such as individual sequences, batches of sequences, and even whole genomes, among which plasmid, viral and organelle genomes (Benson et al. 2009; Choudhuri 2014). An ‘individual sequence’ corresponds to, e.g., a coding region with its corresponding protein translation, a pseudogene, a RNA molecule, whereas a ‘batch of sequences’ can refer to, e.g., raw sequencing read data (Leinonen et al. 2011) or assembled transcriptomic data (Transcriptome shotgun assembly (TSA)) or expressed sequence tags (ESTs) (Benson et al. 2013; Benson et al. 2018). Submitted genomes can be completely assembled and gapless (https://www.ncbi.nlm.nih.gov/genbank/genomesubmit/ Accessed 30 November 2022), incomplete as sets of multiple contigs (Whole Genome Shotgun (WGS)) (https://www.ncbi.nlm.nih.gov/genbank/wgs/ Accessed 30 November 2022) or be so-called metagenome-assembled genomes (MAGs) (https://www.ncbi.nlm.nih.gov/genbank/metagenome/ Accessed 30 November 2022). During submission, submitters are encouraged to introduce metadata, including feature annotations (https://submit.ncbi.nlm.nih.gov/about/bankit/ Accessed 30 November 2022) and cross-references to other types of records (e.g., BioSample, BioProject) (Barrett et al. 2012; Gostev et al. 2012), along with their nucleotide sequences. At the EMBL-EBI repository, all nucleotide sequences and companion metadata are stored in the ENA database (Arita et al. 2021; Harrison et al. 2021), whereas most (but not all) of the nucleotide data today is also apparently included in the NCBI GenBank database (Table 2) (Arita et al. 2021; Sayers et al. 2022).

Table 2. List of the databases from DDBJ, EMBL-EBI and NCBI that compose the records of the INSDC (from www.insdc.org).



Data type	DDBJ	EMBL-EBI	NCBI
Next Generation reads	Sequence Read Archive	European Nucleotide Archive	Sequence Read Archive
Assembled Sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

Databases are dynamic structures that change and complexify over the years in response to the increasing amount of released data (Nadim 2016). Furthermore, public repositories exhibit a lot of redundancy (Pruitt et al. 2005; Chen et al. 2017). For instance, the genomes uploaded to the NCBI GenBank database are also included in its Assembly database (Kitts et al. 2016). Consequently, it is not easy to exactly determine which data is present in a specific database, notably for the NCBI repository (Fig. 22). Historically, GenBank was created in 1979 at the Los Alamos National Laboratory under the name of Los Alamos Sequence Database. In 1982, the database became public and was renamed to GenBank. The GenBank database has been under the responsibility of the NCBI since October 1992 (Benson et al. 1993; Choudhuri 2014). Therefore, the NCBI formerly contained only three nodes in its system: individual nucleotide sequences and protein protein translations included in GenBank, and the associated literature included in MEDLINE (now known as PubMed) (Benson et al. 1990; Schuler et al. 1996; Mrozek et al. 2013). Since late 1980s, the GenBank and ENA databases, along with the DNA Data Bank of Japan (DDBJ) from the National Institute of Genetics (NIG) (Okido et al. 2022), are part of the International Nucleotide Sequence Database Collaboration (INSDC). This collaboration has the purpose to share and standardize the format and the annotation of nucleotide (and protein) sequence data and their subsidiary metadata. Daily exchange between the three multi-database resource partners (i.e., NCBI, EMBL-EBI and NIG) warrants worldwide coverage of this infrastructure (Burks et al. 1985; Karsch-Mizrachi et al. 2012; Arita et al. 2021). To ensure the disponibility of newly added sequences from submitters and from the INSDC, a new version of GenBank is released every two months (https://www.ncbi.nlm.nih.gov/genbank/ Accessed 30 November 2022). In the mid 2000s, the INSDC established the Sequence Read Archive (SRA) for storage of next-generation sequencing (NGS) raw read data (Leinonen et al. 2011; Kodama et al. 2012; Katz et al. 2022). As suggested by Table 2, SRA records are not included in GenBank, which further supports the idea mentioned above that GenBank does not include all nucleotide data.

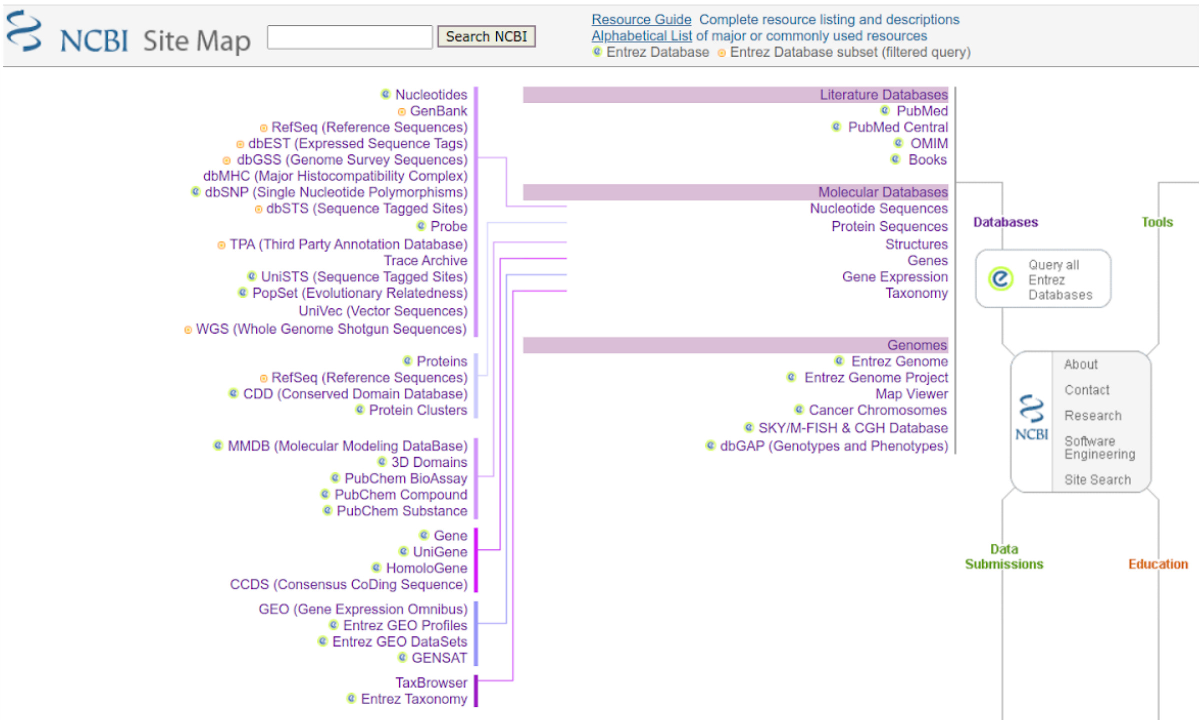

NCBI Site Map
Search NCBI
Resource Guide Complete resource listing and descriptions
Alphabetical List of major or commonly used resources
Entrez Database
Entrez Database subset (filtered query)
Nucleotides
GenBank
RefSeq (Reference Sequences)
dbEST (Expressed Sequence Tags)
dbGSS (Genome Survey Sequences)
dbMHC (Major Histocompatibility Complex)
dbSNP (Single Nucleotide Polymorphisms)
dbSTS (Sequence Tagged Sites)
Probe
TPA (Third Party Annotation Database)
Trace Archive
UniSTS (Sequence Tagged Sites)
PopSet (Evolutionary Relatedness)
UniVec (Vector Sequences)
WGS (Whole Genome Shotgun Sequences)
Proteins
RefSeq (Reference Sequences)
CDD (Conserved Domain Database)
Protein Clusters
MMDB (Molecular Modeling DataBase)
3D Domains
PubChem BioAssay
PubChem Compound
PubChem Substance
Gene
UniGene
HomoloGene
CCDS (Consensus CoDing Sequence)
GEO (Gene Expression Omnibus)
Entrez GEO Profiles
Entrez GEO DataSets
GENSAT
TaxBrowser
Entrez Taxonomy
Literature Databases
PubMed
PubMed Central
OMIM
Books
Molecular Databases
Nucleotide Sequences
Protein Sequences
Structures
Genes
Gene Expression
Taxonomy
Genomes
Entrez Genome
Entrez Genome Project
Map Viewer
Cancer Chromosomes
SKY/M-FISH & CGH Database
dbGAP (Genotypes and Phenotypes)
Databases
Tools
Query all Entrez Databases
NCBI
About
Contact
Research
Software Engineering
Site Search
Data Submissions
Education

Figure 22. Screenshot from NCBI Site Map (www.ncbi.nlm.nih.gov/Sitemap/index.html) showing the complexity of the NCBI Entrez databases. Entrez is the text-based search and retrieval system used by the NCBI.

This superposition of different types of nucleotide sequence data and the collaboration over the years illustrates the dynamic and the complexity of public repositories. For instance, GenBank was the former nucleotide database of the NCBI (Benson et al. 1990; Schuler et al. 1996). Now, it is included in a higher structure, commonly named ‘Nucleotide database’, which notably contains all sequences of

GenBank, RefSeq (discussed in the next section), INSDC and SRA (https://www.ncbi.nlm.nih.gov/nucleotide/ Accessed 30 November 2022).

1.3.4. Standardization and curation

In April 1999, the NCBI launched the Reference Sequence (RefSeq) project, which aims to provide users with a collection of non-redundant genomic DNA, transcript and protein sequences (Maglott et al. 2000; Pruitt et al. 2005; Haft et al. 2018). RefSeq records are based on sequences submitted to INSDC, which undergo a quality assurance procedure to ensure sequence quality, completeness and the absence of contamination (O’Leary, Wright, Brister, Ciufo, Pruitt, et al. 2016; Haft et al. 2018; Li et al. 2021). To reduce redundancy, for bacterial and archaeal genomes, the NCBI has implemented the Prokaryotic Genome Annotation Pipeline (PGAP), which is used to generate structural and functional annotation of genome records from the INSDC (O’Leary et al. 2016). This pipeline uses prediction methods to identify protein-coding and RNA genes associated with sequence alignments (e.g., HMM profiles) and manual curation to assign annotation. Furthermore, the pipeline is frequently updated to improve annotation and standard quality of the genomes (O’Leary, Wright, Brister, Ciufo, Pruitt, et al. 2016; Haft et al. 2018; Li et al. 2021). Complete and WGS genomes uploaded on GenBank can be annotated with PGAP during submission. Genomes that do not meet annotation and sequence quality thresholds are not included in the RefSeq database (O’Leary, Wright, Brister, Ciufo, Haddad, et al. 2016) The minimum standard annotations required for a prokaryotic genome are: 1) at least one copy of each structural RNA (i.e., 5S, 16S, 23S), 2) at least one copy of each tRNA, 3) a ratio of protein-coding genes to genome length close to 1 and 4) no gene completely contained in another gene (Klimke et al. 2011). All criteria that exclude a genome from RefSeq are summarized through this following link: https://www.ncbi.nlm.nih.gov/assembly/help/anomnotrefseq/. As previously explained, a genome uploaded on GenBank is also included in the Assembly database. If this GenBank genome is further selectionned to be in the RefSeq database, the RefSeq version of the genome is also copied in the Assembly database. Both genomes have an identical ID number but a GenBank genome is prefixed by ‘GCA_’ while the RefSeq genome has ‘GCF_’ (Kitts et al. 2016).

Consequently, this is another source of redundancy in the NCBI repository. In contrast to GenBank, RefSeq records are regularly reannotated with PGAP, leading to the suppression of some low-quality genomes or proteins (Tatusova et al. 2016; Li et al. 2021).

Similar curation efforts have been implemented by EMBL-EBI with the Ensembl project (Howe et al. 2021) and the Swiss-Prot section of the UniProt Knowledgebase (UniProtKB) (Poux et al. 2017). The Ensembl project was launched in 1999 to automatically annotate the human genome (Butler 2000). Since then, it provides high-quality annotated genomes from bacteria, protists, fungi, plant and metazoa (Howe et al. 2021). Genomes of the Ensembl project are notably collected from the INSDC and the ENA database (Zerbino et al. 2018; Howe et al. 2021). The UniProtKB repository provides users with a set of functionally annotated protein sequences (The UniProt Consortium 2021). It is divided into two sections: 1) the reviewed Swiss-Prot entries, which contains manually curated and annotated protein sequences, some of them tracing back to the historical Swiss-Prot database, and 2) the unreviewed TrEMBL entries containing automatically annotated protein sequences (Poux et al. 2017; The UniProt Consortium 2021).

1.3.5. The NCBI Taxonomy database

In bioinformatic studies, especially those involving phylogenetic analyses, it is crucial to link sequence data to the source organisms through an appropriate taxonomy. Although there exist various taxonomic databases (Wang et al. 2007; McDonald et al. 2012; Yilmaz et al. 2014; Balvočiūtė and Huson 2017; Parks et al. 2018; Rinke et al. 2021), the NCBI Taxonomy database (Schoch et al. 2020) remains the main source for taxonomic records (Sakamoto and Ortega 2021). In 1991, the NCBI launched the first version of its Taxonomy project where GenBank nucleotide and protein sequences were linked to the source organism. During this period, there was no consensus for taxonomic classification, thus the three INSDC partners independently maintained their own taxonomic nomenclature. For the sake of consistency, INSDC partners agreed in 1997 to use the NCBI Taxonomy nomenclature as the only source of classification (Federhen 2012). This taxonomic database is manually curated by NCBI scientists following up-to-date primary literature. The NCBI nomenclature follows the rules of four principal codes: 1) the International Code of Nomenclature for algae, fungi and plants (Turland et al. 2018), 2) the International Code of Nomenclature of Prokaryotes (ICNP) (ICNP 2019), 3) the International Code of Zoological Nomenclature (Ride 1999), and 4) the International Code of Virus Classification and Nomenclature (Walker et al. 2019). Recently, the International Committee on Systematics of Prokaryotes (ICSP) added some rules to the ICNP. The resulting new prokaryotic nomenclature was implemented in the NCBI Taxonomy (Oren and Garrity 2021). Briefly, the NCBI Taxonomy corresponds to a single list of taxa from across all domains of life, which are hierarchically arranged. The lineage for a specific organism is usually characterized by seven main taxonomic ranks: superkingdom (= domain), phylum, class, order, family, genus and species. Besides, some additional expanded ranks can be used, e.g., superfamily, subspecies. Perhaps surprisingly, in the NCBI Taxonomy, the well-known ‘kingdom’ rank is only used for Metazoa (animals), Viridiplantae (green plants), Fungi, and high-level groups of viruses (e.g., Pararnavirae, including HIV-1). Each node (taxon) of the taxonomic tree is referred to by an unique TaxId (taxonomy identifier). If there is a duplicated taxon in a lineage (i.e., an identical name used for two different, often successive, ranks), a different TaxId is used (Schoch et al. 2020). An example of such ambiguity is the mosquito genus *Anopheles* (TaxId 7164), a subgenus of which is also called *Anopheles* (TaxId 44482). In a related but somewhat worst case, totally unrelated organisms can share the same genus name because they were described in two distinct codes (hemihomonyms)(Starobogatov 1991). Hence, both *Bacillus atticus atticus* (insect) and *Bacillus subtilis* (bacteria) have the same genus name. However, the TaxId for Bacillus (insect) is ‘55087’ and ‘1386’ for Bacillus (bacteria). Even if TaxIds formally allow researchers to distinguish duplicate taxa, such ambiguities are misleading and even can be dangerous in practice. Consequently, the NCBI Taxonomy is regularly updated, notably to reduce the occurrence of duplicate taxa, even if such changes are disruptive for bioinformatic applications relying on a stable taxonomy, including those seeking to identify contaminant sequences in public genomes.

1.4. Genomic contamination in public databases

Despite the efforts to build clean and accurate reference databases, the exponential release of public prokaryotic genomes often comes with contamination issues (Mukherjee et al. 2015; Steinegger and Salzberg 2020; Orakov et al. 2021). “Genome contamination” means the inclusion of foreign sequences along with the sequences of the genuine organism. This presence of contaminant DNA can lead to false interpretations in comparative genomics (Arakawa 2016; Koutsovoulos et al. 2016) or phylogenomic studies (Schierwater et al. 2009; Finet et al. 2010; Philippe et al. 2011; Laurin-Lemay et al. 2012). Furthermore, these contaminated sequences can spread through databases over time (Breitwieser et al. 2019; Steinegger and Salzberg 2020). The problem of genomic contamination is discussed in detail in the recent review of Luc Cornet and Denis Baurain published in *Genome Biology* (Cornet and Baurain 2022).

1.4.1. Sources of genomic contamination

In this review, the authors have described the different causes that lead to the introduction of foreign DNA sequences during the sequencing process (Fig. 23). These issues can be either **biological** “[..] contamination of an axenic culture by unwanted organism(s) [...] sequencing of chimeric organisms [...] or the presence of plain taxonomic errors in reference databases”, **experimental** “[...] inclusion of unwanted DNA either during DNA extraction or sequencing on shared platforms [...]”, or **computational** during in-silico processing of data “The risk of in-silico contamination is higher when the data comes from metagenomic analyses [...] such data can lead to chimeric sequences by merging similar genomic regions during metagenomic assembly [...] metagenomic binning (i.e., the partition of sequences from the constitutive organisms into individual Metagenome-Assembled Genomes – MAGs) also results in some degree of contamination by lumping in a single MAG contigs reconstructed from different organisms”. The authors further state that genomic contaminations induced by those causes can be classified into redundant and non-redundant contaminations. A contamination is considered as **redundant** when “a genomic segment is present multiple times in a genome assembly, due to inclusion of homologous genomic regions from foreign organism(s)”, whereas it is called **non-redundant** when “an extra genomic segment is present in the assembly”. The latter situation can be divided into two sub-cases “1) a genuine genomic segment is lacking in the target organism (i.e., the completeness is not optimal) and is replaced by a foreign genomic region harbouring (some of) the expected genes or 2) an extra genomic region, for which no homologous region exists in the target organism, is present due to the inclusion of a taxonomically distinct organism (e.g., genomic regions from another kingdom)”.

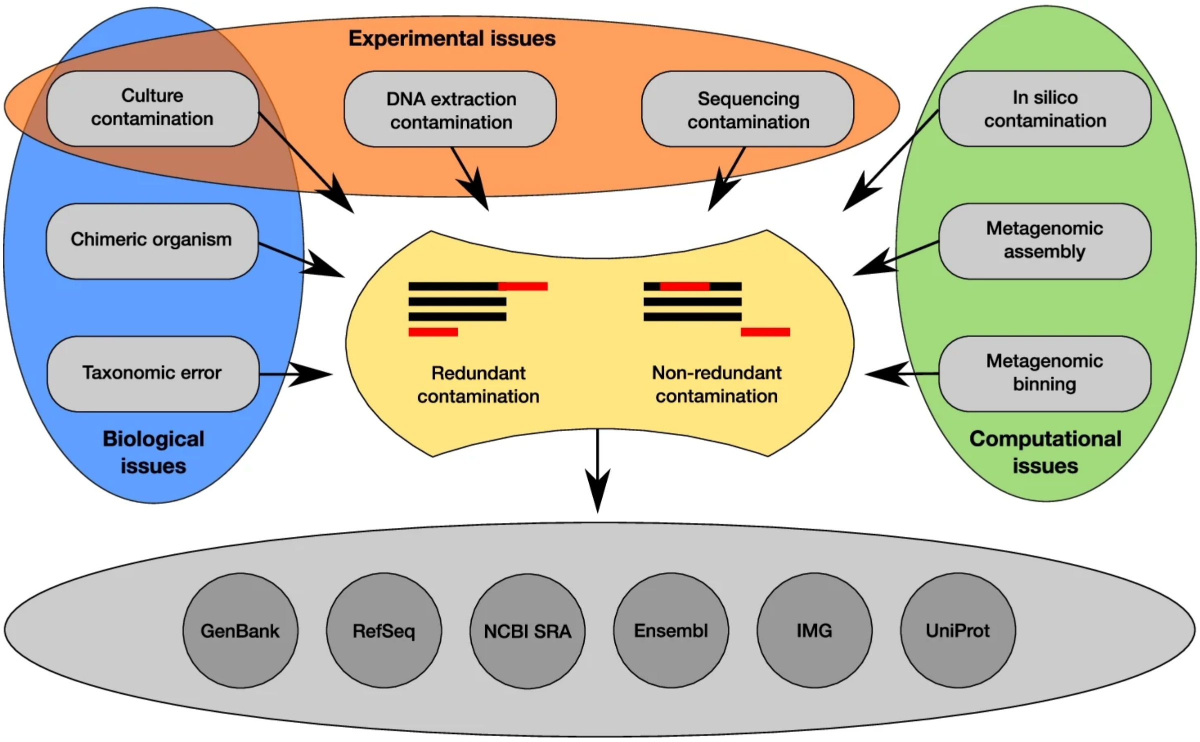


Figure 23. Sources of genomic contamination (from Cornet and Baurain 2022). “Three types of issues lead to contamination of genomic sequence data: biological, experimental and computational. The contamination of “pure” cultures can be due to both experimental (e.g., accidental introduction of contaminating microorganisms) and biological causes (e.g., the presence of an endosymbiont). Redundant contamination occurs when a genomic segment is present multiple times in a genome (e.g., multiple SSU rRNAs from different organisms). Non-redundant contamination occurs when a genomic region of the main organism, the expected one, is replaced by the corresponding region of a foreign organism (e.g., the SSU rRNA of the main organism is replaced by the SSU rRNA from a foreign organism). An extra DNA segment, not part of the main organism but belonging to a contaminant, would also be considered as a non-redundant contamination (e.g.,

eukaryotic DNA in a bacterial genome). A mixed scenario is also possible, as represented in the redundant contamination part of the figure".

1.4.2. Contamination detection algorithms

In order to efficiently detect genomic contamination in public repositories, at least 17 algorithms have been developed during the last years (Fig. 24). Cornet and Baurain classified these tools into two main categories, depending on whether they use a reference database or not. Besides, database-dependent methods can use a genome-wide approach or instead rely on estimators based on gene markers .

1.4.2.1. Database-free algorithms

BlobTools, Anvi'o, ProDeGe and PhylOligo are the four database-free tools (Fig. 24). Their algorithms rely on DNA content to partition sequences in order to detect contamination. Indeed, BlobTools use Guanosine+Cytosine (GC) content for partition, while the other three programs use k-mer (i.e., substrings of DNA sequence between 4 and 9 nt long) frequencies. Furthermore, they (except for PhylOligo) also rely on taxonomy for sequence labeling and program calibration. ProDeGe only works on prokaryotic genomes, while BlobTools, Anvi'o and PhylOligo work on both prokaryotes and eukaryotes (Eren et al. 2015; Koutsovoulos et al. 2016; Tennessen et al. 2016; Mallet et al. 2017; Challis et al. 2020). Cornet and Baurain assert that "Database-free tools can detect both redundant and non-redundant contaminations". However, they claim that "the programs [...] require a case-by-case inspection by the user and are thus difficult to use for large-scale projects".

1.4.2.2. Reference database-dependent algorithms

Out of the 13 database-dependent tools, seven (SINA, ContEst16S, Forty-Two, ConFindR, CheckM, EukCC and BUSCO) rely on highly conserved gene markers to assess the level of redundant and non-redundant (only for Forty-Two) genomic contamination (Fig. 24). Indeed, Cornet and Baurain explain in their review that "These genes are present in a single copy in nearly all organisms and the presence of multiple copies is thus indicative of such type of contamination". SINA and ContEst16S use single-locus SSU rRNA (small subunit ribosomal ribonucleic acid) genes. However, Cornet and Baurain say that “The use of this single locus is not frequent because it entails a higher risk of missing contaminants”. In contrast, the other five tools use multi-locus genes for contamination assessment. Forty-Two and ConFindR use ribosomal proteins while CheckM, EukCC and BUSCO use phylogenetic placement to select lineage-specific gene markers. The latter algorithms, as well as Forty-Two (to some extent), offer the advantage to estimate the completeness of the genomes. ContEst16S, CondFindR and CheckM work on prokaryotes, SINA, Forty-Two and BUSCO on prokaryotes and eukaryotes, whereas EukCC only works for eukaryotic genomes (Pruesse et al. 2012; Parks et al. 2015; Lee et al. 2017; Simion et al. 2017; Low et al. 2019; Saary et al. 2020; Manni et al. 2021).

The last six database-dependent tools (Conterminator, Kraken, CLARK, CONSULT, BASTA, GUNC) compare the entire genome against a reference database. Conterminator, Kraken, CLARK and CONSULT align long k-mers (at least 21 nt) against the database (Wood and Salzberg 2014; Ounit et al. 2015; Wood et al. 2019; Steinegger and Salzberg 2020; Rachtman et al. 2021), while BASTA and GUNC use BLAST (Camacho et al. 2009) or DIAMOND blast (Buchfink et al. 2015) to perform gapped alignment against the database. Like the MEGAN algorithm (Huson et al. 2007), BASTA uses a LCA (Lowest Common Ancestor) labeling approach to classify sequences. GUNC works on prokaryotic genomes, whereas BASTA can handle both prokaryotic and eukaryotic genomes (Kahlke and Ralph 2019; Orakov et al. 2021).

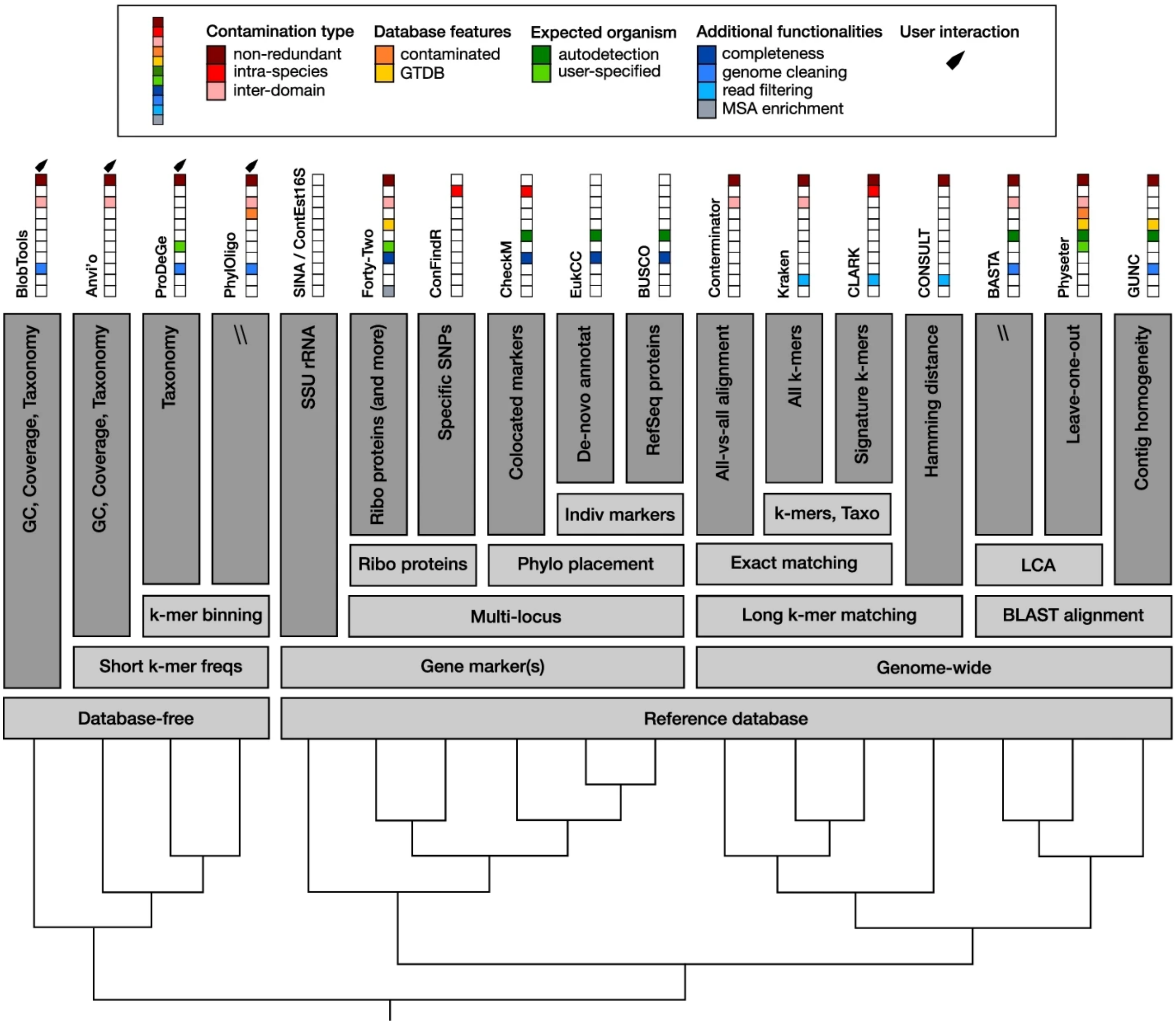


Figure 24. Overview of algorithms (from Cornet and Baurain 2022). “The algorithms are clusterized based on their operating principles, as described in the section “Contamination detection algorithms”. Squares on the top of the figure represent specific features of the algorithms. Non-redundant means that the software can detect contaminant genes without equivalent in the surveyed genome. Intra-species means that the algorithm can detect contamination at the species level. Inter-domain means that the algorithm can detect prokaryotic and eukaryotic contamination simultaneously. Database features show that the algorithm can use the GTDB Taxonomy and/or a moderately contaminated reference database. Expected organism indicates whether the algorithm can detect the main organism by itself and/or if the user can specify it. Additional functionalities list interesting peculiar functions of the programs, such as outputting the completeness of a genome, cleaning a genome from its contaminants, filtering reads based on their taxonomy (positive filtering), or enriching Multiple Sequence Alignments (MSAs) in orthologous

sequences while controlling the taxonomy”. (*) The Physeter algorithm is discussed in detail at Chapter 1 of the Results section.

1.5. Objectives and outline of the thesis

As introduced above, one of the dichotomies that distinct the two prokaryotic domains, Bacteria and Archaea, lies in their cell-wall composition. Although they may exhibit different architectures (e.g., monoderm, diderm), almost all bacterial cell walls bear a mesh-like polymer called peptidoglycan (or murein). In contrast, the paracrystalline S-layer is the most commonly encountered cell-wall in Archaea, even if some lineages of Euryarchaeota do have a polymer in their cell wall. Among those, Methanopyrales and Methanobacteriales feature the so-called pseudomurein, a structural analogue of the peptidoglycan.

This thesis is a part of the research line initiated during the PhD work of Raphael Léonard, entitled "Bacterial cell-wall architecture: from automated genome selection to evolution of genes and traits”, of which I am a co-author of the article “Was the Last Bacterial Common Ancestor a Monoderm after All?” (see **Annexes**) published in *Genes (Basel)* on the 18th February 2022. In this article, we inferred the cell-wall composition of the last bacterial common ancestor (LBCA) and proposed a scenario for bacterial cell-wall evolution based on phylogenomics, phenotypic data and single-gene phylogenies of the genes lying in the *dcw* cluster and those associated with the formation of the outer membrane.

In the present thesis, we further investigate the evolution of prokaryotic cell walls. More precisely, we aim to elucidate the genetic commonalities between Bacteria and the pseudomurein-containing Archaea (i.e., Methanopyrales and Methanobacteriales) that might have driven the evolution of the pseudomurein cell wall in the latter. For this purpose, public genomic data from the NCBI RefSeq database will be mined, in particular to study the phylogeny of different protein families involved in cell-wall biosynthesis.

1.5.1. Chapter 1 - Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics

As only public data will be used during this work, it is fundamental that our phylogenetic interpretation is not driven by potential genomic contamination.

Although NCBI RefSeq was built as a non-redundant and high-quality sequence and genome database, it has been suggested that it is not completely devoid of contamination. In this chapter, we developed *Physeter*, a reference database-dependent contamination detection algorithm and then applied it to thousands of RefSeq genomes. On this occasion, we confirmed that about 0.9% of the complete genomes in RefSeq are contaminated by foreign sequences. Since *Physeter* relies on a reference database itself derived from public repositories, we implemented a leave-one-out strategy, which allows us to reduce the impact of potentially contaminated genomes in the reference database.

The manuscript corresponding to this chapter entitled “Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics” was published on the 22th October 2021 in *Frontiers in Microbiology*.

The first version of *Physeter* was also used in the article of Javier Cordoba published in *Genes (Basel)* on the 29th May 2021 (see **Annexes**, of which I am also a co-author) to detect contamination in a transcriptome meta-assembly of the complex green alga *Euglena gracilis*. A revised pipeline of *Physeter* was then integrated into the Nextflow workflow *CRACOT*, which simulates contamination events to evaluate the accuracy of different contamination detection tools (one of these being *Physeter*). The bioRxiv version of the latter manuscript submitted to *Genome Biology* (https://doi.org/10.1101/2022.11.14.516442), of which I am a co-author, is presented in the **Annexes** of this thesis.

1.5.2. Chapter 2 - An Extended Reservoir of Class-D Beta-Lactamases in Non-Clinical Bacterial Strains

The cell-wall peptidoglycan plays a crucial role in the survival of bacteria. Indeed, it helps them to maintain their cell shape and protect them from internal turgor pressure. Moreover, it mediates the interactions between the cell and its environment. Therefore, peptidoglycan is one the main targets of antimicrobial drugs, such as beta-lactam antibiotics. In response, bacteria have developed resistance to beta-lactam antibiotics by synthesizing enzymes (i.e., beta-lactamases) that inhibit the action of these antimicrobial agents. This chapter focuses on the study of the class-D beta-lactamases. Here, we have mined more that 80,000 RefSeq genomes to assess the real distribution of the class-D beta-lactamases among the bacterial domain and study the phylogenetic relationships within the class-D family.

The first purpose of this chapter was to perform a pilot phylogenetic study of a protein family for which the nomenclature was doubtful, in order to set up analytical guidelines to be applied to the different protein families discussed in Chapter 3, in particular with respect to handling of sequence redundancy. The corresponding work was published on the 21th March 2022 in *Microbiology Spectrum*.

1.5.3. Chapter 3 - Origin and Evolution of Pseudomurein Biosynthetic Gene Clusters

The pseudomurein cell-wall is an oddity within the archaeal domain. Indeed, although some Euryarchaeota also possess a cell-wall polymer, only the pseudomurein is structurally similar to the bacterial peptidoglycan. Furthermore, some genomic studies have shown that homologs of certain genes involved in peptidoglycan biosynthesis are also found in the genomes of Methanopyrales and Methanobacteriales. In this chapter, we try to address the main question of this thesis: do peptidoglycan and pseudomurein have a common origin? More precisely, are the genes involved in their biosynthesis genuinely similar and, if so, what events have led to their current diversity and distribution?

This chapter corresponds to the bioRxiv version of a manuscript (https://doi.org/10.1101/2022.11.30.518518) that has just been submitted to *Molecular Biology and Evolution*.

1.6. References

Afzal-Shah M, Woodford N, Livermore DM. 2001. Characterization of OXA-25, OXA-26, and OXA-27, molecular class D beta-lactamases associated with carbapenem resistance in clinical isolates of Acinetobacter baumannii. *Antimicrob. Agents Chemother.* 45:583–588.

Albers S-V, Meyer BH. 2011. The archaeal cell envelope. *Nat. Rev. Microbiol.* 9:414–426.

Alfredson DA, Korolik V. 2005. Isolation and expression of a novel molecular class D beta-lactamase, OXA-61, from Campylobacter jejuni. *Antimicrob. Agents Chemother.* 49:2515–2518.

Ambler RP. 1980. The structure of beta-lactamases. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 289:321–331.

Angeles DM, Scheffers D-J. 2021. The Cell Wall of Bacillus subtilis. *Curr. Issues Mol. Biol.* 41:539–596.

Antonelli A, D'Andrea MM, Vaggelli G, Docquier J-D, Rossolini GM. 2015. OXA-372, a novel carbapenem-hydrolysing class D β-lactamase from a Citrobacter freundii isolated from a hospital wastewater plant. *J. Antimicrob. Chemother.* 70:2749–2756.

Antunes NT, Fisher JF. 2014. Acquired Class D β-Lactamases. *Antibiot. Basel Switz.* 3:398–434.

Aoki H, Okuhara M. 1980. Natural beta-lactam antibiotics. *Annu. Rev. Microbiol.* 34:159–181.

Arakawa K. 2016. No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc. Natl. Acad. Sci. U. S. A.* 113:E3057.

Arbing MA, Chan S, Shin A, Phan T, Ahn CJ, Rohlin L, Gunsalus RP. 2012. Structure of the surface layer of the methanogenic archaean Methanosarcina acetivorans. *Proc. Natl. Acad. Sci. U. S. A.* 109:11812–11817.

Arita M, Karsch-Mizrachi I, Cochrane G. 2021. The international nucleotide sequence database collaboration. *Nucleic Acids Res.* 49:D121–D124.

Atrih A, Bacher G, Allmaier G, Williamson MP, Foster SJ. 1999. Analysis of peptidoglycan structure from vegetative cells of Bacillus subtilis 168 and role of PBP 5 in peptidoglycan maturation. *J. Bacteriol.* 181:3956–3966.

Babic M, Hujer AM, Bonomo RA. 2006. What's new in antibiotic resistance? Focus on beta-lactamases. *Drug Resist. Updat. Rev. Comment. Antimicrob. Anticancer Chemother.* 9:142–156.

Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, Allen EE, Banfield JF. 2006. Lineages of acidophilic archaea revealed by community genomic analysis. *Science* 314:1933–1935.

Ball AP, Gray JA, Murdoch JMcM. 1978. The Natural Penicillins — Benzylpenicillin (Penicillin G) and Phenoxymethylpenicillin (Penicillin V). In: Ball AP, Gray JA, Murdoch JMcM, editors. Antibacterial Drugs Today. Dordrecht: Springer Netherlands. p. 6–18. Available from: https://doi.org/10.1007/978-94-011-8004-7_3

Balvočiūtė M, Huson DH. 2017. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? *BMC Genomics* 18:114.

Bapteste E, Brochier C, Boucher Y. 2005. Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. *Archaea Vanc. BC* 1:353–363.

Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt KD, Resenchuk S, Tatusova T, et al. 2012. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.* 40:D57–D63.

Baurain D, Wilmotte A, 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.

Bebrone C. 2007. Metallo-beta-lactamases (classification, activity, genetic organization, structure, zinc coordination) and their superfamily. *Biochem. Pharmacol.* 74:1686–1701.

Beceiro A, Bou G. 2004. Class C β-Lactamases: an increasing problem worldwide. *Rev. Res. Med. Microbiol.* [Internet] 15. Available from: https://journals.lww.com/revmedmicrobiol/Fulltext/2004/10000/Class_C___Lactamases__an_increasing_problem.3.aspx

Beeby M, Gumbart JC, Roux B, Jensen GJ. 2013. Architecture and assembly of the Gram-positive cell wall. *Mol. Microbiol.* 88:664–672.

Bennett JW, Chung KT. 2001. Alexander Fleming and the discovery of penicillin. *Adv. Appl. Microbiol.* 49:163–184.

Benson D, Boguski M, Lipman DJ, Ostell J. 1990. The National Center for

Biotechnology Information. *Genomics* 6:389–391.

Benson D, Lipman DJ, Ostell J. 1993. GenBank. *Nucleic Acids Res.* 21:2963–2965.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Res.* 41:D36-42.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, Sayers EW. 2018. GenBank. *Nucleic Acids Res.* 46:D41–D47.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2009. GenBank. *Nucleic Acids Res.* 37:D26-31.

Berzigotti S, Benlafya K, Sépulchre J, Amoroso A, Joris B. 2012. Bacillus licheniformis BlaR1 L3 loop is a zinc metalloprotease activated by self-proteolysis. *PloS One* 7:e36400.

Bhattacharjee MK. 2016a. Antimetabolites: Antibiotics That Inhibit Nucleotide Synthesis. In: Bhattacharjee MK, editor. Chemistry of Antibiotics and Related Drugs. Cham: Springer International Publishing. p. 95–108. Available from: https://doi.org/10.1007/978-3-319-40746-3_4

Bhattacharjee MK. 2016b. Antibiotics That Inhibit Protein Synthesis. In: Bhattacharjee MK, editor. Chemistry of Antibiotics and Related Drugs. Cham: Springer International Publishing. p. 129–151. Available from: https://doi.org/10.1007/978-3-319-40746-3_6

Bhattacharjee MK. 2016c. Antibiotics That Inhibit Cell Wall Synthesis. In: Bhattacharjee MK, editor. Chemistry of Antibiotics and Related Drugs. Cham: Springer International Publishing. p. 49–94. Available from: https://doi.org/10.1007/978-3-319-40746-3_3

Boniface A, Bouhss A, Mengin-Lecreulx D, Blanot D. 2006. The MurE synthetase from Thermotoga maritima is endowed with an unusual D-lysine adding activity. *J. Biol. Chem.* 281:15680–15686.

Bonnet R, Marchandin H, Chanal C, Sirot D, Labia R, De Champs C, Jumas-Bilak E, Sirot J. 2002. Chromosome-encoded class D beta-lactamase OXA-23 in Proteus mirabilis. *Antimicrob. Agents Chemother.* 46:2004–2006.

Bonomo RA. 2017. β-Lactamases: A Focus on Current Challenges. *Cold Spring Harb. Perspect. Med.* 7:a025239.

Bou G, Oliver A, Martínez-Beltrán J. 2000. OXA-24, a novel class D beta-lactamase with carbapenemase activity in an Acinetobacter baumannii clinical strain. *Antimicrob. Agents Chemother.* 44:1556–1561.

Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. 2019. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* 29:954–960.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12:59–60.

Burkovski A. 2013. Cell envelope of corynebacteria: structure and influence on pathogenicity. *ISRN Microbiol.* 2013:935736.

Burks C, Fickett JW, Goad WB, Kanehisa M, Lewitter FI, Rindone WP, Swindell CD, Tung C-S, Bilofsky HS. 1985. CABIOS REVIEW: The GenBank nucleic acid sequence database. *Bioinformatics* 1:225–233.

Bush K. 2018. Past and Present Perspectives on β-Lactamases. *Antimicrob. Agents Chemother.* 62:e01076-18.

Bush K, Bradford PA. 2016. β-Lactams and β-Lactamase Inhibitors: An Overview. *Cold Spring Harb. Perspect. Med.* 6:a025247.

Bush K, Jacoby GA, Medeiros AA. 1995. A functional classification scheme for beta-lactamases and its correlation with molecular structure. *Antimicrob. Agents Chemother.* 39:1211–1233.

Butler D. 2000. Ensembl gets a Wellcome boost. *Nature* 406:333–333.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Cantelli G, Bateman A, Brooksbank C, Petrov AI, Malik-Sheriff RS, Ide-Smith M, Hermjakob H, Flicek P, Apweiler R, Birney E, et al. 2022. The European Bioinformatics Institute (EMBL-EBI) in 2021. *Nucleic Acids Res.* 50:D11–D19.

Cantelli G, Cochrane G, Brooksbank C, McDonagh E, Flicek P, McEntyre J, Birney E, Apweiler R. 2021. The European Bioinformatics Institute: empowering cooperation in response to a global health crisis. *Nucleic Acids Res.* 49:D29–D37.

Cantón R, González-Alba JM, Galán JC. 2012. CTX-M Enzymes: Origin and Diffusion. *Front. Microbiol.* 3:110.

Carcione D, Siracusa C, Sulejmani A, Leoni V, Intra J. 2021. Old and New Beta-Lactamase Inhibitors: Molecular Structure, Mechanism of Action, and Clinical Use. *Antibiotics* 10:995.

Castanheira M, Simner PJ, Bradford PA. 2021. Extended-spectrum β-lactamases: an update on their characteristics, epidemiology and detection. *JAC-Antimicrob.*

Resist. 3:dlab092.

Cavalier-Smith T. 1987. The origin of eukaryotic and archaebacterial cells. *Ann. N. Y. Acad. Sci.* 503:17–54.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 GenesGenomesGenetics* 10:1361–1374.

Chatton E. 1925. Pansporella perplexa, Amoebien a spores protegees parasite des Daphnies. Reflexions sur la biologie et la phylogenie des Protozoaires. *Ann. Sci. Nat. Zool. Paris*:8: 5-84.

Chen Q, Zobel J, Verspoor K. 2017. Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. *Database* 2017:baw163.

Cho H, Wivagg CN, Kapoor M, Barry Z, Rohs PDA, Suh H, Marto JA, Garner EC, Bernhardt TG. 2016. Bacterial cell wall biogenesis is mediated by SEDS and PBP polymerase families functioning semi-autonomously. *Nat. Microbiol.* 1:16172.

Choudhuri S. 2014. Chapter 5 - Data, Databases, Data Format, Database Search, Data Retrieval Systems, and Genome Browsers. In: Choudhuri S, editor. Bioinformatics for Beginners. Oxford: Academic Press. p. 77–131. Available from: https://www.sciencedirect.com/science/article/pii/B9780124104716000050

Colombo M-L, Hanique S, Baurin SL, Bauvois C, De Vriendt K, Van Beeumen JJ, Frère J-M, Joris B. 2004. The ybxI gene of Bacillus subtilis 168 encodes a class D beta-lactamase of low activity. *Antimicrob. Agents Chemother.* 48:484–490.

Comolli LR, Baker BJ, Downing KH, Siegerist CE, Banfield JF. 2009. Three-dimensional analysis of the structure and ecology of a novel, ultra-small archaeon. *ISME J.* 3:159–167.

Cornet L, Baurain D. 2022. Contamination detection in genomic data: more is not enough. *Genome Biol.* 23:60.

Cummins C, Ahamed A, Aslam R, Burgin J, Devraj R, Edbali O, Gupta D, Harrison PW, Haseeb M, Holt S, et al. 2022. The European Nucleotide Archive in 2021. *Nucleic Acids Res.* 50:D106–D110.

Dajkovic A, Tesson B, Chauhan S, Courtin P, Keary R, Flores P, Marlière C, Filipe

SR, Chapot-Chartier M-P, Carballido-Lopez R. 2017. Hydrolysis of peptidoglycan is modulated by amidation of meso-diaminopimelic acid and Mg(2+) in Bacillus subtilis. *Mol. Microbiol.* 104:972–988.

Davies J, Davies D. 2010. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev. MMBR* 74:417–433.

De Simeis D, Serra S. 2021. Actinomycetes: A Never-Ending Source of Bioactive Compounds-An Overview on Antibiotics Production. *Antibiotics* 10:483.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends Genet. TIG* 30:418–426.

Doi Y, Paterson DL. 2007. Detection of plasmid-mediated class C beta-lactamases. *Int. J. Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis.* 11:191–197.

Dridi B, Fardeau M-L, Ollivier B, Raoult D, Drancourt M. 2012. Methanomassiliicoccus luminyensis gen. nov., sp. nov., a methanogenic archaeon isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* 62:1902–1907.

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform. Int. Conf. Genome Inform.* 23:205–211.

Egan AJF, Errington J, Vollmer W. 2020. Regulation of peptidoglycan synthesis and remodelling. *Nat. Rev. Microbiol.* 18:446–460.

Eiamphungporn W, Schaduangrat N, Malik AA, Nantasenamat C. 2018. Tackling the Antibiotic Resistance Caused by Class A β-Lactamases through the Use of β-Lactamase Inhibitory Protein. *Int. J. Mol. Sci.* 19:2222.

Elander RP. 2003. Industrial production of beta-lactam antibiotics. *Appl. Microbiol. Biotechnol.* 61:385–392.

Eme L, Spang A, Lombard J, Stairs CW, Ettema TJG. 2017. Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* 15:711–723.

Eraso JM, Markillie LM, Mitchell HD, Taylor RC, Orr G, Margolin W. 2014. The highly conserved MraZ protein is a transcriptional regulator in Escherichia coli. *J. Bacteriol.* 196:2053–2066.

Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi’o: an advanced analysis and visualization platform for ’omics data. *PeerJ* 3:e1319.

Espaillat A, Forsmo O, El Biari K, Björk R, Lemaitre B, Trygg J, Cañada FJ, de Pedro MA, Cava F. 2016. Chemometric Analysis of Bacterial Peptidoglycan Reveals

Atypical Modifications That Empower the Cell Wall against Predatory Enzymes and Fly Innate Immunity. *J. Am. Chem. Soc.* 138:9193–9204.

Evans BA, Amyes SGB. 2014. OXA β-lactamases. *Clin. Microbiol. Rev.* 27:241–263.

Fagan RP, Fairweather NF. 2014. Biogenesis and functions of bacterial S-layers. *Nat. Rev. Microbiol.* 12:211–222.

Federhen S. 2012. The NCBI Taxonomy database. *Nucleic Acids Res.* 40:D136-143.

Feng S, Wu Z, Liang W, Zhang X, Cai X, Li J, Liang L, Lin D, Stoesser N, Doi Y, et al. 2022. Prediction of Antibiotic Resistance Evolution by Growth Measurement of All Proximal Mutants of Beta-Lactamase. *Mol. Biol. Evol.* 39:msac086.

Fernandes R, Amador P, Prudêncio C. 2013. β-Lactams: chemical structure, mode of action and mechanisms of resistance. *Rev. Res. Med. Microbiol.* [Internet] 24. Available from: https://journals.lww.com/revmedmicrobiol/Fulltext/2013/01000/__Lactams__chemical_structure,_mode_of_action_and.2.aspx

Finet C, Timme RE, Delwiche CF, Marlétaz F. 2010. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Curr. Biol. CB* 20:2217–2222.

Fleming A. 1929. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. *Br. J. Exp. Pathol.* 10:226–236.

Formanek H. 1985. Three-Dimensional Models of the Carbohydrate Moieties of Murein and Pseudomurein. 40:555–561.

Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, et al. 1980. The phylogeny of prokaryotes. *Science* 209:457–463.

Francoleon DR, Boontheung P, Yang Y, Kin U, Ytterberg AJ, Denny PA, Denny PC, Loo JA, Gunsalus RP, Loo RRO. 2009. S-layer, surface-accessible, and concanavalin A binding proteins of Methanosarcina acetivorans and Methanosarcina mazei. *J. Proteome Res.* 8:1972–1982.

Frère JM. 1995. Beta-lactamases and bacterial resistance to antibiotics. *Mol. Microbiol.* 16:385–395.

Galleni M, Lamotte-Brasseur J, Rossolini GM, Spencer J, Dideberg O, Frère JM. 2001. Standard numbering scheme for class B beta-lactamases. *Antimicrob. Agents Chemother.* 45:660–663.

Gaynes R. 2017. The Discovery of Penicillin—New Insights After More Than 75 Years of Clinical Use. *Emerg. Infect. Dis.* 23:849–853.

Goffin C, Ghuysen JM. 1998. Multimodular penicillin-binding proteins: an enigmatic family of orthologs and paralogs. *Microbiol. Mol. Biol. Rev. MMBR* 62:1079–1093.

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, et al. 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 86:6661–6665.

Golemi D, Maveyraud L, Vakulenko S, Samama JP, Mobashery S. 2001. Critical involvement of a carbamylated lysine in catalytic function of class D beta-lactamases. *Proc. Natl. Acad. Sci. U. S. A.* 98:14280–14285.

Golemi-Kotra D, Cha JY, Meroueh SO, Vakulenko SB, Mobashery S. 2003. Resistance to beta-lactam antibiotics and its mediation by the sensor domain of the transmembrane BlaR signaling pathway in Staphylococcus aureus. *J. Biol. Chem.* 278:18419–18425.

Golyshina OV, Timmis KN. 2005. Ferroplasma and relatives, recently discovered cell wall-lacking archaea making a living in extremely acid, heavy metal-rich environments. *Environ. Microbiol.* 7:1277–1288.

Gostev M, Faulconbridge A, Brandizi M, Fernandez-Banet J, Sarkans U, Brazma A, Parkinson H. 2012. The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acids Res.* 40:D64–D70.

Graham DE, Huse HK. 2008. Methanogens with pseudomurein use diaminopimelate aminotransferase in lysine biosynthesis. *FEBS Lett.* 582:1369–1374.

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

Gupta RS. 1998. 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.

Guy L, Ettema TJG. 2011. The archaeal “TACK” superphylum and the origin of eukaryotes. *Trends Microbiol.* 19:580–587.

Haaber J, Penadés JR, Ingmer H. 2017. Transfer of Antibiotic Resistance in Staphylococcus aureus. *Trends Microbiol.* 25:893–905.

Haeggman S, Löfdahl S, Burman LG. 1997. An allelic variant of the chromosomal gene for class A beta-lactamase K2, specific for Klebsiella pneumoniae, is the ancestor of SHV-1. *Antimicrob. Agents Chemother.* 41:2705–2709.

Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, Li W, Chitsaz F, Derbyshire MK, Gonzales NR, et al. 2018. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* 46:D851–D860.

Hall BG, Barlow M. 2005. Revised Ambler classification of β-lactamases. *J. Antimicrob. Chemother.* 55:1050–1051.

Hammoudi Halat D, Ayoub Moubareck C. 2020. The Current Burden of Carbapenemases: Review of Significant Properties and Dissemination among Gram-Negative Bacteria. *Antibiotics* 9.

Hardt K, Joris B, Lepage S, Brasseur R, Lampen JO, Frère JM, Fink AL, Ghuysen JM. 1997. The penicillin sensory transducer, BlaR, involved in the inducibility of beta-lactamase synthesis in Bacillus licheniformis is embedded in the plasma membrane via a four-alpha-helix bundle. *Mol. Microbiol.* 23:935–944.

Harrison PW, Ahamed A, Aslam R, Alako BTF, Burgin J, Buso N, Courtot M, Fan J, Gupta D, Haseeb M, et al. 2021. The European Nucleotide Archive in 2020. *Nucleic Acids Res.* 49:D82–D85.

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

Hartmann E, König H. 1994. A novel pathway of peptide biosynthesis found in methanogenic Archaea. *Arch. Microbiol.* 162:430–432.

Healy VL, Lessard IA, Roper DI, Knox JR, Walsh CT. 2000. Vancomycin resistance in enterococci: reprogramming of the D-ala-D-Ala ligases in bacterial peptidoglycan biosynthesis. *Chem. Biol.* 7:R109-119.

Heather JM, Chain B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics* 107:1–8.

Heimerl T, Flechsler J, Pickl C, Heinz V, Salecker B, Zweck J, Wanner G, Geimer S, Samson RY, Bell SD, et al. 2017. A Complex Endomembrane System in the Archaeon Ignicoccus hospitalis Tapped by Nanoarchaeum equitans. *Front. Microbiol.* 8:1072.

Hesse L, Bostock J, Dementin S, Blanot D, Mengin-Lecreulx D, Chopra I. 2003. Functional and biochemical analysis of Chlamydia trachomatis MurC, an enzyme displaying UDP-N-acetylmuramate:amino acid ligase activity. *J.*

Bacteriol. 185:6507–6512.

Hoiczyk E, Hansel A. 2000. Cyanobacterial cell walls: news from an unusual prokaryotic envelope. *J. Bacteriol.* 182:1191–1199.

Höltje JV. 1998. Growth of the stress-bearing and shape-maintaining murein sacculus of Escherichia coli. *Microbiol. Mol. Biol. Rev. MMBR* 62:181–203.

Howe KL, Achuthan P, Allen James, Allen Jamie, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, et al. 2021. Ensembl 2021. *Nucleic Acids Res.* 49:D884–D891.

Hsueh Y-H, Huang K-Y, Kunene SC, Lee T-Y. 2017. Poly-γ-glutamic Acid Synthesis, Gene Regulation, Phylogenetic Relationships, and Role in Fermentation. *Int. J. Mol. Sci.* 18:2644.

Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Res.* 17:377–386.

Hutchings MI, Truman AW, Wilkinson B. 2019. Antibiotics: past, present and future. *Curr. Opin. Microbiol.* 51:72–80.

ICNP. 2019. International Code of Nomenclature of Prokaryotes. *Int. J. Syst. Evol. Microbiol.* 69:S1–S111.

Ithurbide S, Gribaldo S, Albers S-V, Pende N. 2022. Spotlight on FtsZ-based cell division in Archaea. *Trends Microbiol.* 30:665–678.

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* 86:9355–9359.

Jacoby GA. 2009. AmpC beta-lactamases. *Clin. Microbiol. Rev.* 22:161–182.

Jago M, Heatley NG. 1961. Some biological properties of cephalosporin C and a derivative. *Br. J. Pharmacol. Chemother.* 16:170–179.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236–1240.

Joris B, Dusart J. 2012. The induction of beta-lactamases in Eubacteria. In: Nova Science Publishers.

Kahlke T, Ralph PJ. 2019. BASTA – Taxonomic classification of sequences and sequence bins using last common ancestor estimations. *Methods Ecol. Evol.* 10:100–103.

Kandler O, Konig H. 1993. Chapter 8 Cell envelopes of archaea: Structure and chemistry. In: Kates M, Kushner DJ, Matheson AT, editors. New Comprehensive Biochemistry. Vol. 26. Elsevier. p. 223–259. Available from: https://www.sciencedirect.com/science/article/pii/S0167730608602574

Kapoor G, Saigal S, Elongavan A. 2017. Action and resistance mechanisms of antibiotics: A guide for clinicians. *J. Anaesthesiol. Clin. Pharmacol.* 33:300–305.

Karsch-Mizrachi I, Nakamura Y, Cochrane G. 2012. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* 40:D33-37.

Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O’Sullivan C. 2022. The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* 50:D387–D390.

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

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

Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A, et al. 2016. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* 44:D73–D80.

Klimke W, O’Donovan C, White O, Brister JR, Clark K, Fedorov B, Mizrachi I, Pruitt KD, Tatusova T. 2011. Solving the Problem: Genome Annotation Standards before the Data Deluge. *Stand. Genomic Sci.* 5:168–193.

Klingl A, Pickl C, Flechsler J. 2019. Archaeal Cell Walls. In: Kuhn A, editor. Bacterial Cell Walls and Membranes. Cham: Springer International Publishing. p. 471–493. Available from: https://doi.org/10.1007/978-3-030-18768-2_14

Kodama Y, Shumway M, Leinonen R. 2012. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* 40:D54-56.

König H, Hartmann E, Kärcher U. 1993. Pathways and Principles of the Biosynthesis of Methanobacterial Cell Wall Polymers. *Syst. Appl. Microbiol.* 16:510–517.

Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker AA, Blaxter M. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. *Proc. Natl. Acad. Sci. U. S. A.* 113:5053–5058.

Kreisl P, Kandler O. 1986. Chemical structure of the cell wall polymer of methanosarcina. *Syst. Appl. Microbiol.* 7:293–299.

Lakaye B, Dubus A, Lepage S, Groslambert S, Frère JM. 1999. When drug inactivation renders the target irrelevant to antibiotic resistance: a case story with beta-lactams. *Mol. Microbiol.* 31:89–101.

Lang BF, Gray MW, Burger G. 1999. Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.* 33:351–397.

Langworthy TA, Smith PF, Mayberry WR. 1972. Lipids of Thermoplasma acidophilum. *J. Bacteriol.* 112:1193–1200.

Lasek-Nesselquist E, Gogarten JP. 2013. The effects of model choice and mitigating bias on the ribosomal tree of life. *Mol. Phylogenet. Evol.* 69:17–38.

Laurin-Lemay S, Brinkmann H, Philippe H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. *Curr. Biol. CB* 22:R593-594.

Leahy SC, Kelly WJ, Altermann E, Ronimus RS, Yeoman CJ, Pacheco DM, Li D, Kong Z, McTavish S, Sang C, et al. 2010. The genome sequence of the rumen methanogen Methanobrevibacter ruminantium reveals new possibilities for controlling ruminant methane emissions. *PloS One* 5:e8926.

Lee I, Chalita M, Ha S-M, Na S-I, Yoon S-H, Chun J. 2017. ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences. *Int. J. Syst. Evol. Microbiol.* 67:2053–2057.

Leinonen R, Sugawara H, Shumway M. 2011. The sequence read archive. *Nucleic Acids Res.* 39:D19-21.

Leisner JJ. 2020. The Diverse Search for Synthetic, Semisynthetic and Natural Product Antibiotics From the 1940s and Up to 1960 Exemplified by a Small Pharmaceutical Player. *Front. Microbiol.* 11:976.

Leonard DA, Bonomo RA, Powers RA. 2013. Class D β-lactamases: a reappraisal after five decades. *Acc. Chem. Res.* 46:2407–2415.

Léonard RR, Sauvage E, Lupo V, Perrin A, Sirjacobs D, Charlier P, Kerff F, Baurain D. 2022. Was the Last Bacterial Common Ancestor a Monoderm after All? *Genes* 13:376.

Leonelli S. 2015. What Counts as Scientific Data? A Relational Framework. *Philos. Sci.* 82:810–821.

Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire MK, Durkin AS, et al. 2021. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.* 49:D1020–D1028.

Liakopoulos A, Mevius D, Ceccarelli D. 2016. A Review of SHV Extended-Spectrum β-Lactamases: Neglected Yet Ubiquitous. *Front. Microbiol.* 7:1374.

Llarrull LI, Toth M, Champion MM, Mobashery S. 2011. Activation of BlaR1 protein of methicillin-resistant Staphylococcus aureus, its proteolytic processing, and recovery from induction of resistance. *J. Biol. Chem.* 286:38148–38158.

Lobanovska M, Pilla G. 2017. Penicillin's Discovery and Antibiotic Resistance: Lessons for the Future? *Yale J. Biol. Med.* 90:135–145.

López-Pelegrín M, Cerdà-Costa N, Martínez-Jiménez F, Cintas-Pedrola A, Canals A, Peinado JR, Marti-Renom MA, López-Otín C, Arolas JL, Gomis-Rüth FX. 2013. A novel family of soluble minimal scaffolds provides structural insight into the catalytic domains of integral membrane metallopeptidases. *J. Biol. Chem.* 288:21279–21294.

Low AJ, Koziol AG, Manninger PA, Blais B, Carrillo CD. 2019. ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. *PeerJ* 7:e6995.

Luo Y, Pfister P, Leisinger T, Wasserfallen A. 2001. The genome of archaeal prophage PsiM100 encodes the lytic enzyme responsible for autolysis of Methanothermobacter wolfeii. *J. Bacteriol.* 183:5788–5792.

Luo Y, Pfister P, Leisinger T, Wasserfallen A. 2002. Pseudomurein endoisopeptidases PeiW and PeiP, two moderately related members of a novel family of proteases produced in Methanothermobacter strains. *FEMS Microbiol. Lett.* 208:47–51.

Macalady JL, Vestling MM, Baumler D, Boekelheide N, Kaspar CW, Banfield JF. 2004. Tetraether-linked membrane monolayers in Ferroplasma spp: a key to survival in acid. *Extrem. Life Extreme Cond.* 8:411–419.

Maglott DR, Katz KS, Sicotte H, Pruitt KD. 2000. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* 28:126–128.

Magnet S, Dubost L, Marie A, Arthur M, Gutmann L. 2008. Identification of the L,D-transpeptidases for peptidoglycan cross-linking in Escherichia coli. *J. Bacteriol.* 190:4782–4785.

Mahapatra S, Crick DC, Brennan PJ. 2000. Comparison of the

UDP-N-acetylmuramate:L-alanine ligase enzymes from Mycobacterium tuberculosis and Mycobacterium leprae. *J. Bacteriol.* 182:6827–6830.

Maitra A, Nukala S, Dickman R, Martin LT, Munshi T, Gupta A, Shepherd AJ, Arnvig KB, Tabor AB, Keep NH, et al. 2021. Characterization of the MurT/GatD complex in Mycobacterium tuberculosis towards validating a novel anti-tubercular drug target. *JAC-Antimicrob. Resist.* 3:dlab028.

Mallet L, Bitard-Feildel T, Cerutti F, Chiapello H. 2017. PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies. *Bioinforma. Oxf. Engl.* 33:3283–3285.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* 38:4647–4654.

Margulis L. 1970. Origin of eukaryotic cells: Evidence and research implications for a theory of the origin and evolution of microbial, plant and animal cells on the precambrian Earth. Yale University Press

Martínez-Torró C, Torres-Puig S, Marcos-Silva M, Huguet-Ramón M, Muñoz-Navarro C, Lluch-Senar M, Serrano L, Querol E, Piñol J, Pich OQ. 2021. Functional Characterization of the Cell Division Gene Cluster of the Wall-less Bacterium Mycoplasma genitalium. *Front. Microbiol.* 12:695572.

Matagne A, Dubus A, Galleni M, Frère JM. 1999. The beta-lactamase cycle: a tale of selective pressure and bacterial ingenuity. *Nat. Prod. Rep.* 16:1–19.

Matias VRF, Al-Amoudi A, Dubochet J, Beveridge TJ. 2003. Cryo-transmission electron microscopy of frozen-hydrated sections of Escherichia coli and Pseudomonas aeruginosa. *J. Bacteriol.* 185:6112–6118.

Matias VRF, Beveridge TJ. 2005. Cryo-electron microscopy reveals native polymeric cell wall structure in Bacillus subtilis 168 and the existence of a periplasmic space. *Mol. Microbiol.* 56:240–251.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6:610–618.

Meeske AJ, Riley EP, Robins WP, Uehara T, Mekalanos JJ, Kahne D, Walker S, Kruse AC, Bernhardt TG, Rudner DZ. 2016. SEDS proteins are a widespread family of bacterial cell wall polymerases. *Nature* 537:634–638.

Meyer BH, Albers S-V. 2020. Archaeal Cell Walls. In: eLS. p. 1–14. Available from: https://doi.org/10.1002/9780470015902.a0000384.pub3

Miethke M, Pieroni M, Weber T, Brönstrup M, Hammann P, Halby L, Arimondo PB, Glaser P, Aigle B, Bode HB, et al. 2021. Towards the sustainable discovery and development of new antibiotics. *Nat. Rev. Chem.* 5:726–749.

Mingorance J, Tamames J. 2004. The bacterial dcw gene cluster: an island in the genome? In: Vicente M, Tamames J, Valencia A, Mingorance J, editors. Molecules in Time and Space: Bacterial Shape, Division and Phylogeny. Boston, MA: Springer US. p. 249–271. Available from: https://doi.org/10.1007/0-306-48579-6_13

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41:e121.

Mohammadi T, van Dam V, Sijbrandi R, Vernet T, Zapun A, Bouhss A, Diepeveen-de Bruin M, Nguyen-Distèche M, de Kruijff B, Breukink E. 2011. Identification of FtsW as a transporter of lipid-linked cell wall precursors across the membrane. *EMBO J.* 30:1425–1432.

Mohammadi T, Sijbrandi R, Lutters M, Verheul J, Martin NI, den Blaauwen T, de Kruijff B, Breukink E. 2014. Specificity of the transport of lipid II by FtsW in Escherichia coli. *J. Biol. Chem.* 289:14707–14718.

Mojica MF, Rossi M-A, Vila AJ, Bonomo RA. 2022. The urgent need for metallo-β-lactamase inhibitors: an unattended global threat. *Lancet Infect. Dis.* 22:e28–e34.

Morlot C, Straume D, Peters K, Hegnar OA, Simon N, Villard A-M, Contreras-Martel C, Leisico F, Breukink E, Gravier-Pelletier C, et al. 2018. Structure of the essential peptidoglycan amidotransferase MurT/GatD complex from Streptococcus pneumoniae. *Nat. Commun.* 9:3180.

Mrozek D, Małysiak-Mrozek B, Siążnik A. 2013. search GenBank: interactive orchestration and ad-hoc choreography of Web services in the exploration of the biomedical resources of the National Center For Biotechnology Information. *BMC Bioinformatics* 14:73.

Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. 2015. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand.*

Genomic Sci. 10:18.

Münch D, Roemer T, Lee SH, Engeser M, Sahl HG, Schneider T. 2012. Identification and in vitro analysis of the GatD/MurT enzyme-complex catalyzing lipid II amidation in Staphylococcus aureus. *PLoS Pathog.* 8:e1002509.

Munita JM, Arias CA. 2016. Mechanisms of Antibiotic Resistance. *Microbiol. Spectr.* 4:VMBF-0016-2015.

Murray-Rust P. 2008. Open Data in Science. *Nat. Preced.* [Internet]. Available from: https://doi.org/10.1038/npre.2008.1526.1

Naas T, Oueslati S, Bonnin RA, Dabos ML, Zavala A, Dortet L, Retailleau P, Iorga BI. 2017. Beta-lactamase database (BLDB) - structure and function. *J. Enzyme Inhib. Med. Chem.* 32:917–919.

Nadim T. 2016. Data Labours: How the Sequence Databases GenBank and EMBL-Bank Make Data. *Sci. Cult.* 25:496–519.

Newton GG, Abraham EP. 1955. Cephalosporin C, a new antibiotic containing sulphur and D-alpha-aminoadipic acid. *Nature* 175:548.

Ngadjeua F, Braud E, Saidjalolov S, Iannazzo L, Schnappinger D, Ehrt S, Hugonnet J-E, Mengin-Lecreulx D, Patin D, Ethève-Quelquejeu M, et al. 2018. Critical Impact of Peptidoglycan Precursor Amidation on the Activity of l,d-Transpeptidases from Enterococcus faecium and Mycobacterium tuberculosis. *Chem. Weinh. Bergstr. Ger.* 24:5743–5747.

Niemetz R, Kärcher U, Kandler O, Tindall BJ, König H. 1997. The cell wall polymer of the extremely halophilic archaeon Natronococcus occultus. *Eur. J. Biochem.* 249:905–911.

Nobs S-J, MacLeod FI, Wong HL, Burns BP. 2022. Eukarya the chimera: eukaryotes, a secondary innovation of the two domains of life? *Trends Microbiol.* 30:421–431.

Nöldeke ER, Muckenfuss LM, Niemann V, Müller A, Störk E, Zocher G, Schneider T, Stehle T. 2018. Structural basis of cell wall peptidoglycan amidation by the GatD/MurT complex of Staphylococcus aureus. *Sci. Rep.* 8:12953.

Okido T, Kodama Y, Mashima J, Kosuge T, Fujisawa T, Ogasawara O. 2022. DNA Data Bank of Japan (DDBJ) update report 2021. *Nucleic Acids Res.* 50:D102–D105.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence

(RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44:D733-745.

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 Res.* 44:D733–D745.

Orakov A, Fullam A, Coelho LP, Khedkar S, Szklarczyk D, Mende DR, Schmidt TSB, Bork P. 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* 22:178.

Oren A, Garrity GM. 2021. Valid publication of the names of forty-two phyla of prokaryotes. *Int. J. Syst. Evol. Microbiol.* 71:005056.

Ounit R, Wanamaker S, Close TJ, Lonardi S. 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 16:236.

Page MGP. 2012. Beta-Lactam Antibiotics. In: Dougherty TJ, Pucci MJ, editors. Antibiotic Discovery and Development. Boston, MA: Springer US. p. 79–117. Available from: https://doi.org/10.1007/978-1-4614-1400-1_3

Palzkill T. 2013. Metallo-β-lactamase structure and function. *Ann. N. Y. Acad. Sci.* 1277:91–104.

Papp-Wallace KM, Endimiani A, Taracila MA, Bonomo RA. 2011. Carbapenems: past, present, and future. *Antimicrob. Agents Chemother.* 55:4943–4960.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36:996–1004.

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

Pasquetto IV, Randles BM, Borgman CL. 2017. On the Reuse of Scientific Data. *Data Sci. J.* 16:8.

Paterson DL, Bonomo RA. 2005. Extended-spectrum beta-lactamases: a clinical update. *Clin. Microbiol. Rev.* 18:657–686.

Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, et al. 2022. InterPro in 2022. *Nucleic Acids Res.*:gkac993.

Pazos M, Peters K. 2019. Peptidoglycan. In: Kuhn A, editor. Bacterial Cell Walls and Membranes. Cham: Springer International Publishing. p. 127–168. Available from: https://doi.org/10.1007/978-3-030-18768-2_5

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D, Philippe H, Brinkmann H, Lavrov DV, et al. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* 9:e1000602.

Philippon A, Arlet G, Labia R, Iorga BI. 2022. Class C β-Lactamases: Molecular Characteristics. *Clin. Microbiol. Rev.* 35:e0015021.

Pilhofer M, Rappl K, Eckl C, Bauer AP, Ludwig W, Schleifer K-H, Petroni G. 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.

Poirel L, Naas T, Nordmann P. 2010. Diversity, epidemiology, and genetics of class D beta-lactamases. *Antimicrob. Agents Chemother.* 54:24–38.

Poole K. 2004. Resistance to beta-lactam antibiotics. *Cell. Mol. Life Sci. CMLS* 61:2200–2223.

Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server: 2018 update. *Nucleic Acids Res.* 46:W200–W204.

Poux S, Arighi CN, Magrane M, Bateman A, Wei C-H, Lu Z, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B, et al. 2017. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinforma. Oxf. Engl.* 33:3454–3460.

Pruesse E, Peplies J, Glöckner FO. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinforma. Oxf. Engl.* 28:1823–1829.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 33:D501-504.

Queenan AM, Bush K. 2007. Carbapenemases: the versatile beta-lactamases. *Clin. Microbiol. Rev.* 20:440–458.

Rachel R, Engel AM, Huber R, Stetter K-O, 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 KO, Baumeister W. 1988. The Structure of the Surface Protein of Thermotoga maritima. In: Sleytr UB, Messner P, Pum D, Sára M, editors. Crystalline Bacterial Cell Surface Layers. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 83–86.

Rachel R, Wyschkony I, Riehl S, Huber H. 2002. The ultrastructure of Ignicoccus: evidence for a novel outer membrane and for intracellular vesicle budding in an archaeon. *Archaea Vanc. BC* 1:9–18.

Rachtman E, Bafna V, Mirarab S. 2021. CONSULT: accurate contamination removal using locality-sensitive hashing. *NAR Genomics Bioinforma.* 3:lqab071.

Rahlwes KC, Sparks IL, Morita YS. 2019. Cell Walls and Membranes of Actinobacteria. *Subcell. Biochem.* 92:417–469.

Ramsey C, MacGowan AP. 2016. A review of the pharmacokinetics and pharmacodynamics of aztreonam. *J. Antimicrob. Chemother.* 71:2704–2712.

Ranjit C, Noll KM. 2016. Distension of the toga of Thermotoga maritima involves continued growth of the outer envelope as cells enter the stationary phase. *FEMS Microbiol. Lett.* 363:fnw218.

Rasmussen BA, Keeney D, Yang Y, Bush K. 1994. Cloning and expression of a cloxacillin-hydrolyzing enzyme and a cephalosporinase from Aeromonas sobria AER 14M in Escherichia coli: requirement for an E. coli chromosomal mutation for efficient expression of the class D enzyme. *Antimicrob. Agents Chemother.* 38:2078–2085.

Rawat D, Nair D. 2010. Extended-spectrum β-lactamases in Gram Negative Bacteria. *J. Glob. Infect. Dis.* 2:263–274.

Raymann K, Brochier-Armanet C, Gribaldo S. 2015. The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci.*:201420858.

Real G, Henriques AO. 2006. Localization of the Bacillus subtilis murB gene within the dcw cluster is important for growth and sporulation. *J. Bacteriol.* 188:1721–1732.

Ride W ed. 1999. International Code of Zoological Nomenclature. London: International Trust for Zoological Nomenclature

Rinke C, Chuvochina M, Mussig AJ, Chaumeil P-A, Davín AA, Waite DW, Whitman WB, Parks DH, Hugenholtz P. 2021. A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat. Microbiol.* 6:946–959.

Rodrigues-Oliveira T, Belmok A, Vasconcellos D, Schuster B, Kyaw CM. 2017.

Archaeal S-Layers: Overview and Current State of the Art. *Front. Microbiol.* 8:2597.

Ruiz N. 2008. Bioinformatics identification of MurJ (MviN) as the peptidoglycan lipid II flippase in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* 105:15553–15557.

Saary P, Mitchell AL, Finn RD. 2020. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biol.* 21:244.

Sakamoto T, Ortega JM. 2021. Taxallnomy: an extension of NCBI Taxonomy that produces a hierarchically complete taxonomic tree. *BMC Bioinformatics* 22:388.

Salverda MLM, De Visser JAGM, Barlow M. 2010. Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS Microbiol. Rev.* 34:1015–1036.

Samuel BS, Hansen EE, Manchester JK, Coutinho PM, Henrissat B, Fulton R, Latreille P, Kim K, Wilson RK, Gordon JI. 2007. Genomic and metabolic adaptations of Methanobrevibacter smithii to the human gut. *Proc. Natl. Acad. Sci. U. S. A.* 104:10643–10648.

Sauvage E, Kerff F, Terrak M, Ayala JA, Charlier P. 2008. The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis. *FEMS Microbiol. Rev.* 32:234–258.

Sawa T, Kooguchi K, Moriyama K. 2020. Molecular diversity of extended-spectrum β-lactamases and carbapenemases, and antimicrobial resistance. *J. Intensive Care* 8:13.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, et al. 2011. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 39:D38–D51.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, et al. 2022. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 50:D20–D26.

Scheffers D-J, Pinho MG. 2005. Bacterial cell wall synthesis: new insights from localization studies. *Microbiol. Mol. Biol. Rev. MMBR* 69:585–607.

Schierwater B, Kolokotronis S-O, Eitel M, DeSalle R. 2009. The Diploblast-Bilateria Sister hypothesis: parallel revolution of a nervous systems may have been a simple step. *Commun. Integr. Biol.* 2:403–405.

Schleifer KH, Kandler O. 1972. Peptidoglycan types of bacterial cell walls and their taxonomic implications. *Bacteriol. Rev.* 36:407–477.

Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, et al. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* 2020:baaa062.

Schofield LR, Beattie AK, Tootill CM, Dey D, Ronimus RS. 2015. Biochemical Characterisation of Phage Pseudomurein Endoisopeptidases PeiW and PeiP Using Synthetic Peptides. *Archaea Vanc. BC* 2015:828693.

Schuler GD, Epstein JA, Ohkawa H, Kans JA. 1996. [10] Entrez: Molecular biology database and retrieval system. In: Methods in Enzymology. Vol. 266. Academic Press. p. 141–162. Available from: https://www.sciencedirect.com/science/article/pii/S0076687996660121

Silhavy TJ, Kahne D, Walker S. 2010. The bacterial cell envelope. *Cold Spring Harb. Perspect. Biol.* 2:a000414.

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, et al. 2017. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr. Biol. CB* 27:958–967.

Slesarev AI, Mezhevaya KV, Makarova KS, Polushin NN, Shcherbinina OV, Shakhova VV, Belova GI, Aravind L, Natale DA, Rogozin IB, et al. 2002. The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. *Proc. Natl. Acad. Sci. U. S. A.* 99:4644–4649.

Sleytr UB, Schuster B, Egelseer E-M, Pum D. 2014. S-layers: principles and applications. *FEMS Microbiol. Rev.* 38:823–864.

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, et al. 1997. Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. *J. Bacteriol.* 179:7135–7155.

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

Sperandeo P, Martorana AM, Polissi A. 2019. Lipopolysaccharide Biosynthesis and Transport to the Outer Membrane of Gram-Negative Bacteria. In: Kuhn A, editor. Bacterial Cell Walls and Membranes. Cham: Springer International Publishing. p. 9–37. Available from: https://doi.org/10.1007/978-3-030-18768-2_2

Stanier RY, Van Niel CB. 1962. The concept of a bacterium. *Arch. Mikrobiol.* 42:17–35.

Starobogatov YI. 1991. Problems in the nomenclature of higher taxonomic categories. *Bull. Zool. Nomencl.* 48:6–18.

Steber J, Schleifer KH. 1979. N-Glycyl-glucosamine: A novel constituent in the cell wall of Halococcus morrhuae. *Arch. Microbiol.* 123:209–212.

Steinegger M, Salzberg SL. 2020. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol.* 21:115.

Subedi BP, Martin WF, Carbone V, Duin EC, Cronin B, Sauter J, Schofield LR, Sutherland-Smith AJ, Ronimus RS. 2021. Archaeal pseudomurein and bacterial murein cell wall biosynthesis share a common evolutionary ancestry. *FEMS Microbes* 2:xtab012.

Supuran CT. 2017. Special Issue: Sulfonamides. *Molecules* 22:1642.

Szarecka A, Lesnock KR, Ramirez-Mondragon CA, Nicholas HBJ, Wymore T. 2011. The Class D beta-lactamase family: residues governing the maintenance and diversity of function. *Protein Eng. Des. Sel. PEDS* 24:801–809.

Taguchi A, Welsh MA, Marmont LS, Lee W, Sjodt M, Kruse AC, Kahne D, Bernhardt TG, Walker S. 2019. FtsW is a peptidoglycan polymerase that is functional only in complex with its cognate penicillin-binding protein. *Nat. Microbiol.* 4:587–594.

Tamames J. 2001. Evolution of gene order conservation in prokaryotes. *Genome Biol.* 2:RESEARCH0020.

Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* 44:6614–6624.

Tennessen K, Andersen E, Clingenpeel S, Rinke C, Lundberg DS, Han J, Dangl JL, Ivanova N, Woyke T, Kyrpides N, et al. 2016. ProDeGe: a computational protocol for fully automated decontamination of genomes. *ISME J.*

10:269–272.

The UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49:D480–D489.

Tooke CL, Hinchliffe P, Bragginton EC, Colenso CK, Hirvonen VHA, Takebayashi Y, Spencer J. 2019. β-Lactamases and β-Lactamase Inhibitors in the 21st Century. *J. Mol. Biol.* 431:3472–3500.

Toth M, Antunes NT, Stewart NK, Frase H, Bhattacharya M, Smith CA, Vakulenko SB. 2016. Class D β-lactamases do exist in Gram-positive bacteria. *Nat. Chem. Biol.* 12:9–14.

Toth M, Stewart NK, Smith C, Vakulenko SB. 2018. Intrinsic Class D β-Lactamases of Clostridium difficile. *mBio* 9:e01803-18.

Trachtenberg S. 1998. Mollicutes-wall-less bacteria with internal cytoskeletons. *J. Struct. Biol.* 124:244–256.

Turland NJ, Wiersema JH, Barrie FR, Greuter W, Hawksworth DL, Herendeen PS, Knapp S, Kusber W-H, Li D-Z, Marhold K, et al. 2018. International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017. *Regnum Veg. Vol. 159* [Internet]. Available from: http://hdl.handle.net/10141/622572

Ventola CL. 2015. The antibiotic resistance crisis: part 1: causes and threats. *P T Peer-Rev. J. Formul. Manag.* 40:277–283.

Visweswaran GRR, Dijkstra BW, Kok J. 2010. Two major archaeal pseudomurein endoisopeptidases: PeiW and PeiP. *Archaea Vanc. BC* 2010:480492.

Vollmer W, Blanot D, de Pedro MA. 2008. Peptidoglycan structure and architecture. *FEMS Microbiol. Rev.* 32:149–167.

Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Dempsey DM, Dutilh BE, Harrach B, Harrison RL, Hendrickson RC, Junglen S, et al. 2019. Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019). *Arch. Virol.* 164:2417–2429.

Walther-Rasmussen J, Høiby N. 2006. OXA-type carbapenemases. *J. Antimicrob. Chemother.* 57:373–383.

Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl.*

Environ. Microbiol. 73:5261–5267.

Weiss MC, Preiner M, Xavier JC, Zimorski V, Martin WF. 2018. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genet.* 14:e1007518.

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

Williams TA, Embley TM. 2014. Archaeal “dark matter” and the origin of eukaryotes. *Genome Biol. Evol.* 6:474–481.

Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504:231–236.

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

Woese CR, Kandlert O, Wheelis ML. 1990. Towards a natural system of organisms : Proposal for the domains Archaea, Bacteria and Eucarya. 87:4576–4579.

Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20:257.

Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46.

Y. Demchenko, Z. Zhao, P. Grosso, A. Wibisono, C. de Laat. 2012. Addressing Big Data challenges for Scientific Data Infrastructure. *4th IEEE Int. Conf. Cloud Comput. Technol. Sci. Proc.*:614–617.

Yanagihara Y, Kamisango K, Yasuda S, Kobayashi S, Mifuchi I, Azuma I, Yamamura Y, Johnson RC. 1984. Chemical compositions of cell walls and polysaccharide fractions of spirochetes. *Microbiol. Immunol.* 28:535–544.

Yang LL, Haug A. 1979. Structure of membrane lipids and physico-biochemical properties of the plasma membrane from Thermoplasma acidophilum, adapted to growth at 37 degrees C. *Biochim. Biophys. Acta* 573:308–320.

Yao X, Jericho M, Pink D, Beveridge T. 1999. Thickness and elasticity of gram-negative murein sacculi measured by atomic force microscopy. *J. Bacteriol.* 181:6865–6875.

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. 2014. The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Res.* 42:D643–D648.

Yoon E-J, Jeong SH. 2021. Class D β-lactamases. *J. Antimicrob. Chemother.* 76:836–864.

Zaman SB, Hussain MA, Nye R, Mehta V, Mamun KT, Hossain N. 2017. A Review on Antibiotic Resistance: Alarm Bells are Ringing. *Cureus* 9:e1403.

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

Zenke R, von Gronau S, Bolhuis H, Gruska M, Pfeiffer F, Oesterhelt D. 2015. Fluorescence microscopy visualization of halomucin, a secreted 927 kDa protein surrounding Haloquadratum walsbyi cells. *Front. Microbiol.* 6:249.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res.* 46:D754–D761.

2. Results

2.1. Chapter 1 - Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics

2.1.1. Manuscript

frontiers in Microbiology

PERSPECTIVE
published: 22 October 2021
doi: 10.3389/fmicb.2021.755101

Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics

Valérian Lupo[1,2], Mick Van Vlierberghe[1], Hervé Vanderschuren[3], Frédéric Kerff[2], Denis Baurain[1] and Luc Cornet[1,3]*

[1] InBioS-PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, Liège, Belgium, [2] InBioS, Center for Protein Engineering, University of Liège, Liège, Belgium, [3] Plant Genetics, TERRA Teaching and Research Center, Gembloux Agro-Bio Tech, University of Liège, Liège, Belgium

Contaminating sequences in public genome databases is a pervasive issue with potentially far-reaching consequences. This problem has attracted much attention in the recent literature and many different tools are now available to detect contaminants. Although these methods are based on diverse algorithms that can sometimes produce widely different estimates of the contamination level, the majority of genomic studies rely on a single method of detection, which represents a risk of systematic error. In this work, we used two orthogonal methods to assess the level of contamination among National Center for Biotechnological Information Reference Sequence Database (RefSeq) bacterial genomes. First, we applied the most popular solution, CheckM, which is based on gene markers. We then complemented this approach by a genome-wide method, termed Physeter, which now implements a *k*-folds algorithm to avoid inaccurate detection due to potential contamination of the reference database. We demonstrate that CheckM cannot currently be applied to all available genomes and bacterial groups. While it performed well on the majority of RefSeq genomes, it produced dubious results for 12,326 organisms. Among those, Physeter identified 239 contaminated genomes that had been missed by CheckM. In conclusion, we emphasize the importance of using multiple methods of detection while providing an upgrade of our own detection tool, Physeter, which minimizes incorrect contamination estimates in the context of unavoidably contaminated reference databases.

Keywords: sequencing, assembly, contamination, genomes, databases, NCBI RefSeq, phylogenomics

OPEN ACCESS

Edited by:
Baolei Jia,
Chung-Ang University, South Korea

Reviewed by:
Terence Murphy,
National Center for Biotechnology Information (NLM), United States
Jonathan Badger,
National Cancer Institute (NCI), United States

****Correspondence:***
Denis Baurain
denis.baurain@uliege.be

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *07 August 2021*
Accepted: *04 October 2021*
Published: *22 October 2021*

Citation:
Lupo V, Van Vlierberghe M, Vanderschuren H, Kerff F, Baurain D and Cornet L (2021) Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics. Front. Microbiol. 12:755101. doi: 10.3389/fmicb.2021.755101

INTRODUCTION

Genome contamination, defined here as the accidental inclusion of sequences from other organisms or the misclassification of sequences in public repositories, is a problem having attracted much attention in the recent literature (see for instance, Kahlke and Ralph, 2018; Lu and Salzberg, 2018; Breitwieser et al., 2019; Low et al., 2019). Hence, it is notoriously known that contamination of genome-scale datasets can lead to false conclusions, and such cases have been reported in

Abbreviations: RefSeq, Reference Sequence Database; LCA, Last Common Ancestor; IMG, Integrated Microbial Genome; NCBI, National Center for Biotechnological Information; GTDB, Genome Taxonomy Database.

numerous publications (e.g., Laurin-Lemay et al., 2012; Merchant et al., 2014; Koutsovoulos et al., 2016). Nowadays, many algorithms are available to detect contaminants in complete genomes, e.g., Kraken 2 (Wood et al., 2019), CheckM (Parks et al., 2015), Physeter (Cornet et al., 2018), ConFindR (Low et al., 2019), and BASTA (Kahlke and Ralph, 2018). By studying the phenomenon in Cyanobacteria, we have shown that different methods sometimes yield widely different estimates of the contamination level (Cornet et al., 2018). As this result is explained by differences between the respective algorithms or databases, we argued that the use of multiple methods is the best way to detect contaminant sequences (Cornet et al., 2018). In contrast, relying on a single method of detection, even if very well designed and popular, always bears a danger of systematic error, which can eventually lead to the spread of sequences of incorrect taxonomy into public databases. The objective of this Perspective is to highlight the importance of using multiple methods of detection when assessing contamination in genomic studies.

To this end, we investigated the results of the most cited tool (3,532 citations as of September 2021 according to Google Scholar) in the field of contamination detection, CheckM (Parks et al., 2015). The latter is frequently the only method used in genome-scale studies, for example in the Genome Taxonomy Database (GTDB) project, in which specific genomes are selected as type organisms for the community (Parks et al., 2018). We chose to estimate the contamination level of bacterial genomes from the reference sequence database of the National Center for Biotechnological Information (NCBI), Reference Sequence Database (RefSeq; O'Leary et al., 2016; Haft et al., 2018), not only because this resource is frequently used by many researchers (Nasko et al., 2018), but also because it has been reported to be affected by sequence contamination (Cornet et al., 2018; Breitwieser et al., 2019; Pasolli et al., 2019; Zhu et al., 2019). Here, we first evaluated the contamination level of this database using CheckM, and then compared these estimates, for 12,326 results that we considered as potentially dubious, to those obtained with an upgrade of Physeter, a decontamination tool introduced in Cornet et al. (2018).

CHECKM YIELDS POTENTIALLY DUBIOUS RESULTS FOR 12,326 GENOMES IN NCBI REFSEQ

CheckM estimates the contamination level in a given genome by counting duplications of single-copy and taxon-specific gene markers (Parks et al., 2015). This requires a phylogenetic placement of the genome, based on ribosomal protein genes, in order to determine its taxon and derive the appropriate marker set (Parks et al., 2015). However, for 12,326 bacterial genomes among the 111,088 of RefSeq (Haft et al., 2018), this first step of the algorithm yields a dubious taxon, which has the potential to affect the contamination estimate. In detail, CheckM results were considered dubious for at least one, frequently several, of the four following reasons (**Supplementary Table 1**: https://doi.org/10.6084/m9.figshare.13139810): (1) the CheckM taxon obtained by phylogenetic placement is ambiguous when compared to the NCBI taxon, even if closely related (e.g., same phylum; 9,257 cases), (2) the CheckM taxon is of a too high level (e.g., "bacteria") to be useful in practice (2,967 cases), (3) the CheckM taxon is "incorrect" (e.g., different phylum) with respect to the NCBI taxon or both taxa are uninformative (77 cases), and (4) the estimated contamination level is ≥20% (25 cases), which is the upper tested limit of detection for CheckM (per documentation). In the latter case, CheckM results can be erroneous because its phylogenetic placement is affected by an array of supernumerous ribosomal genes belonging to the contaminants. Owing to these reasons, the current release of CheckM produces reliable estimates for only 14 phyla whereas these are questionable for 38 phyla (**Figure 1**). However, the accuracy of CheckM on the remaining 98,801 genomes of RefSeq has not been investigated here.

PHYSETER AS A SECOND ESTIMATOR OF THE CONTAMINATION LEVEL

We then used Physeter to estimate the contamination level of the 12,326 dubious genomes. Physeter features a MEGAN-like (Huson et al., 2007) Last Common Ancestor (LCA) algorithm that uses DIAMOND blastx (Buchfink et al., 2015) results to compute its estimates. Here, we upgraded its heuristics to overcome the unavoidable presence of contaminated genomes in reference databases. In practice, a sliding window splits the reference database into 10 partitions, and Physeter returns the median contamination level of 10 independent estimations, each one based on 90% of the database. This *k*-fold approach allowed us to identify false positive results only driven by a few contaminated genomes in the reference database (**Figure 2A**). For instance, the assemblies GCF_003612345.1 and GCF_003611835.1 have a low median level of contamination, even if some independent estimations (**Figure 2A**) show a higher level. The opposite is also observed (**Figure 2A**), with some contaminated genomes leading to false negative results (see **Supplementary Additional File 1**). Overall, Physeter minimizes the estimation biases due to overlooked contamination while maintaining the diversity of the reference database (**Supplementary Figure 1**).

TAXONOMIC ERRORS AND RARE GENOMES

According to Physeter, 107 RefSeq genomes (among the 12,326) presented very low levels of the organism expected from the associated NCBI taxon. First, these "taxonomic errors" may correspond to genomes that are misclassified by the NCBI (e.g., GCF_900453015.1). Such misclassifications should also be considered as contamination because misclassified genomes are susceptible to be incorporated in downstream studies under a

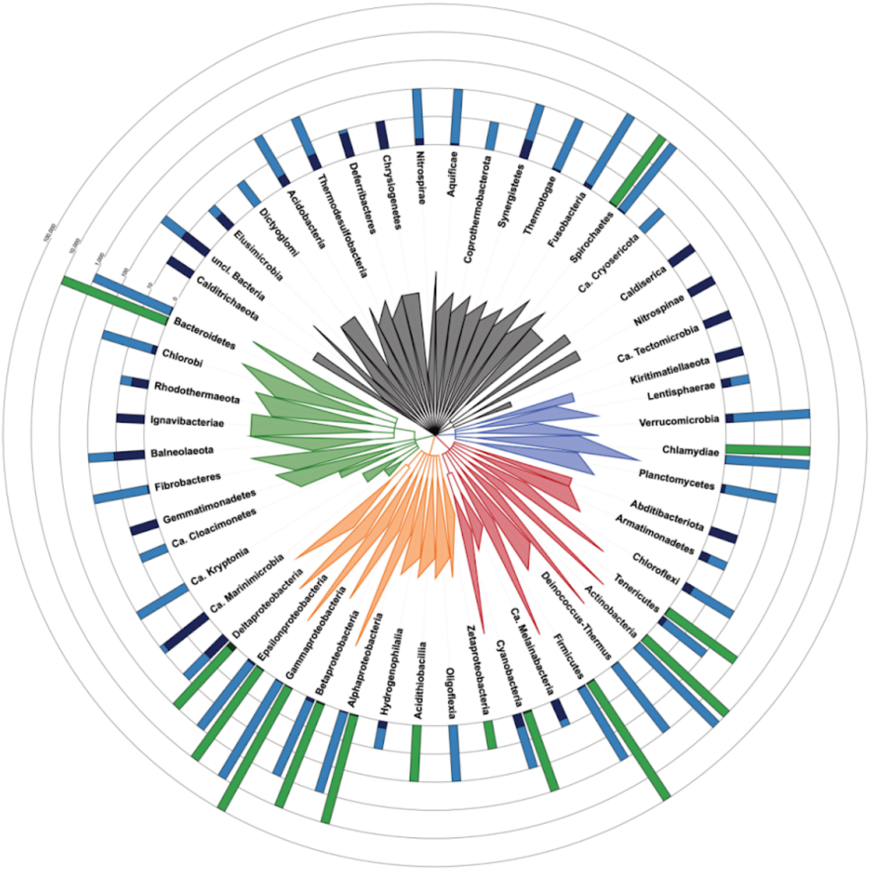


FIGURE 1 | Taxonomic tree of the bacterial domain showing the fraction of contaminated genomes in each phylum with each method. Taxon identifiers of the 111,088 RefSeq bacterial genomes were passed to NCBI Common Tree tools to construct the tree [parameters: (1) include unranked taxa, (2) expand all]. Tree visualization was performed with iTOL and branches were collapsed at the taxonomic levels reported in the tree. Triangles are proportional to taxonomic depth. Proteobacteria are colored in orange, FCB group in green, Terrabacteria in red, PVC group in blue and the other phyla in dark gray. Green barplots are for genomes evaluated with CheckM and blue barplots are for Physeter. The fraction of genomes with a contamination level <5% is shown in a light color whereas those ≥5% are shown in a dark color. The number of genomes evaluated with each method is indicated by the height of the barplot on a ceiled logarithmic scale. For simplicity, the estimates for Ca. Saccharibacteria (2 contaminated and 12 uncontaminated genomes), candidate division NC10 (2 contaminated genomes), Ca. Atribacteria (2 contaminated genomes), and Ca. Bipolaricaulota (1 contaminated genome) are included in unclassified Bacteria. Completely contaminated phyla (e.g., Caldiserica, Nitrospinae, and Kiritimatiellaeota) are generally represented by very few genomes (i.e., one to three genomes). Among the more extensively studied phyla (11 to 37,487 genomes), some appear to be extremely contaminated, such as Balneolaeota, Synergistetes, and Chloroflexi, with, respectively, 54.5, 33.3, 16.9% of contaminated genomes, whereas other phyla are characterized by a very low contamination level, including Cyanobacteria (2.8%), Gammaproteobacteria (0.6%), or Chlamydiae (0.3%).

wrong taxonomy, which could be very damaging to biological conclusions (Laurin-Lemay et al., 2012). Second, taxonomic errors can also stem from genomes that are so contaminated that the sequences of the expected organism are overwhelmed by the foreign sequences (e.g., GCF_003264215.1). Third, some genomes belong to a taxon that is so rare in genome databases that they only match themselves, which is not allowed by the Physeter algorithm and thus leads to low levels of the expected organism (e.g., GCF_000226295.1), including 45 genomes tagged as "unclassified Bacteria" by the NCBI. In practice, distinguishing between the three cases is very difficult. Among the 107 genomes, 65 were left unclassified by CheckM (i.e., identified as "bacteria"

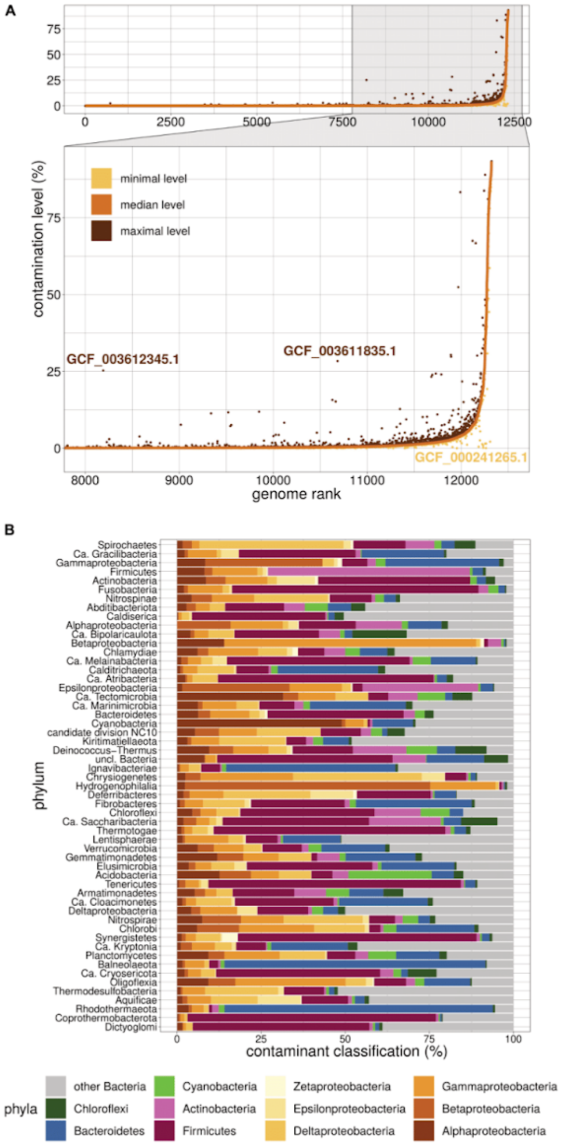


FIGURE 2 | Overview of Physeter properties. **(A)** Distribution of contamination levels assessed by Physeter in *k*-fold mode. Genomes are ranked from the lowest to highest median level of contamination. Median levels are shown in a solid orange line, while minimal and maximal levels are represented as yellow and brown dots, respectively. GCF_003612345.1 and GCF_003611835.1 are examples of genomes having a low median level of contamination with some independent estimations showing a higher contamination level. The opposite case is illustrated with GCF_000241265.1. **(B)** Taxonomic distribution of contaminating sequences within each phylum. The relative contributions of each contaminating phylum were first averaged by genome over all 10 *k*-folds, then these genome-wise averaged values were averaged by tested phylum over all genomes.

or "root") with a low level of contamination (median 1.1%), whereas Physeter found high contamination levels (median 14.6%) for these 65 cases. To deal with those 107 problematic genomes, we re-ran Physeter using the GTDB taxonomy (Parks et al., 2018) as an alternative and let the tool determine the main organism itself, just like CheckM usually does (see **Supplementary Table 1**). In theory, the use of GTDB should help us to discriminate between taxonomic errors and rare genomes, though in practice it does not. This is so because 76 genomes (among the 107) are representative genomes in GTDB, which have been decontaminated based on CheckM results alone. On the other hand, Physeter's auto-detection mode is not compatible with its self-match skipping feature. Therefore we cannot make a decision on these 107 complex cases. The take-home message of this section is that estimating the contamination level in the case of rare genomes or taxonomic errors is very difficult, especially when interconnected tools are used.

THE CASE FOR CORROBORATED ESTIMATES

Based on the recommendations established by the Genomic Standards Consortium (Bowers et al., 2017), we used a threshold of 5% to decide if a genome is contaminated. CheckM and Physeter results can only be compared in the context of this specific cutoff, since the two algorithms are very different and hardly comparable in terms of contaminant percentage. Moreover, while CheckM is based on taxon-specific marker sets, Physeter probes the whole genomes. Nevertheless, the results can be divided into four categories based on the maximum contamination threshold of 5%: (1) both methods identify <5% of contaminants (11,759 genomes), (2) CheckM alone identifies ≥5% of contaminants (384 genomes), (3) Physeter alone identifies ≥5% of contaminants (133 genomes), and (4) both methods identify ≥5% of contaminants (46 genomes). The two methods are thus in agreement for 95.77% of the 12,326 dubious genomes. The discrepancies were expected based on our previous results on Cyanobacteria, where we compared six different detection methods (Cornet et al., 2018). Even if numerically minor, they confirm the importance of using multiple methods of detection when estimating contamination levels. Schematically, the intersection of the methods (i.e., corroboration) increases the certainty that a given genome is contaminated, hence reducing false positives, whereas the union maximizes the power of detection, hence reducing false negatives. The choice of the intersection or of the union is dependent on the goal of study, as both options have their drawbacks, either more false negatives or more false positives, respectively. At this stage, it is difficult to decide "which method is right" between CheckM and Physeter. One way would be to perform a metagenomic binning on the genomes for which they disagree. However, sequencing reads are not publicly available for more than half of these genomes (only 41.3 and 45.1% for category 2 and 3, respectively), and these genomes being lowly contaminated, the foreign bins are too small to be accurately classified by any tool.

Physeter presents the advantage of labeling the individual sequences and thus offers the possibility to explore the taxonomy of the contaminants. These are very diversified, with a median of 45 different contaminant phyla per phylum (over the 10 *k*-fold replicates). Firmicutes appear to be the major contaminant of various phyla (**Figure 2B**), such as Tenericutes (75.1% of the contaminant sequences), Fusobacteria (73.3%), Synergistetes (70.9%), or Thermotogae (68.9%). Reciprocally, the major contaminant of Firmicutes genomes are Actinobacteria (60.1%). Biological traits like sheath thickness or the abundance of co-living organisms can explain the nature of the contaminants and the fact that some taxa have a higher propensity for contamination, the latter being also affected by uneven sampling of lifestyles in RefSeq (e.g., lots of clinical samples).

DISCUSSION

In this study, we have only looked at bacterial genomes contaminated by other bacterial sequences. However, the situation can be more complex, for instance in metagenomic samples including small eukaryotes where contaminations can remain unnoticed by most algorithms to the exception of Kraken (Wood et al., 2019), BlobToolKit (Challis et al., 2020), a workflow developed for eukaryotes, and Physeter (Cornet et al., 2018). As a case in point, we provide a protocol to construct a database containing representative genomes from the three domains of life and study contamination in complex samples with Physeter (see **Supplementary Additional File 2**). Based on the results of the present study, even the most curated database publicly available, RefSeq, includes 1,395 significantly (≥5%) contaminated genomes (considering the union of CheckM and Physeter results), which translates to 1.25% of the genomes. This low percentage should not be considered as a comforting result because even a single contaminated genome can lead to false interpretations (Bemm et al., 2016). Perhaps more critical, since nearly all contamination detection tools use databases derived from public repositories as references [RefSeq (Haft et al., 2018) for Kraken (Wood et al., 2019), Integrated Microbial Genomes (IMG; Markowitz et al., 2012) for CheckM (Parks et al., 2015), Ensembl (Hubbard et al., 2002) for the first version of Physeter (Cornet et al., 2018), RefSeq (Haft et al., 2018) for ConFindR (Low et al., 2019), RefSeq (Haft et al., 2018) for BASTA (Kahlke and Ralph, 2018)], the reliability of the detection hinges on the quality of these public databases. To our knowledge, Physeter is the only software able to robustly detect contaminations at a genome-wide scale when using a moderately contaminated database as a reference.

Considering the low level of contaminated genomes in RefSeq, one could conclude that the risk to include contaminants in a study, due to reliance on a single method of detection, is also low. Nevertheless, researchers are by essence more interested in particularities than by generalities, and even small amounts of contaminants have the potential to lead to exciting but false conclusions. That is why we argue that a "second opinion" should be considered when searching for contaminating sequences, especially as long as genome reference databases are not completely devoid of contamination (Pasolli et al., 2019; Zhu et al., 2019).

METHODS

111,088 genomes were downloaded from RefSeq on the 9th of March 2019, regardless of their sequencing status. These genomes were analyzed with CheckM using the typical automatic workflow option lineag_wf. CheckM automatically places the queried genomes in a reference tree through the concatenation of predicted ribosomal proteins. The completeness and contamination levels are then estimated by searching for lineage specific marker genes provided with the software. CheckM uses 5,656 genomes from a decontaminated version of IMG dating from 2015 (Parks et al., 2015).

For Physeter analyses, we first built a DIAMOND blastx database corresponding to the 177,288 genomes of the Kraken2 database (Wood et al., 2019; **Supplementary Table 2**: https://doi.org/10.6084/m9.figshare.13139819). This very comprehensive database is composed for a large part of RefSeq genomes, after curation by the authors (Wood et al., 2019). Yet it only includes bacterial genomes, which prevents us from analyzing archaeal genomes here. Moreover, CheckM indicated that 685 genomes of this database are contaminated, which motivated our choice of a leave-one-out approach. The queried genomes were then split into pseudo-reads of 250 nt, BLASTed against the protein database, and labeled by computing the LCA of each pseudo-read based on its best hits (excluding self-matches), provided that they yielded a bit-score $\geq$ 80 and within 95% of the bit-score of the first hit (MEGAN-like algorithm; Huson et al., 2007). As in Cornet et al. (2018), we chose to set the minimal number of best hits to 1 for computing LCAs. For the 107 misclassified genomes on the NCBI, we ran Physeter using a local mirror of the GTDB taxonomy (Parks et al., 2018; release 202) instead of the NCBI Taxonomy. Taxa were attributed through the "auto-detect" option and the "labeller" was constructed using all available GTDB phyla, except for Proteobacteria, which were split into their constituting classes instead.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://doi.org/10.6084/m9.figshare.13139810.v2; https://doi.org/10.6084/m9.figshare.13139819.v1; https://metacpan.org/dist/Bio-MUST-Apps-Physeter.

AUTHOR CONTRIBUTIONS

LC and DB conceived the study. LC, VL, MV, and DB developed Physeter. VL performed all analyses and drew the figures. LC supervised the study. LC, VL, and DB wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

VL and MV are supported by FRIA fellowships of the Belgian National Fund for Scientific Research (F.R.S.-FNRS). FK is a Research Associate employed by the F.R.S.-FNRS. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

ACKNOWLEDGMENTS

The authors warmly thank David Colignon (ULiège) for his help with the optimal use of the computing clusters. The authors are also grateful to Raphaël R. Léonard for his help with genome download.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.755101/full#supplementary-material

REFERENCES

Bemm, F., Weiß, C. L., Schultz, J., and Förster, F. (2016). Genome of a tardigrade: horizontal gene transfer or bacterial contamination? *Proc. Natl. Acad. Sci. U. S. A.* 113, E3054–E3056. doi: 10.1073/pnas.1525116113

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35, 725–731. doi: 10.1038/nbt.3893

Breitwieser, F. P., Pertea, M., Zimin, A. V., and Salzberg, S. L. (2019). Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* 29, 954–960. doi: 10.1101/gr.245373.118

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Challis, R., Richards, E., Rajan, J., Cochrane, G., and Blaxter, M. (2020). BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3* 10, 1361–1374. doi: 10.1534/g3.119.400908

Cornet, L., Meunier, L., Vlierberghe, M. V., Léonard, R. R., Durieu, B., Lara, Y., et al. (2018). Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLoS One* 13:e0200323. doi: 10.1371/journal.pone.0200323

Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., et al. (2018). RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* 46, D851–D860. doi: 10.1093/nar/gkx1068

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., et al. (2002). The Ensembl genome database project. *Nucleic Acids Res.* 30, 38–41. doi: 10.1093/nar/30.1.38

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/gr.5969107

Kahlke, T., and Ralph, P. J. (2018). BASTA – Taxonomic classification of sequences and sequence bins using last common ancestor estimations. *Methods Ecol. Evol.* 10, 100–103. doi: 10.1111/2041-210X.13095

Koutsovoulos, G., Kumar, S., Laetsch, D. R., Stevens, L., Daub, J., Conlon, C., et al. (2016). No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. *Proc. Natl. Acad. Sci. U. S. A.* 113, 5053–5058. doi: 10.1073/pnas.1600338113

Laurin-Lemay, S., Brinkmann, H., and Philippe, H. (2012). Origin of land plants revisited in the light of sequence contamination and missing data. *Curr. Biol.* 22, R593–R594. doi: 10.1016/j.cub.2012.06.013

Low, A. J., Koziol, A. G., Manninger, P. A., Blais, B., and Carrillo, C. D. (2019). ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. *PeerJ* 7:e6995. doi: 10.7717/peerj.6995

Lu, J., and Salzberg, S. L. (2018). Removing contaminants from databases of draft genomes. *PLoS Comput. Biol.* 14:e1006277. doi: 10.1371/journal.pcbi.1006277

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., et al. (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* 40, D115–D122. doi: 10.1093/nar/gkr1044

Merchant, S., Wood, D. E., and Salzberg, S. L. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2:e675. doi: 10.7717/peerj.675

Nasko, D. J., Koren, S., Phillippy, A. M., and Treangen, T. J. (2018). RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.* 19:165. doi: 10.1186/s13059-018-1554-6

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

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. doi: 10.1038/nbt.4229

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* 176, 649–662.e20. doi: 10.1016/j.cell.2019.01.001

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *bioRxiv* [preprint]. doi: 10.1101/762302

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.* 10:5477. doi: 10.1038/s41467-019-13443-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

2.1.2. Supplementary Material

Supplementary Figures

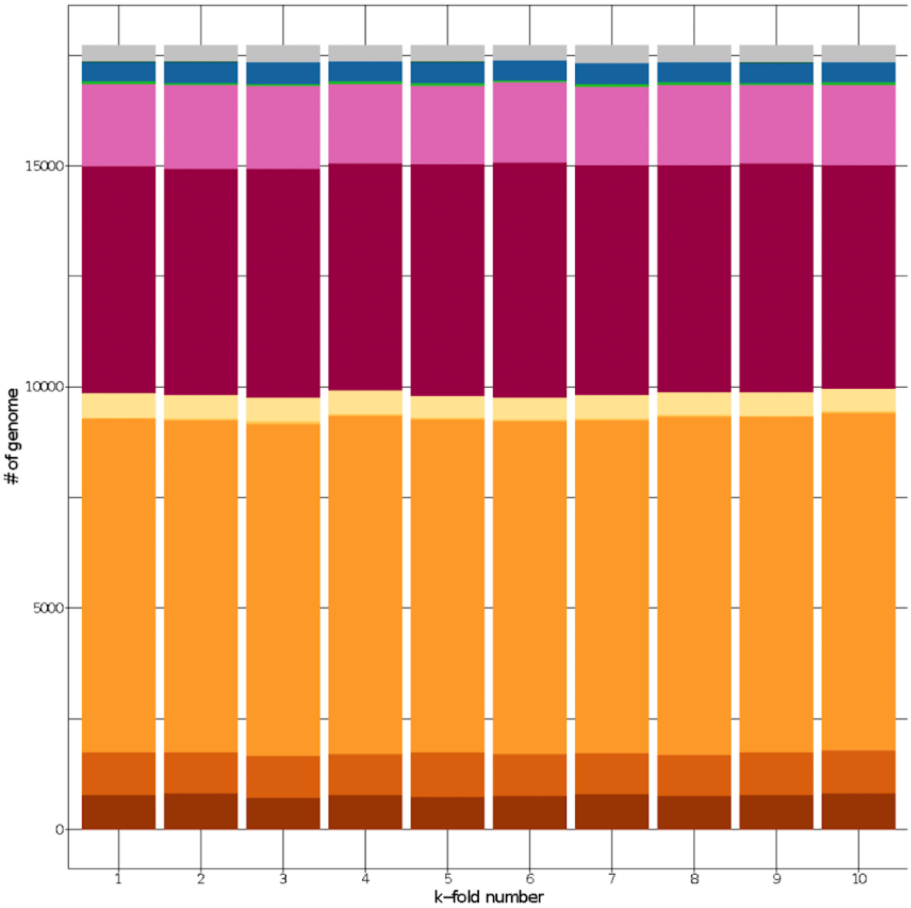


Fig S1. **Taxonomic diversity of 10 equal-sized partitions randomly generated by Physeter in an examplative *k*-fold analysis**. Each of these subsets is left out in turn so as to estimate the sensitivity of the results towards database composition.

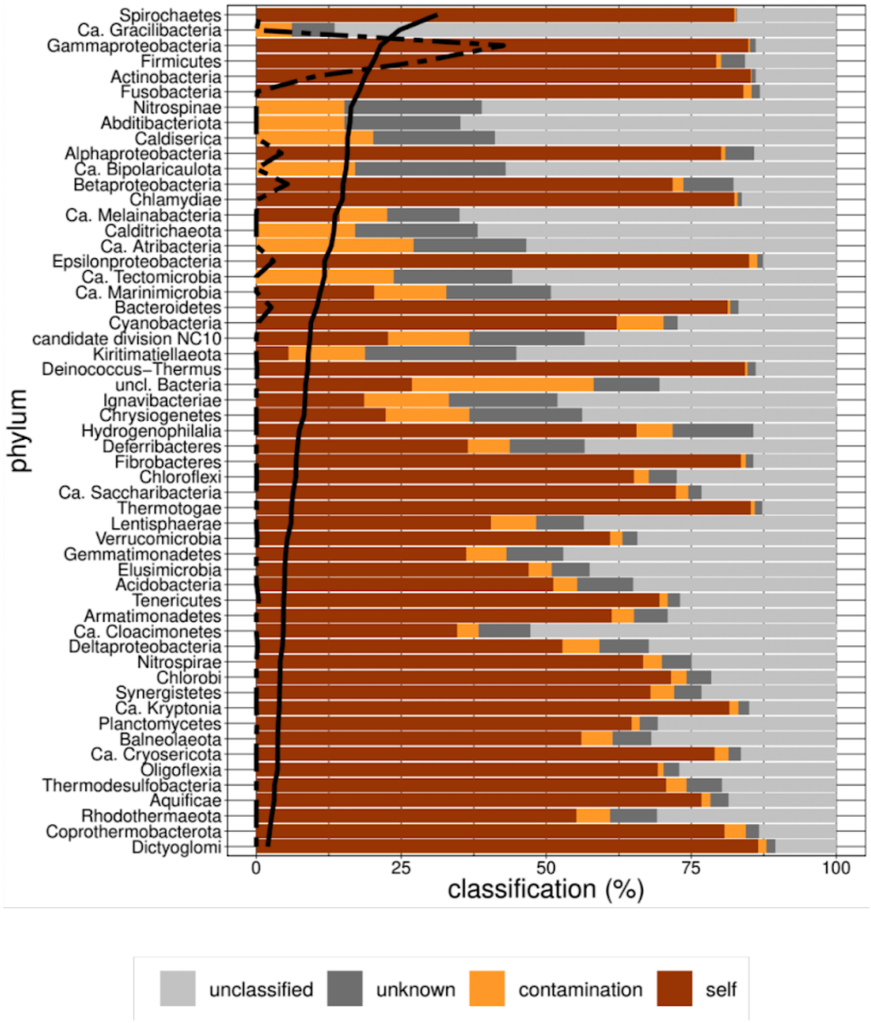


Fig S2. **Relative distribution of the four classification categories for each phylum**. Percentages are the average of the respective median values computed in *k*-fold mode for each genome within the phylum. The average median value of the average number of hits used to compute a LCA is represented with a solid black line, which is further used to rank the phyla. The dashed black line shows the fraction of genomes represented by each phylum in the database. Such a fraction computed on cumulative lengths (in amino acids) rather than numbers of genomes (counts) would have yielded a very similar curve.

Additional file 1: Technical explanation of Physeter algorithm

Background

Physeter is a command-line tool that uses DIAMOND blastx (Buchfink et al., 2015) reports to assess the level of contamination of a genome assembly (see manual). To ensure maximum sensitivity, it is advised to split the genome to analyze into pseudo-reads of 250-250 nt (Cornet et al., 2018). Like BASTA (Kahlke & Ralph, 2018), it is based on a Last Common Ancestor (LCA) algorithm to assign a taxonomy to each pseudo-read of the genome. Its algorithm consists in accumulating the hits for each query, (2) assigning a lineage to these hits (based on NCBI Taxonomy), (3) computing a LCA that is then used to taxonomically annotate the pseudo-read and (4) classifying the pseudo-read to compute the ratio of contaminated pseudo-reads in the genome. Our LCA algorithm was also inspired by MEGAN (Huson et al., 2007) for hit accumulation, using a bit score threshold expressed as a percentage of the highest bit score of the current pseudo-read.

Algorithm description

Classic mode

The first step of the Physeter algorithm is to parse the DIAMOND blastx report (where queries are pseudo-reads). For each query, Physeter evaluates whether the first hit passes all the specified thresholds (i.e., length, percent of identity, bit score). If so, it starts accumulating hits according to the --tax-min-hits and --tax-max-hits (see manual) thresholds, i.e., minimum and maximum numbers of hits to accumulate in order to compute the LCA. If --tax-min-hits and --tax-max-hits are both set to 1, then Physeter only uses the best hit to assign taxonomy (BEST HIT MODE). The highest bit score among the hits is used to initialize the bit score threshold. This bit score threshold itself is computed by multiplying the highest bit score by the --tax-score-mul (MEGAN-LIKE MODE). During hit accumulation, if the genome has a NCBI GCA/GCF accession (see manual), the hits corresponding to the organism are ignored. In contrast, this is not possible when using either the --exp-tax or --auto-detect option (designed for custom genome assemblies; see below). Hit accumulation can stop for different reasons: 1) minimum of hit is not reached, thus no LCA is computed and Physeter goes to the next query, 2) maximum of hit is reached, if minimum of hit is reached too, therefor LCA is computed, if not, no LCA is computed, 3) the bit score of a hit is lower than the bit score threshold and like point 2) LCA is compute or not either if minimum of hit is reached or not. Then, Physeter uses a local mirror of the NCBI Taxonomy to fetch the lineages of all accumulated hits in order to compute the LCA. The optional --tax-min-lca-freq threshold can be applied to discard minor lineages incongruent with those encountered in majority. This threshold works at any taxonomic level, which makes it very efficient at determining the most precise LCA. For diagnostic purposes, Physeter keeps track of the LCA assigned to each query (or lack of) and the number of hits used in the taxonomical computation.

The second step is to determine if the assigned taxonomy (LCAs) of the pseudo-reads corresponds or not to the organism taxonomy. The organism taxonomy can be determined in three ways: 1) based on its NCBI GCA/GCF accession in the case of public genome assemblies, 2) user-specified using the --exp-tax option for custom genomes for which one approximately knows the taxonomy, 3) through auto-detection (--auto-detect option) based on the most abundant LCA identified during the first step. In contrast to CheckM (Parks et al., 2015), which uses ribosomal phylogenetic placement followed by the detection of clade-specific sets of about hundreds marker genes to evaluate its contamination level, Physeter considers the entire set of pseudo-reads of the genome under analysis. To this end, the organism taxonomy and the pseudo-read LCAs are remapped at a higher taxonomic level, using a taxonomic labeller defined as a list of high-ranking NCBI taxa. For some ambiguous taxa with the same label at different taxonomic levels (e.g., Actinobacteria), the --greedy-taxa option can be used to decide which level to use when remapping pseudo-reads (see manual). After labelling, pseudo-reads are classified into one of four categories: 1) 'self' if the labels are identical between the organism and the pseudo-read, 2) 'contaminated' if the labels are different, 3) 'unknown' if no label could be assigned to the pseudo-read (e.g., if the LCA is too high-ranking, such as 'cellular organisms', 4) 'unclassified' if no LCA could be computed due to a lack of hits to the reference DIAMOND database. Interestingly, the sensitivity of Physeter is not affected by the number of hits used to compute LCAs. Indeed, while the unclassified fraction increases with the number of hits, some groups with very high fractions of classified sequences are also among those with the highest numbers of hits (Fig. S1, e.g., Firmicutes). The eight taxonomic groups with no genome fraction identified as "self" (Fig. S1, e.g., Nitrospinae) are rare phyla represented by a maximum of two genomes (five with one and three with two genomes). Since our approach does not take into account self hits, zero to one reference genomes are available for classification of these organisms, which leads to the observed lack of "self". As expected, the abundance of genomes from a given phylum positively influences the number of hits, but only moderately. Hence, even if highly represented phyla attract more hits, some less represented phyla (e.g., Spirochaetes) are also characterized by high numbers of hits (Fig. S1).

k-fold mode

The *k*-fold mode allows users to systematically ignore subsets of the DIAMOND database, so as to identify the reference genomes leading to false detection. The list of NCBI GCA/GCF accessions used to construct the database is passed to Physeter using --kfold option. Then, accessions are shuffled and split into 10 equal-sized subsets. The functioning of the algorithm described in *Classic mode* stays the same except that Physeter runs 10 times and, for each run, hits that belong to the subset to be ignored are skipped during hit accumulation. Finally, non-parametric statistics are computed for each sequence category.

Dataset

The DIAMOND reference database used in this study is based on the Kraken2 database (Wood et al., 2019) and contains 177,288 genomes. The 111,907 tested genomes were downloaded from RefSeq on the 9th of March 2019.

To download and run Physeter, see https://metacpan.org/dist/Bio-MUST-Apps-Physeter.

References

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Cornet, L., Meunier, L., Vlierberghe, M. V., Léonard, R. R., Durieu, B., Lara, Y., Misztak, A., Sirjacobs, D., Javaux, E. J., Philippe, H., Wilmotte, A., & Baurain, D. (2018). Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLOS ONE*, *13*(7), e0200323. https://doi.org/10.1371/journal.pone.0200323

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386. https://doi.org/10.1101/gr.5969107

Kahlke, T., & Ralph, P. J. (2018). BASTA – Taxonomic classification of sequences and sequence bins using last common ancestor estimations. *Methods in Ecology and Evolution*, *10*(1), 100–103. https://doi.org/10.1111/2041-210X.13095

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *BioRxiv*, 762302. https://doi.org/10.1101/762302

Additional file 2: Running Physeter on complex samples

Installation

For installation and dependencies, see manual at: https://metacpan.org/dist/Bio-MUST-Apps-Physeter.

Input files

Install a local mirror of the *NCBI Taxonomy* or the *GTDB Taxonomy*.

```
$ setup-taxdir.pl --taxdir=ncbi-taxdump/
$ setup-taxdir.pl --taxdir=gtdb-taxdump/ --source=gtdb
```

Building the DIAMOND database

Get prokaryote proteome download links at: https://figshare.com/articles/dataset/Datasets_for_L_onard_et_al_ToRQuEMaDA_Tool_for_Retrieving_Queried_Eubacteria_Metadata_and_Dereplicating_Assemblies/13238936/2.

Decompress the prokaryote archive file.

```
$ tar -xf tqmd_datasets.tar.gz
```

Download and decompress bacterial and archaeal proteomes.

```
$ for f in `cut -f4
  tqmd_datasets/tables/bacteria-151-tax-links.tsv \
  tqmd_datasets/tables/archaea-86-tax-links.tsv`; do wget \
  ${f}/*protein.faa.gz ; done
$ gunzip *.faa.gz
```

Rename sequence identifiers.

```
$ ls *.faa | perl -nle '($gcf) = m/(GC[AF]\_\d{9}\.\d{1})/ ;
  print "$_\t$gcf"' > file.idm
$ inst-abbr-ids.pl --id-prefix-mapper=file.idm \
  --id-regex=:DEF *.faa
```

Concatenate prokaryote files.

```
$ cat *-abbr.faa > prokaryote.faa
$ rm -f GCF*.faa
```

Download eukaryote proteomes at: https://doi.org/10.6084/m9.figshare.13573424.

Decompress eukaryote proteome files.

```
$ tar -xf Data_set_2.tar.gz
```

Rename sequence identifiers.

```
$ cd Data_set_2/
$ perl -i.bak -nle 's/>.*_(\d+)@(.*)$/>\1|\2/; print' *.faa
```

Concatenate prokaryote and eukaryote files.

```
$ cat Data_set_2/*.faa prokaryote.faa > database.faa
```

Build the `DIAMOND` database.

```
$ diamond makedb --in database.faa -d database
```

Running DIAMOND BLASTX

Before running `DIAMOND`, you have to transform the prokaryotic genome files you want to assess into pseudo-read `FASTA` files. Use `inst-split-fas.pl` from the Bio::MUST::Core distribution to do so. In the example below, the genome will be split into 250-base long pseudo-read sequences without overlap. If your genome has a NCBI GCA/GCF accession, name your `outfile` `assembly_accession.fasta` (e.g., `GCF_000006605.1.fasta`).

```
$ inst-split-seq.pl genome.fasta --out=-split
```

Then run `DIAMOND` as follows. Like the `FASTA` file, name your `BLASTX` report as `assembly_accession.blastx` (e.g., `GCF_000006605.1.blastx`). If your genome file does not have a NCBI GCA/GCF accession, both the `FASTA` file and the `BLASTX` report must have the same basename. The `-f tab` option of DIAMOND will generate a tab-separated file corresponding to the `-outfmt 6` of regular *NBCI-BLAST+*. You can adapt the `-p 10` option (number of CPU threads) to suit your system.

```
$ diamond blastx -d database -q split-genome.fasta -o \
  split-genome.blastx -t ./temp -k 50 -e 1e-10 -f tab -p 10
```

Taxonomic labeller

A taxonomic labeler is used by `physeter.pl` to determine at which taxonomic level you consider a pseudo-read sequence as a contaminant. Note that you have to adjust your labeler depending on the used taxonomy. See examples below:

```
$ head phylum-taxa.idl

unclassified Bacteria
unclassified Archaea
Abditibacteriota
```

```
Acidithiobacillia
Acidobacteria
Actinobacteria
Alphaproteobacteria
Aquificae
Armatimonadetes
Bacteroidetes
```

Command-line options of physeter.pl

Classic mode

Once all input files are correctly prepared, you can simply run `physeter.pl` like this:

```
$ physeter.pl *.blastx --outfile=contam.report \
  --taxdir=ncbi-taxdump/ --taxon-list=phylum-taxa.idl
```

Or using GTDB taxonomy.

```
$ physeter.pl *.blastx --outfile=contam.report \
  --taxdir=gtdb-taxdump/ --taxon-list=phylum-taxa.idl
```

The standard output file of `physeter.pl` is a tab-separated file containing the following sections: (1) organism accession or file name, (2) assigned taxon, (3) % self sequences, (4) % contaminated sequences, (5) % unknown taxon sequences, (6) % unclassified sequences, (7) detail of contaminants, (8) mean number of hits used to classify the pseudo-read sequences.

In addition to the Physeter output file, you can generate for each assayed genome a Kraken-like file, an Anvio-like file, a Krona-compatible file or a LCA (Last Common Ancestor) file, the latter providing the taxonomic affiliation of each pseudo-read.

```
$ physeter.pl *.blastx --outfile=contam.report \
  --taxdir=gtdb-taxdump/ --taxon-list=phylum-taxa.idl \
  --kraken --anvio --krona --lca
```

When your pseudo-read FASTA files are not in the working directory, you can specify their localization using the `--fasta-dir` option.

```
$ physeter.pl *.blastx --outfile=contam.report \
  --fasta-dir=split-fasta/ --taxdir=gtdb-taxdump/ \
  --taxon-list=phylum-taxa.idl
```

If your organism does not have a NCBI GCA/GCF accession but you know approximately its taxonomy, you can specify it with the `--exp-tax` option. Note that the specified taxon must be listed in the file provided through the `--taxon-list` option.

```
$ physeter.pl organism.blastx --exp-tax=Firmicutes \
  --outfile=contam.report --taxdir=gtdb-taxdump/ \
  --taxon-list=phylum-taxa.idl
```

Otherwise, use the `--auto-detect` option.

```
$ physeter.pl organism.blastx --auto-detect \
  --outfile=contam.report --taxdir=gtdb-taxdump/ \
  --taxon-list=phylum-taxa.idl
```

In the basic configuration, `physeter.pl` will assess the contamination status of a pseudo-read sequence using only 1 hit (i.e., *best-hit mode*). If you want to use more than 1 hit (i.e., *MEGAN-like mode*), you can use the `--tax-min-hits` and `--tax-max-hits` options. In the *MEGAN-like mode*, a LCA will be inferred for each pseudo-read sequence.

```
$ physeter.pl *.blastx --outfile=contam.report \
  --taxdir=gtdb-taxdump/ --taxon-list=phylum-taxa.idl \
  --tax-min-hits=2 --tax-max-hits=50
```

You can use `--tax-score-mul` and `--tax-min-lca-freq` options to fine tune LCA inference.

```
$ physeter.pl *.blastx --outfile=contam.report \
  --taxdir=gtdb-taxdump/ --taxon-list=phylum-taxa.idl \
  --tax-min-hits=2 --tax-max-hits=50 \
  --tax-score-mul=0.7 --tax-min-lca-freq=0.85
```

Other options can be applied to filter the BLASTX hits used for contamination assessment. Those are `--tax-min-ident`, `--tax-min-len` and `--tax-min-score`.

K-fold mode

The last functionality of `physeter.pl` is the *k-fold mode*. In this mode, the `DIAMOND` database is randomly split into 10 subsets. Then, `physeter.pl` runs 10 times and, for each run, hits from one of the subsets are ignored. The results of the 10 analyses are written in the standard output file. None of the Kraken-like file, Anvio-like file, Krona-coompatible file and LCA file are available when running in *k-fold mode*.

```
$ physeter.pl *.blastx --outfile=contam.report \
  --taxdir=taxdump/ --taxon-list=phylum-taxa.idl \
  --tax-min-hits=2 --tax-max-hits=50 --k-fold=database.gca
```

The `database.gca` file is the list of all NCBI GCA/GCF accessions of the genomes used to build the `DIAMOND` database.

```
$ grep \> database.faa | cut -f1 -d'|' | cut \
  -c2- | sort -u > database.gca
$ head database.gca

1169474
1169539
1169540
1202447
1255295
127563
130081
1321669
13642
1389228

$ tail database.gca

GCF_900095815.1
GCF_900095855.1
GCF_900105895.1
GCF_900120375.1
GCF_900128725.1
GCF_900128965.1
GCF_900129645.1
GCF_900143135.1
GCF_900155405.1
GCF_900155645.1
```

2.2. Chapter 2 - An Extended Reservoir of Class-D Beta-Lactamases in Non-Clinical Bacterial Strains

2.2.1. Manuscript

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

RESEARCH ARTICLE

An Extended Reservoir of Class-D Beta-Lactamases in Non-Clinical Bacterial Strains

Valérian Lupo,[a,b] Paola Sandra Mercuri,[b] Jean-Marie Frère,[b] Bernard Joris,[b] Moreno Galleni,[b] Denis Baurain,[a] Frédéric Kerff[b]

[a]InBioS-PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, Liège, Belgium
[b]InBioS, Center for Protein Engineering, University of Liège, Liège, Belgium

ABSTRACT Bacterial genes coding for antibiotic resistance represent a major issue in the fight against bacterial pathogens. Among those, genes encoding beta-lactamases target penicillin and related compounds such as carbapenems, which are critical for human health. Beta-lactamases are classified into classes A, B, C, and D, based on their amino acid sequence. Class D enzymes are also known as OXA beta-lactamases, due to the ability of the first enzymes described in this class to hydrolyze oxacillin. While hundreds of class D beta-lactamases with different activity profiles have been isolated from clinical strains, their nomenclature remains very uninformative. In this work, we have carried out a comprehensive survey of a reference database of 80,490 genomes and identified 24,916 OXA-domain containing proteins. These were deduplicated and their representative sequences clustered into 45 non-singleton groups derived from a phylogenetic tree of 1,413 OXA-domain sequences, including five clusters that include the C-terminal domain of the BlaR membrane receptors. Interestingly, 801 known class D beta-lactamases fell into only 18 clusters. To probe the unknown diversity of the class, we selected 10 protein sequences in 10 uncharacterized clusters and studied the activity profile of the corresponding enzymes. A beta-lactamase activity could be detected for seven of them. Three enzymes (OXA-1089, OXA-1090 and OXA-1091) were active against oxacillin and two against imipenem. These results indicate that, as already reported, environmental bacteria constitute a large reservoir of resistance genes that can be transferred to clinical strains, whether through plasmid exchange or hitchhiking with the help of transposase genes.

IMPORTANCE The transmission of genes coding for resistance factors from environmental to nosocomial strains is a major component in the development of bacterial resistance toward antibiotics. Our survey of class D beta-lactamase genes in genomic databases highlighted the high sequence diversity of the enzymes that are able to recognize and/or hydrolyze beta-lactam antibiotics. Among those, we could also identify new beta-lactamases that are able to hydrolyze carbapenems, one of the last resort antibiotic families used in human antimicrobial chemotherapy. Therefore, it can be expected that the use of this antibiotic family will fuel the emergence of new beta-lactamases into clinically relevant strains.

KEYWORDS antimicrobial resistance, beta-lactamase, phylogenetic classification, sequence clustering, carbapenemase, OXA

Beta-lactamases are the main enzymes responsible for the resistance of bacteria to beta-lactams, which are among the most common antibiotics used in the fight against pathogenic bacteria. Even before the structure of penicillin was known, Abraham and Chain (1) described "An enzyme from bacteria able to destroy penicillin" and, in the late 1940s and early 1950s, the staphylococcal beta-lactamase became an important source of clinical problems solved by the introduction of methicillin (2, 3). Later, an ever increasing number of these hydrolases have been identified. These can be classified into four classes based on

Editor Monica Adriana Garcia-Solache, Brown University

Ad Hoc Peer Reviewer Peter Oelschlaeger, Western University of Health Sciences

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

The authors declare no conflict of interest.

Received 26 January 2022
Accepted 20 February 2022

their primary structures. Classes A, C, and D are active-serine enzymes (4) while class B consists of metallo-proteins whose active site usually contains 1 or 2 Zn^{++} ions (5, 6).

Beta-lactamases of classes A and D exhibit a very high diversity of amino acid (AA) sequences, with only a very little number of conserved residues within each class (e.g., 29 residues are conserved within class-A beta-lactamases) (7). It is nearly impossible to establish clear relationships between AA sequences and the ability to hydrolyze specific classes of beta-lactam antibiotics. Indeed, it is well known that a single mutation can alter this activity profile in a significant manner (8, 9). Moreover, the literature contains numerous disagreements and errors concerning the kinetic parameters of various enzymes (10). This is probably in part because these parameters are often determined under different experimental conditions and the studied enzymes are not always pure. In consequence, even though clinicians are more interested in specificity profiles, the AA sequences remain the primary tool for proposing a classification of beta-lactamases, as in the case of the Beta-Lactamase Database (BLDB; http://www.bldb.eu/) (11). Concerning class D beta-lactamases, the situation is complicated by the fact that these enzymes can dimerize, which sometimes modifies the activity (12) and that carboxylation of the first conserved motif lysine also increases the activity in most cases (13). Inversely, loss of CO_2 during turn-over of the substrate results in "substrate-induced inactivation," a phenomenon already observed by Ledent et al. (14).

The first two identified class D beta-lactamases exhibited a number of features that differed from those of nearly all beta-lactamases known at the time, notably the ability to efficiently hydrolyze oxacillin and other isoxazolyl penicillins. For this reason, they were named OXA-1 and OXA-2. Unfortunately, it was then decided to name the further class D beta-lactamases homologs "OXA" plus sequence (i.e., increasing) number that follows the chronological order of identification (15). This was sometimes done in spite of a sequence identity below 30% and/or (10) more similarity with the BlaR receptor than with other class D beta-lactamases (16). Class D beta-lactamases were first identified as plasmid-encoded proteins but the corresponding genes were later found to reside on the bacterial chromosome too (10).

Similarity searches using the OXA-2 AA sequence as a query revealed homologous primary structures of unknown function or without true beta-lactamase activity, such as YbxI/BSD-1 in *Bacillus subtilis* (17, 18), or even devoid of any beta-lactamase activity, such as the C-terminal domain (CTD) of the BlaR penicillin-receptor involved in the induction of a class A beta-lactamase in *Bacillus licheniformis* and *Staphylococcus aureus* (19). In the present study, proteins containing a class-D beta-lactamase domain will be further referred to as the "OXA-domain family." Among those, "DBL" will be reserved to demonstrably active class-D beta-lactamases, while characterized class-D beta-lactamase homologs of low activity or with a different function will be termed "pseudo-DBL" proteins. Finally, "DBL-homolog" proteins will define the union of DBL, pseudo-DBL proteins, and other homologs not yet characterized.

It is clear that our present knowledge of the OXA-domain family is biased toward clinically relevant DBLs. The analysis of whole genome sequences of isolated bacteria and metagenome-assembled genomes highlighted that non-pathogenic and environmental bacteria can also harbor beta-lactamase-encoding genes, and thus may behave as reservoirs of emerging new resistance genes identified in nosocomial strains (20, 21). It is likely that these bacteria, which, in many cases, were never exposed to synthetic or semi-synthetic beta-lactams used in human health care or animal husbandry, can encounter other beta-lactam-producing microorganisms in their natural environment and, over the ages, have acquired beta-lactamase genes in their "struggle for life" (22). A significant example could be the carbapenems that can be produced by some *Streptomyces* species (23), probably resulting in the appearance of the carbapenemases that were later transferred to clinical strains (24, 25). The large heterogeneity of the resistance gene repertoire present in bacteria challenges the efficiency of antimicrobial chemotherapy. It also underlines the need to develop new analytical methods allowing a clear and rapid identification of potential new resistance pathways including enzymes that can inactivate both old and new antibiotics.

The goals of the present article were to explore genomic databases to discover how widespread the class D beta-lactamase gene and its homologs were throughout the microbial world and to propose a sequence-based classification of the members of the OXA-domain family derived from their phylogenetic relationships. Starting from 80,490 genomes, we identified a total of 24,916 DBLs and DBL-homolog sequences, which we classified into 64 clusters of proteins. Furthermore, we synthesized and expressed 10 gene sequences sampled from 10 clusters devoid of characterized members and conducted a survey of their activity. This revealed that three of them had an oxacillinase activity, including two able to hydrolyze imipenem, reminding us how environmental bacteria represent an enormous reservoir of resistance factors that can be transferred to clinical strains.

RESULTS

Enlarging the OXA-domain family taxonomic distribution. According to the BLDB (as of July 2019), the 810 described DBL-homologs (including both DBL and pseudo-DBL proteins) have been isolated from bacteria belonging to five different phyla: Proteobacteria (583 Sequences), Spirochaetes (14), Firmicutes (9), Bacteroidetes (1), Fusobacteria (1) and also from some marine metagenomes (2). Two-hundred sequences have no source organism and are all plasmid-encoded. Most of these DBL-homologs are found in Proteobacteria, essentially in the genera *Acinetobacter* (411 sequences) and *Campylobacter* (91), which are part of the Gammaproteobacteria and Epsilonproteobacteria, respectively. Some are also found in Betaproteobacteria (41) but not in the other Proteobacteria classes.

A HMM profile constructed from an alignment of 470 DBL from NCBI Pathogen Detection server allowed us to identify 24,916 OXA-domain family AA sequences distributed across 20,342 organisms (on a total of 80,490 screened genome assemblies found in NCBI RefSeq). Nearly all those organisms (99.4%) belonged to the aforementioned five phyla, whereas the small remaining fraction (0.6%) came from eight additional bacterial phyla: Cyanobacteria (65 Sequences), Actinobacteria (36), Chlorobi (10), Chlamydiae (6), Verrucomicrobia (6), Chloroflexi (2), Balneolaeota (1), and planctomycetes (1). Moreover, some sequences were identified in additional classes of proteobacteria: Alphaproteobacteria (Holosporales) and *Deltaproteobacteria* (Desulfovibrionales). In contrast, no sequences of the OXA-domain family were found in Archaea. In this work, we wanted to characterize the protein sequences themselves and, to do so, we deduplicated the 24,916 sequences and observed that they represented only 3,510 unique sequences (i.e., 100% identical at the AA level), indicating that many of them were multispecies enzymes. Indeed, it is known that the NCBI RefSeq database is unevenly biased toward clinical strains (26). Hence, 3459 of the unique sequences (98.5%) were found in several species of the same genus (e.g., WP_001046004.1 was found in 952 *Acinetobacter* species) while 51 unique sequences (1.5%) were found in more than one genus. These results show that the redundancy is mostly due to the number of species in NCBI RefSeq belonging to the same genus. In a second step, these 3,510 unique sequences were deduplicated at a global identity level of 95%, and the 1,413 resulting sequences (hereafter termed "representative" sequences) were used to infer a phylogenetic tree (see Materials and Methods).

OXA-domain family proteins include BlaR homologs. A distribution of sequence length showed that the 24,916 OXA-domain family sequences formed three populations, one shorter than 350 AAs with an average size of 271 AAs (typical DBL length), one longer than 550 AAs with an average size of 587 AAs (typical BlaR membrane receptor length) and one intermediate-length population with an average size of 449 AAs (Fig. 1a). Mapping sequence length onto the tree revealed that representative sequences of intermediate length are scarce (five sequences) and not clustered, whereas long sequences do cluster in two distinct groups (Fig. 1b). A sequence similarity analysis showed that three of the five intermediate-length sequences are actually

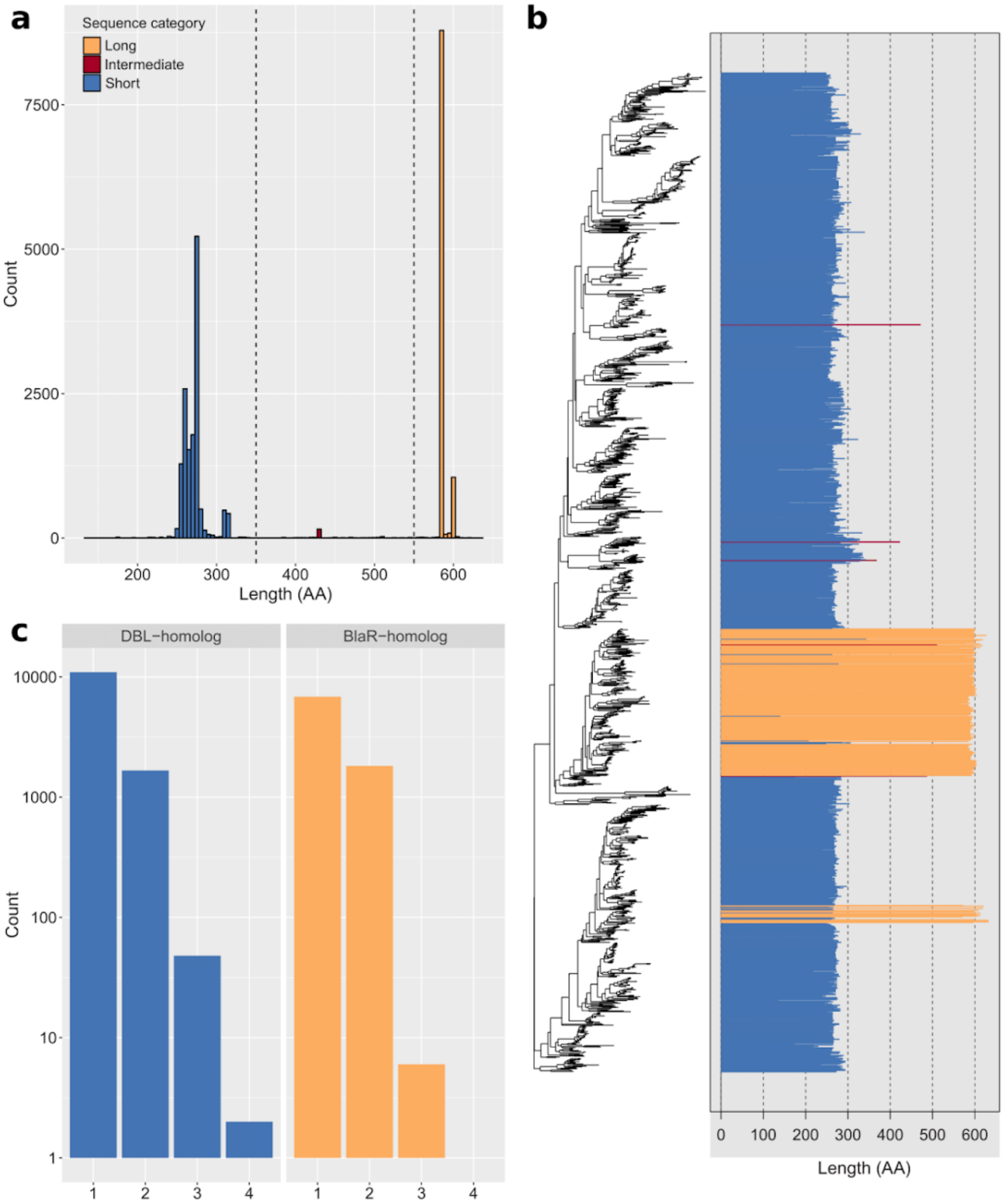


FIG 1 Classification of OXA-domain family protein sequences as DBL-homologs or BlaR-homologs. (a) Length distribution of the 24,916 OXA-domain family protein sequences. Sequences shorter than 350 AAs are colored in blue, sequences longer than 550 AAs are in orange, while sequences between 350 and 550 AAs are in red. (b) Length distribution of the representative sequences mapped onto the phylogenetic tree. The tree was constructed from a matrix of 1,413 representative sequences × 188 unambiguously aligned AAs using RAxML under the LG+F+G4 model. (c) Distribution of the number of DBL-homolog and BlaR-homolog sequences per organism. Blue bar plots represent DBL-homolog sequences while orange bar plots represent BlaR-homolog sequences. The *y* axis is in $\log_{10}$ units.

DBL-homologs while two are BlaR homologs. Regarding the two groups of long sequences, the larger one is formed of sequences found in Firmicutes, with a majority in *Staphylococcus*, *Clostridioides*, and *Bacillus*. According to the annotation results, these sequences are actually BlaR homologs. The second group contains sequences found in Oxalobacteraceae (Betaproteobacteria) and annotation results at first showed no close similarity with DBL nor BlaR. However, detailed *in silico* functional analysis (InterProScan and pepwindowall; Data set S1) eventually revealed that 14 of these representative sequences indeed have a class D active site and the BlaR1 peptidase M56 domain, whereas three have both a class D active site and a class C beta-lactamase active site, like in the LRA-13 fusion enzyme (20), but these exhibit a low sequence identity to the latter (around 60%). To facilitate subsequent discussion, the three intermediate-length DBL homologues and the three OXA-class C fusion proteins were considered as DBL-homologs, whereas the two intermediate-length sequences more similar to BlaR and the two groups of long sequences were considered as BlaR-homologs.

In Firmicutes, beside the 10,496 BlaR-homologs, we also found 1,383 DBL-homologs. According to the annotation results, 374 are homologous to low-activity pseudo-DBL proteins found in *Bacillus* (17, 18) and 956 are homologous to the two intrinsic pseudo-DBL (CDD) of *Clostridium difficile* (27).

In general, surveyed bacteria possess only either one DBL-homolog protein (9,964 strains) or one BlaR-homolog protein (5,874 strains). In 1,665 and 1,813 strains, we found two DBL-homologs or two BlaR-homologs, respectively, and rarely more than two DBL-homologs (27) or BlaR-homologs (5). In addition, 963 strains simultaneously possess one DBL-homolog and one BlaR-homolog, while 10 strains show more than one DBL-homolog and one BlaR-homolog or the opposite (Fig. 1c). Interestingly, strains that harbor more than one DBL-homolog (ignoring BlaR homologs) mostly belong to Pseudomonadales, and more specifically the genera *Acinetobacter* and *Pseudomonas*.

Gene genetic context. Among the 24,916 DBL-homolog and BlaR-homolog protein sequences initially identified, only 23,833 corresponding genes (found on 23,093 contigs) could actually be fetched from complete genomes. Three reasons explain the 1,083 missing sequences: (i) the genome has been suppressed during the study, (ii) the sequence has been suppressed or removed at the submitter's request and could not be found in the genome annotation (gff) file, and (iii) no link between the protein and any gene exists in the NCBI. According to the NCBI annotation pipeline, the contigs are classified "chromosome" in 960 cases, "plasmid" in 273 cases and "genomic" in 21,860 cases. This rather uninformative "genomic" classification led us to predict the genetic context of each OXA-domain family protein sequence using the dedicated PlasFlow pipeline. With this strategy, 15,515 contigs were classified as "chromosome" (67.2%), 5,660 as "plasmid" (24.5%), whereas 1,918 remained unclassified. These unclassified contigs correspond to 9.20% of the DBL-homolog genes (1,327 cases) and 5.79% of the BlaR-homolog genes (608 cases). In addition, 1,177 contigs were congruently classified (either as chromosome or plasmid) by both pipelines, and only four had different labels, thereby confirming the accuracy of PlasFlow for contig classification. DBL-homolog and BlaR-homolog genes are mostly chromosome-encoded (Fig. 2a), with 10,078 (69.89% of DBL-homolog genes) and 6,153 (58.62% of BlaR-homolog genes) cases, respectively, whereas the genes are plasmid-encoded in 2,159 and 3,508 cases, respectively. Thus, 14.97% of DBL-homolog genes and 33.42% of BlaR-homolog genes lie on a plasmid.

The majority of bacteria carrying a DBL-homolog gene on a plasmid belong to six genera of Gammaproteobacteria: *Acinetobacter* (724), *Klebsiella* (590), *Escherichia* (212), *Shigella* (200), *Enterobacter* (196), and *Pseudomonas* (85). The remaining plasmid-encoded sequences are distributed across the other classes of Proteobacteria (Alpha- [35], Beta- [46], and Gammaproteobacteria [61]), while a few can also be found in Firmicutes (6), Cyanobacteria (3), and Bacteroidetes (1).

To assess the transfer potential of DBL-homolog and BlaR-homolog genes, and therefore the propensity of emergence of a new resistance, we looked for transposase genes in the vicinity of these genes (Fig. 2b). We noticed that DBL-homolog and BlaR-homolog

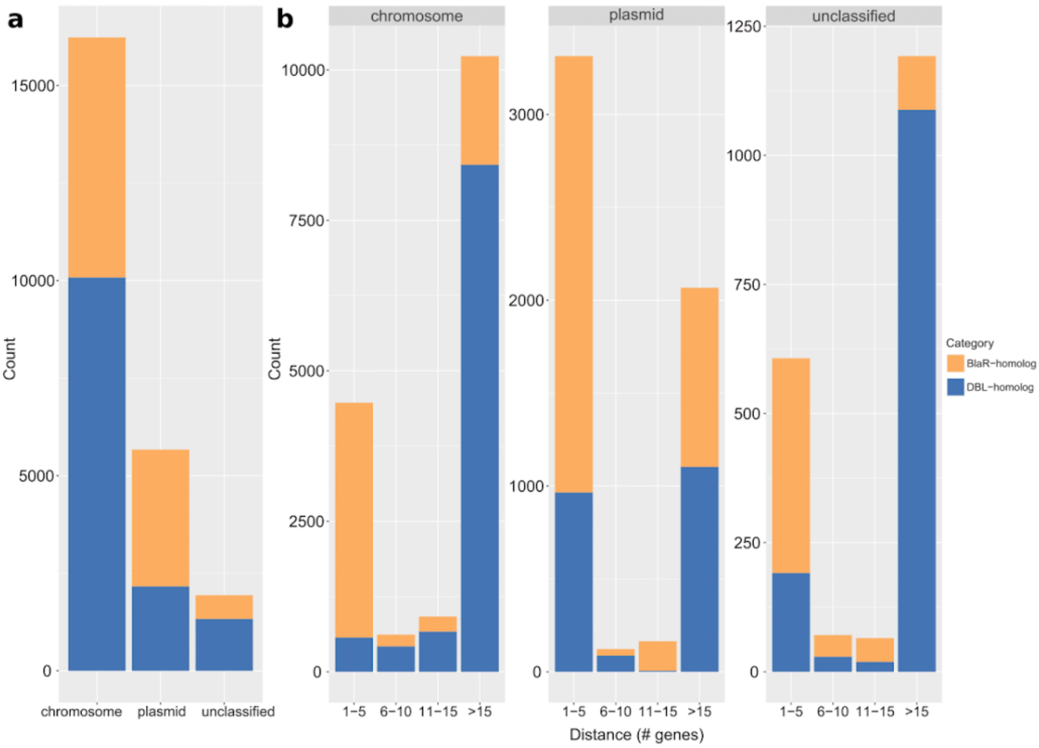


FIG 2 DBL-homolog and BlaR-homolog genes in their genetic context. (a) Distribution of DBL- and BlaR-homolog genes according to the type of encoding molecule. (b) Distribution of the distances between DBL- and BlaR-homolog genes and transposase genes across the classified contigs. The distance is measured as a range of genes centered on the gene (DBL- or BlaR-homolog) of interest. DBL- and BlaR-homolog genes are colored in blue and orange, respectively.

genes are either close to transposase genes (distance from one to five genes) or very distant (more than 15 genes) in each genetic context. Concerning BlaR-homologs, 63.4% of the genes are close to at least one transposase gene when chromosome-encoded and 67% when plasmid-encoded. Regarding DBL-homolog genes, only 5.6% and 44.7% are close to transposase genes on chromosomes and plasmids, respectively. The majority of DBL-homolog genes encoded on chromosomes near a transposase gene (568) are found in *Acinetobacter* (395), which is also the genus in which we identified most DBL-homologs (see section OXA-domain family proteins include BlaR homologs). However, when the genes are plasmid-encoded, those close to a transposase gene (965) are mainly found in *Klebsiella* (345), then *Acinetobacter* (189), *Shigella* (177), and *Escherichia* (112). Furthermore, three DBL-homolog genes in Cyanobacteria (two on a chromosome and one on plasmid) and one chromosome-encoded gene in Balnealaoeta are close to a transposase gene, which suggests that they might have been acquired by gene transfer. In contigs not classified by PlasFlow, we observed a higher prevalence of DBL-homolog genes than BlaR-homolog genes, and these DBL-homologs are very distant from transposase genes. As this pattern is similar to the pattern observed for chromosomes (Fig. 2b), it indicates that unclassified contigs likely correspond to chromosomes.

Signal peptide and transmembrane segment prediction. Most DBL-homolog sequences are characterized by a signal peptide (SP), as predicted by SignalP (Table 1). The Sec, Lipo, and Tat SPs were identified in 65%, 22%, and 3% of DBL-homolog unique sequences, respectively (see "Enlarging the OXA-domain family taxonomic distribution").

TABLE 1 Distribution of predicted signal peptides in DBL-homolog and BlaR-homolog unique sequences further broken down by the number of predicted transmembrane (TM) domains (0, cytoplasmic, 1, monotopic, > 1, polytopic)

	DBL-homologs				BlaR-homologs			
	Signal peptide (SP)			No SP	Signal peptide (SP)			No SP
# TM	Sec	Lipo	Tat	Other	Sec	Lipo	Tat	Other
0	1,660	587	83	195	0	0	0	0
1	70	4	2	49	0	0	0	1
> 1	0	0	0	11	2	1	0	845

The rest of the sequences (OTHER-SP, 9%) are either transmembrane proteins or have no SP. DBL-homolog sequences with a Sec-SP are mainly found in the genera *Pseudomonas*, *Burkholderia*, *Campylobacter*, *Klebsiella*, and *Legionella*, while DBL-homolog sequences with a Lipo-SP were mostly identified in *Acinetobacter* and *Leptospira*. DBL-homologs with a Tat-SP seem to be more specific to Alphaproteobacteria (*Bradyrhizobium*) whereas the "OTHER-SP" prediction is mainly associated with intrinsic pseudo-DBLs (CDD-1 and CDD-2 enzymes) of *Clostridioides* (27).

Beside signal peptide prediction of SignalP, the transmembrane segment (TM) prediction was used to distinguish between membrane proteins and cytoplasmic proteins. Whenever a SP is predicted in DBL-homolog sequences, the TM prediction indicates no TM or, rarely, one TM domain (monotopic) (Table 1). When no TM domain is detected, it may indicate that the corresponding DBL-homolog is excreted outside the cell or into the periplasmic space (for diderm bacteria). In contrast, when one TM domain is predicted, the protein is more likely to be anchored in the cytoplasmic membrane. In the majority of DBL-homologs with no SP predicted, no TM domain is detected, and these are possibly cytoplasmic proteins. Nevertheless, some exceptions exist, with 60 unique sequences (on 255 unique DBL-homolog proteins with no SP) presenting one or more TM domains (polytopic), a configuration which remains to be explained.

Almost all the BlaR-homolog proteins have no SP predicted and have, as expected, more than one TM domain (Table 1). However, only three polytopic proteins were predicted with Sec-SP or Lipo-SP instead of OTHER-SP. This can be explained by a wrong attribution by SignalP. Indeed, SignalP gives a probability for each possible SP and then chooses the highest value for the prediction but, for those sequences, the probabilities for OTHER-SP and Sec-SP/Lipo-SP are both close to 0.5.

Prevalence of DBL-homolog genes in clinical strains. Acquired resistance in clinical bacterial strains is a very important concern, but determining the clinical origin of a given bacterial isolate only based on the metadata of the corresponding genome assembly is still challenging to automate at a large scale. Indeed, BioSample reports from the NCBI can contain such information but these remain difficult to analyze due to the lack of a controlled vocabulary. To overcome this difficulty, we used a script that standardizes all the words of a BioSample report. Thus, 20,317 BioSample accessions were associated with the 20,342 bacterial assemblies containing DBL-homolog or BlaR-homolog genes for a total of 223 unique standardized words. Note that 4,658 BioSample reports did not contain any word. BioSamples with a positive clinical score (see Materials and Methods for details) were considered as clinical strains while those with negative scores were not. Furthermore, we decided to not classify BioSamples with a null score (essentially due to the aforementioned lack of words). Using this strategy, 3,192 bacteria were classified as *clinical*, 2,810 as *non-clinical*, and 14,340 could not be classified. Around 28% of gene sequences belong to classified strains, of which 55% clinical strains and, among those "clinical genes," 73% are DBL-homolog sequences (Table 2). Clinical DBL-homolog genes encoded on a plasmid are exclusively present in Gammaproteobacteria, mostly in *Acinetobacter* (176), *Klebsiella* (135), and *Enterobacter* (93), while DBL-homolog genes encoded on chromosomes are mostly found in Proteobacteria

TABLE 2 Distribution of DBL- and BlaR-homolog sequences in clinical, non-clinical and unclassified strains, further broken by type of encoding molecule (chromosome, plasmid, or unclassified)

Encoding molecule	Clinical		Non-clinical		Unclassified	
	DBL-homologs	BlaR-homologs	DBL-homologs	BlaR-homologs	DBL-homologs	BlaR-homologs
Chromosome	2,080	459	1,716	696	6,282	4,998
Plasmid	512	515	150	305	1497	2,688
Unclassified	234	67	222	36	871	505

(2,016) and some in Firmicutes (43), Bacteroidetes (10), Spirochaetes (8), Actinobacteria (2), and Verrucomicrobia (1).

Clustering and DBL-homolog selection. Over 600 combinations of clustering parameters were tested on the OXA-domain family phylogenetic tree (see SQL database) and the clustering with the highest entropy and the lowest number of singletons (i.e., clusters of size one) was retained (x set at 0.20 and inflation at 1.5; see Materials and Methods). This specific clustering solution has a computed entropy of 0.762 and a score of 0.52. It contains 64 clusters, including 19 singletons, with the larger cluster having 207 representative putative sequences (cluster 15) among 1,413 (Table S1). In general, there is little taxonomic diversity within each cluster. Indeed, the majority of these clusters (28) contain sequences from organisms belonging to the same phylum or class.

Annotating the unique sequences using BDLB reference sequences at an identity threshold set to 100% (see Materials and Methods) allowed us to tag 340 unique sequences, corresponding to 307 reference sequences (304 DBL/pseudo-DBL and three BlaRs) among 813 BLDB sequences. When decreasing the identity threshold to 99%, 623 unique sequences were tagged with 653 reference sequences (650 DBL/pseudo-DBL and three BlaRs), while at 90%, 1,269 unique sequences were tagged with 801 reference sequences. All those tagged sequences are distributed across 18 clusters, regardless of the identity threshold. Interestingly, up to half of the reference sequences tag cluster 60 (i.e., 168 sequences at 100%; 363 at 99%; 452 at 90%). The main genus of this cluster composed of 66 representative sequences (standing for a total of 3,472 sequences) is *Acinetobacter*, which is the host organism for 99.7% of the sequences. Irrespective of the high-redundancy of cluster 60, the latter genus is known to harbor various chromosome-encoded DBL (10).

Assessment of the beta-lactamase activity in uncharacterized clusters. To test the beta-lactamase activity of some of the 46 non-annotated clusters, 10 DBL-homolog sequences were selected for expression and production. Clusters were sorted from the largest to the smallest (considering all and not only representative sequences), then one sequence from the first 10 clusters with no DBL found in the BLDB, a sequence length between 250 and 350 AAs and no mutation in the three conserved motifs defining the class D active site. Thus, the 10 DBL-homolog (termed OXAVL01 to 10) were selected from clusters 14, 22, 23, 28, 30, 39, 41, 42, 44, and 57 (Table S2). OXAVL01 has the two lysines of its active site mutated but these mutations are shared by all the sequences in cluster 14. According to the clinical score (see Prevalence of DBL-homolog genes in clinical strains), none of those DBL-homologs belong to a clinical strain (six classified as non-clinical and four as unclassified). Seven of those sequences are chromosome-encoded while no localization could be associated to OXAVL05, OXAVL09, and OXAVL10.

The OXAVL01-10 genes were cloned in the pET24a(+) plasmid under the control of the strong T7 promoter and introduced in *Escherichia coli*. The production of OXAVL01-10 was induced by IPTG and evaluated by SDS-PAGE and beta-lactam hydrolysis. No apparent over-expression of OXAVL01, OXAVL03, OXAVL05, OXAVL07, and OXAVL09 (OXA-1091) was observed in the soluble or insoluble fractions of *E. coli* (DE3) grown at 18°C and 37°C. For OXAVL04, OXAVL08, and OXAVL10, a large production of the beta-lactamases was found only in the insoluble fractions at both culture temperatures, likely indicating the formation of inclusion bodies. Only OXAVL02 (OXA-1089) and OXAVL06 (OXA-1090) were overproduced as soluble enzymes at 18°C.

TABLE 3 Beta-lactamase activity of crude extract (CE) for cells expressing active DBL-homologs[a]

	V_0 (nmol.min^{-1}.mgP^{-1})			
CE	**Nitrocefin**	**Ampicillin**	**Oxacillin**	**Imipenem**
OXAVL02	9.5	70	18	4
OXAVL03	1	1	NH	NH
OXAVL04	0.7	NH	NH	NH
OXAVL05	2.	NH	NH	NH
OXAVL06	6	85	100	7
OXAVL09	4	7	4	NH
OXAVL10	0.5	NH	NH	NH

[a] The measurements were performed in 25 mM HEPES buffer (pH 7) at 30°C. NH, no hydrolysis.

The evaluation of the beta-lactamase activity on crude cell extracts (Table 3) showed that only OXAVL02 and OXAVL06 were able to hydrolyze all beta-lactams tested, including imipenem. OXAVL09 was active versus nitrocefin, ampicillin, and oxacillin but not imipenem. OXAVL03 was able to hydrolyze nitrocefin and ampicillin. Cell extracts of OXAVL04, OXAVL05, and OXAVL10 were active only against nitrocefin. These results may only be indicative of the true spectrum of activity because of the low fraction of soluble enzymes present in some cases. The DBL-homolog enzymes were not produced in an active form in the strains bearing the plasmid pOXAVL01, pOXAVL07, or pOXAVL08.

OXAVL02 and OXAVL06 have carbapenemase activity. Because crude extracts of OXAVL02 and OXAVL06 were the only ones able to hydrolyze all tested beta-lactams and had the highest level of expression in the soluble fraction, we focused our work on those two hydrolases. The purification of the two enzymes included three chromatographic steps, namely, an anion exchanger, an IMAC affinity chromatography, and a molecular sieve. For OXAVL02, the purification consists in an IMAC column followed by a strong anion exchanger high resolution SOURCE 15Q column. The last step is a size exclusion chromatography (SEC). At the end of the process, we obtained more than 100 mg of pure protein per liter of culture. The three steps of the OXAVL06 purification are a Q Sepharose HP ion exchanger, an IMAC column, and finally a SEC. For OXAVL06, we obtained 10 mg of pure protein per liter of culture.

SEC experiments revealed that the OXAVL02 elutes in three major peaks (Fig. 3a), with one at an elution volume typical of a monomeric DBL (~260 mL). The two additional peaks elute at about 230 mL and 180 mL, which is similar to the elution volume of the dimer and multimer, respectively. The three peaks displayed an oxacillinase activity. Due to the low precision of oligomeric states of the proteins determined by SEC, we further characterized these three peaks using size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS) (Fig. 3b).

The elution was monitored by a UV detector, a MALS detector, and a differential refractometer in line with the SEC column, allowing for the deconvolution of the protein molar masses (MM) of eluting protein complexes. The major peak in the OXAVL02 sample was confirmed to result from an equilibrium between a major monomeric form with an apparent protein MM of 32,000 ± 1,000 Da (theoretical MM [tMM] 31,298 Da) and a dimer at 62,000 ± 2,000 Da (tMM 62,596 Da). Of the two other peaks, the lower elution volume peak (at 180 mL) contained large aggregates (apparent MM $> 3 \times 10^5$ Da), while the higher elution volume peak (230 mL) corresponded to the approximate MM of a dimer at 62,000 ± 2,000 Da (tMM 62,596 Da) in equilibrium with protein aggregates. Similar data were recorded for OXAVL06 (Fig. S2).

A kinetic profile of the two purified DBL-homologs was performed in the presence of 50 mM $NaHCO_3$ (Table 4). Indeed, in the absence of hydrogenocarbonate, their activity generally showed an initial burst, followed by a pronounced slowdown, even when the substrate conversion and product accumulation were quite low. Our data indicates that OXAVL02 displays a lower catalytic efficiency compared to OXAVL06. We observed that both enzymes were not able to hydrolyze amoxicillin, temocillin, cefazolin, and cefotaxime. In addition,

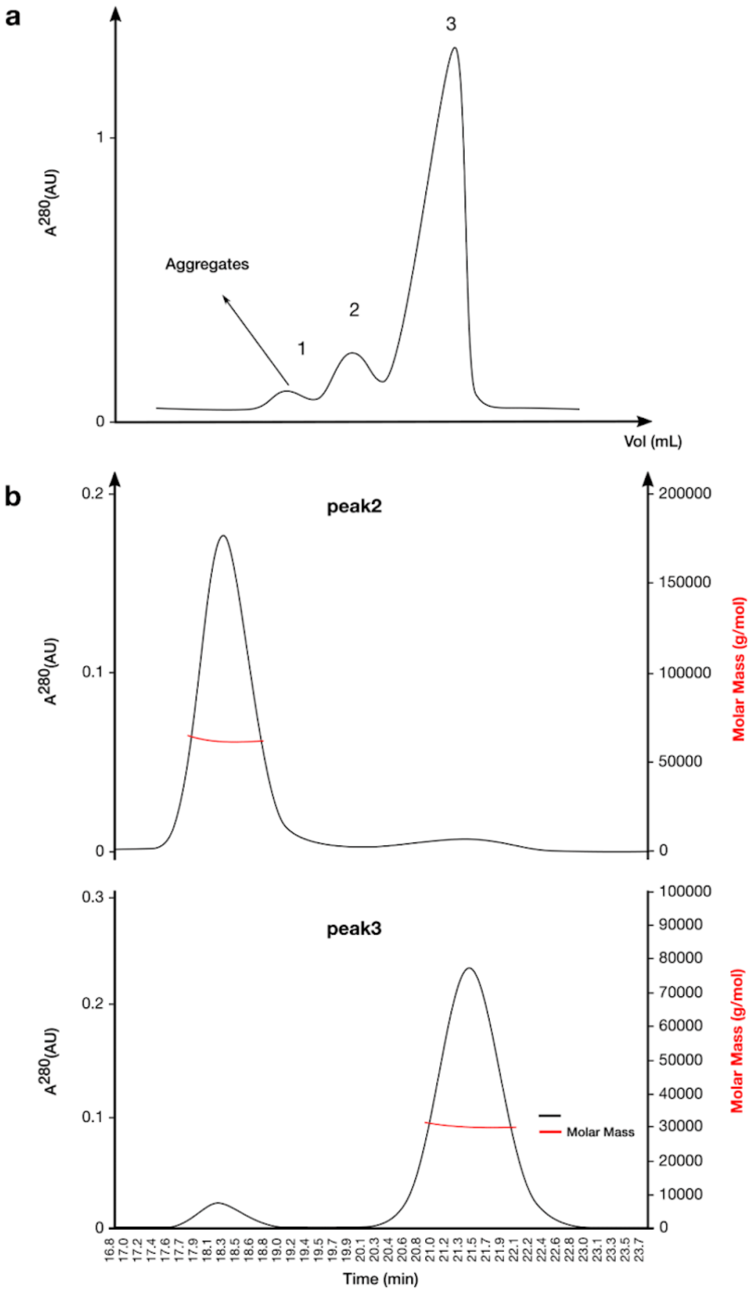


FIG 3 SEC and SEC-MALS analysis performed on the purified OXAVL02. (a) SEC analysis of the purified OXAVL02. (b) Determination of the multimeric state of OXAVL02 (peaks 2 and 3) by SEC-MALS analysis.

TABLE 4 Kinetic parameters of OXAVL02 and OXAVL06 beta-lactamases in 25 mM HEPES pH 7.5 + 50 mM NaCarbonate[a]

	OXAVL02			OXAVL06		
Antibiotics	K_{cat} (s^{-1})	Km (μM)	K_{cat}/Km ($\mu M^{-1}s^{-1}$)	K_{cat} (s^{-1})	Km (μM)	K_{cat}/Km ($\mu M^{-1}s^{-1}$)
Ampicillin	20 ± 2	380 ± 10	0.055 ± 0.007	530 ± 30	270 ± 20	2 ± 0.3
Carbenicillin	22 ± 1	400 ± 20	0.055 ± 0.005	380 ± 20	1400 ± 200	0.25 ± 0.05
Piperacillin	3 ± 0.02	850 ± 30	0.0055 ± 0.0003	NH	NH	NH
Oxacillin	1 ± 0.05	690 ± 40	0.0015 ± 0.0002	90 ± 5	160 ± 20	0.56 ± 0.10
Cephaloridine	>12.5	>400	0.030 ± 0.005	24 ± 3	90 ± 10	0.26 ± 0.06
Nitrocefin	Product Inhibition			350 ± 30	20 ± 0.5	17.5 ± 2
Imipenem	9 ± 1	550 ± 50	0.016 ± 0.002	0.9 ± 0.05	0.4 ± 0.05	2.5 ± 0.03
Meropenem	0.3 ± 0.05	6 ± 0.2	0.05 ± 0.01	NH	NH	NH

[a] NH, no hydrolysis. Each kinetic value is the mean and standard deviation of three different measurements.

OXAVL06 was not active against piperacillin and meropenem. We confirmed also that the two beta-lactamases displayed a carbapenemase activity. Imipenem was among the best substrates (k_{cat}/K_m = 0.016 and 2.5 $\mu M^{-1}s^{-1}$ for OXAVL02 and OXAVL06, respectively). In comparison to values obtained for oxacillin, the k_{cat}/K_m ratios of OXAVL02 for meropenem and imipenem were 30- and 10-fold higher, respectively.

DISCUSSION

No OXA-domain family protein detected in Archaea. The focus of this study was to explore the occurrence of class D beta-lactamases in the prokaryotic world. The 24,916 identified OXA-domain family sequences correspond to 3,510 unique sequences distributed across 20,343 bacterial strains. This highlighted a well-known redundancy in the NCBI RefSeq database toward clinical strains (26) (Fig. S3). The fact that none of these OXA-domain family proteins was detected in Archaea could be expected because Archaea are naturally resistant to beta-lactam antibiotics. Indeed, even when a pseudomurein is present, the cross-linking of the glycan chains does not involve d-Ala-d-Ala and thus does not hinge on the activity of penicillin binding proteins. However, two recent studies identified class A, B, and C beta-lactamase homologues in archaeal genomes and revealed that archaeal class B and C homologues do show a weak beta-lactamase activity (21, 28). Therefore, although we did not detect OXA-domain family proteins in Archaea, it is possible that archaeal OXA-domain family proteins will be identified in further studies, like for the other classes of beta-lactamases.

Contaminated genomes from local NCBI RefSeq database. Identification of new beta-lactamases in some unexpected organisms like Archaea or non-clinical bacterial strains might seem an exciting finding but could also be artifacts. In 2021, Lupo et al. assessed the contamination level of 111,088 bacterial genomes in the NCBI RefSeq database and found that 1% of the genomes were contaminated at a minimal threshold of 5% (26). For the 20,343 genome assemblies used in the current study, 20,200 results were available, indicating that 143 genomes had been suppressed since then. Among these 20,200 bacterial genomes, 114 showed a contamination level ≥5%. Those 114 genomes are distributed across seven phyla: Proteobacteria (78), Firmicutes (29), Verrucomicrobia (2), Cyanobacteria (2), Chloroflexi (1), Chlorobi (1), and Balneolaeota (1). Obviously, conclusions for contaminated genomes should be taken with caution. For example, the only genome containing a DBL-homolog sequence in the Balneolaeota phylum is contaminated. From our data, it is however difficult to identify if this DBL-homolog is part of the contamination or if it is genuinely part of the genome, possibly acquired from an unknown organism by horizontal gene transfer.

OXA-domain family phylogeny and classification. We inferred the phylogenetic tree using a matrix of the 188 most conserved AAs (around two thirds of typical DBL length) from the 1,413 representative OXA-domain family sequences. Those representative sequences resulted from the deduplication of the 3,510 unique sequences at a global identity threshold of 95%, which means that, considering their full length, they are similar to at least 95% of observed identity with member sequences of their deduplication clusters. Then, a phylogenetic clustering of the representative OXA-domain family proteins was computed using the patristic distances taken from the tree (i.e., the sum of the branch lengths between two leaves). This patristic distance quantifies the number of AA substitutions computed by the statistical model of sequence evolution. To select the best clustering parameters, we decided to exclude the clustering solutions with less than 15 clusters. In fact, we noticed that at least half of the OXA-domain family protein sequences regroup into one single large cluster when fewer than 15 clusters are produced. The retained parameters yielded 64 clusters, including 19 of size one (singletons). Despite a larger number of clusters, BLDB reference DBL and pseudo-DBL sequences are distributed in only 16 clusters, while BlaR from *Clostridium difficile* and *Bacillus licheniformis* are found in cluster 15, and BlaR from *Staphylococcus aureus* in cluster 18.

Another objective of this study was to suggest a meaningful classification of OXA-domain family proteins based on their phylogeny. However, the majority of bacterial strains have only one OXA-domain family protein, indicating that these genes are essentially orthologous. Moreover, 50 clusters out of 64 contain sequences from organisms belonging to a single phylum or class, which means that sequence diversity within the OXA-domain family is mostly due to speciation. While it is possible to generate a classification of class D beta-lactamases based on the clustering obtained in this study, our results indicate that a more practical classification should rather include a reference to the species of origin.

Analysis of the BlaR clusters. Because some class D beta-lactamases display more sequence identity with the C-terminal beta-lactam sensing domain of BlaR than with other class D beta-lactamases, it was impossible to avoid retrieving types of proteins in our homology searches. BlaR is also characterized by a N-terminal domain containing four transmembrane helices and a zinc protease module in loop 2 that is activated upon acylation of the C-terminal domain catalytic serine by a beta-lactam antibiotic. This triggers a cascade that eventually leads to the increased expression of either a beta-lactamase or a resistant PBP (19). As a consequence, BlaR has a total length of about 600 AAs. The size was therefore used to discriminate between BlaR-homologs (>550 AAs) and the DBL-homologs (<350 AAs). Clusters 16 and 17 exclusively contain BlaR-homologs, while clusters 8, 15, and 18 contain both BlaR- and DBL-homologs (Table S1). Most BlaR-homologs harbor a polar residue as the third residue of the second conserved motif (Table S1) and contain a N-terminal peptidase domain, two specific features of the BlaR receptor. The only exceptions are a few shorter sequences found in cluster 18; which have been removed from the database since we downloaded them, possibly indicating sequencing errors. Two sequences shorter than 550 AAs and labeled as BlaR-homologs are found in cluster 15. They have the typical conserved motifs of BlaR but their N-terminal domain is truncated and likely not functional. In this study we noticed that BlaR-homolog genes are more frequent on a plasmid nearby a transposase gene. A recent study has shown that *Staphylococcus* species have Tn552-like elements carrying the *bla* operon often located on a plasmid (29). The authors hypothesized that the Tn552 transposon can mediate the transfer of the *bla* operon from a plasmid to the chromosome. This hypothesis would also fit our results showing a high prevalence of BlaR-homolog genes on plasmids and their proximity with transposase genes.

Analysis of all the 62 DBL-homolog clusters. The size of the DBL-homolog proteins is very homogenous and the only sequences longer than 350 AAs are three fusions between a class D and a class C beta-lactamase (cluster 8), possibly homologous to LRA-13 (20), three sequences with an N-terminal extension (up to 423 AAs in total) in cluster 21, and a fusion with a crotonase domain (one sequence in cluster 40) of unknown function. The analysis of the active site motifs (Table S1) shows an almost perfect conservation of the three motifs characteristic of the catalytic site (SxxK, SxV, and KTG), as well as of the tryptophan in the omega loop, which is important for the stabilization of the carboxylated lysine of the first motif. The most variable position is the second motif valine, which is often substituted by another hydrophobic AA. Some clusters do however diverge from this consensus. Indeed, clusters 13, 29, 35, 36, 45, and 54 contain only one or two sequences and have significantly impaired motifs, which are likely not compatible with a beta-lactamase activity. Cluster 12 (eight representative sequences), which also has poorly conserved motifs, with an absence of catalytic serine in most cases, is very unlikely to display beta-lactamase activity. In contrast, cluster 14 (11 representative sequences) has the following conserved motifs: SxxH, SxH/Q, AS/TG. A sequence from this cluster (OXAVL01) was selected for *in vitro* characterization. No beta-lactamase activity was measured on a crude extract, but no overexpression was detected in our assays, preventing us from drawing any definitive conclusion. The conserved motifs are, however, not sufficient to warrant a beta-lactamase activity, as demonstrated by cluster 19 (Table S1), which regroups so far only pseudo-DBL sequences like YbxI or BAC-1 (17, 18).

Probing clusters without class D beta-lactamase representative. Beyond OXAVL01, we have selected nine DBL-homologs among the 45 clusters devoid of reference OXA-domain family proteins to probe their activity. Overall, for seven of the 10 sequences selected for evaluation, a beta-lactamase activity was detected at least on crude extracts (Table 3), including two hydrolases active on imipenem (OXAVL02 and OXAVL06). The enzymatic studies of these two DBL-homolog enzymes confirmed that they both display a beta-lactamase activity and hydrolyze efficiently imipenem but that meropenem is only inactivated by OXAVL02. We also showed that the presence of hydrogenocarbonate enhances their catalytic activity, a sign of the necessary carboxylation of the first motif lysine for optimal activity. As already shown for numerous other class D beta-lactamases, the monomeric form OXAVL02 is in equilibrium with the dimeric form, the monomer being the predominant form of the enzyme at the concentration tested. These results, obtained with randomly selected enzymes, confirm that the environmental strains provide a large reservoir of new resistance genes, which include high potential for resistance to carbapenems, a family of last resort antibacterials. The acquisition of such genes by multi-resistant nosocomial strains therefore represents an important threat for the treatment of the related infections. This phenomenon has already been observed with the chromosome-encoded class A CTX-M-3 from *Kluyvera* spp., which is at the origin of the plasmid-borne CTX-M-1-derived cefotaximases produced by clinical isolates (30). This is a reminder of the importance of an adequate use of the available antibiotics to postpone as much as possible the emergence of new resistance factors.

Predicting activity profiles from amino acid sequences. The most clinically relevant result would be to deduce the activity profile of an enzyme from its AA sequence. However, determining the activity of only one representative DBL-homolog per cluster would not be informative of the specific activity profile of the cluster. In fact, it has been shown that only one mutated AA can alter the activity profile of a DBL (8, 9). Although the sequence similarity between the 1,413 representative sequences and their respective member sequences is high (i.e., at least 95% identity), the identity between the sequences within one of the 45 non-singleton phylogenetic cluster is low (i.e., down to 50%) (Table S3). Furthermore, this similarity is certainly undervalued because it is computed from only 188 unambiguously aligned AAs. Altogether, those arguments support that, for now, the activity profile of a DBL-homolog cannot be predicted only based on its AA sequence. This problem is also true for the other classes of beta-lactamases. Solving this would require a major effort for the high throughput biochemical characterization of the enzymes and the determination of their three-dimensional structure, which is more likely correlated with the substrate specificity than the AA sequence. While biochemical characterization still represents a significant bottleneck, the recent development of the AlphaFold prediction software (31) has put structure determination within reach. Consequently, the use of artificial intelligence to predict the activity profile of enzymes is not as far-fetched as it used to be.

MATERIALS AND METHODS

SQL database. Bioinformatic data generated in this study were stored into a sqlite3 database (Fig. S1). This database was exploited using SQL queries in order to generate additional results and statistics.

Reference class D beta-lactamase sequences and identification of OXA-domain family proteins. A total of 1,617 unique beta-lactamase amino-acid sequences were downloaded from the NCBI

Pathogen Detection server (ftp://ftp.ncbi.nlm.nih.gov/pathogen/) on December 1, 2017. Among those, 470 DBL were retrieved based on metadata and accession numbers. DBL protein sequences were deduplicated using CD-HIT v4.6 (32) with a global sequence identity threshold of 0.98 and then aligned using MAFFT v7.273 (33). An HMM profile was constructed from the DBL alignment using the HMMER package v3.1b2 (34) to identify OXA-domain family proteins in a local prokaryotic protein sequence database. This local database was built on December 7, 2017 using the protein sequences of 80,490 prokaryotic genome assemblies stored in the NCBI RefSeq database. OXA-domain family proteins were graphically selected using the ompa-pa.pl interactive software package (A. Bertrand and D. Baurain; https://metacpan.org/dist/Bio-MUST-Apps-OmpaPa) and taxonomically annotated using the NCBI Taxonomy.

Annotation of OXA-domain family proteins. OXA-domain family proteins were tagged using a BLAST-based annotation script (part of Bio-MUST-Drivers) with an identity threshold from 90% to 100% and an e-value threshold of 1e-20. DBL-homolog sequences used for the annotation were downloaded from the BLDB (http://www.bldb.eu/BLDB.php?prot=D) (11) on July 22, 2019, to which were added three sequences of the membrane receptor BlaR from *Clostridium difficile* (CDT53463.1), *Staphylococcus aureus* (P18357), and *Bacillus licheniformis* (P12287), the bifunctional class C/class D beta-lactamase LRA13-1 (ACH58991.1) (20) and the two intrinsic pseudo-DBLs of *Clostridium difficile* CDD-1 (CZR76508.1) and CDD-2 (SJQ22628.1) (27).

Domain characterization of OXA-domain family proteins. The potential presence of a signal peptide was predicted using local SignalP-5.0b (35). The organism option was set to "Gram+" for sequences belonging to Firmicutes and Actinobacteria and "Gram-" for the other phyla. To improve the prediction of transmembrane helices with local TMHMM v2.0 (36), the signal peptide (if any) was first removed from the original sequences when the cleavage site prediction probability was greater than or equal to 0.6. For sequences of intermediary length (i.e., between 350 and 550 AAs) and some long sequences (i.e., greater than 550 AAs), InterProScan v5.37-76.0 with default parameters and disabled use of the precalculated match lookup (37), along with pepwindowall with default parameters from the EMBOSS web portal (38) were used to distinguish between transmembrane segments and other extensions.

Localization and genetic environment of OXA-domain family proteins. A genetic environment database was built from the bacterial genomes featuring at least one OXA-domain family sequence using GeneSpy "3 in 1" module, as described in the manual (39). Contig accessions were retrieved from the database and the corresponding FASTA files were downloaded using the command-line version of the "efetch" tool from the NCBI Entrez Programming Utilities (E-utilities). PlasFlow v1.1 was used to predict potential plasmid sequences in the contig FASTA files (40).

Clinical strain determination. BioSample reports associated with bacterial organisms were also downloaded using efetch (see above). All words of a report were collected and fed to a script that renamed and standardized them using an OBO (Open Biomedical Ontologies) dictionary. A score was attributed to each standardized word: +1 for a "clinical" word, 0 for an uninformative word and −1 for a non-clinical word. At last, a final score was computed for each BioSample according to its collection of standardized words (see figshare). A bacterial strain was considered as "clinical" when its metadata were associated with a positive score, "non-clinical" for a negative score and not classified for a null score.

Alignment and phylogenetic analysis. After deduplication using CD-HIT v4.6 (32) with a global sequence identity threshold of 0.95, OXA-domain family protein sequences were aligned using MAFFT v7.273 (33). Alignments were then carefully optimized by hand using the program "ed" and alignment columns were manually selected using the program "net," both part of the MUST software package (41). The resulting matrix of 1,413 sequences × 188 unambiguously aligned AAs was used to infer a phylogenetic tree with RAxML v8.1.17 (42) under the LG+F+G4 model. Support values were initially estimated through 100 fast bootstrap pseudo-replicates with RAxML then transformed into transfer bootstrap expectation (TBE) values using the booster algorithm (43).

Phylogenetic clustering. To produce clusters of related OXA-domain family sequences, the phylogenetic tree was first converted to a phylo4 object using the readNewick function of the phylobase R package (44). Then, a patristic distance matrix (dist.mat) was computed using the distTip function from the adephylo R package (45) and an adjacency matrix (adj.mat) was computed as follows: $ajd.mat = \frac{dist.mat}{lim.p} < 1$ with $lim.p = max(dist.mat) \times x$ and x varying from 0.10 to 0.50. Clustering was performed by passing the adjacency matrix to the mcl function of the MCL R package (46) with the addLoops option set to FALSE, allow1 set to TRUE and the inflation parameter varying from 1.0 to 3.0 by increments of 0.5. The best parameter combination was chosen by maximizing a score composed of the normalized entropy and the fraction of monophyletic clusters, following the method of Califice et al. (47). However, combinations yielding less than 15 clusters were discarded, regardless of their score, in order to avoid the grouping of most OXA-domain family sequences into a single cluster and retain the potential to provide a meaningful classification.

DBL-homolog genes selection for lab validation. Based on the phylogenetic clustering of OXA-domain family proteins, 10 representative protein sequences (hereafter referred to as OXAVL01 to OXAVL10) spread among different clusters corresponding to DBL-homologs were selected as probes for the functional diversity. The criteria of selection were: (i) the sequence must belong to a cluster with more than five DBL-homologs and no DBL found in the BLDB; (ii) the length of the sequence must lie between 250 and 350 AAs and the sequence must have no mutation in the three conserved motifs defining the class D active site (SxxK, SxV, KT/SG) (except if a mutation is shared by all the sequences of the cluster); (iii) the sequence must be present in a bacterial species where no DBL is described according to the BLDB.

Gene synthesis and expression plasmids. Signal peptides of the 10 selected sequences were removed and replaced by the PelB leader sequence (48). Then, the corresponding genes were synthesized after codon optimization for expression in *E. coli*. Expression plasmids of OXAVL01 to OXAVL10 were purchased from Twist Bioscience (San Francisco, CA, USA). Briefly, the synthesized genes were cloned into pET24a(+) (Novagen-Merck KGaA, Darmstadt, DE) and inserted between BamHI (at the 5′ end of the gene) and XhoI (at 3′ end of the gene) restriction sites. All the enzymes were produced by *E. coli* BL21(DE3) (Fisher Scientific SAS Illkirch Cedex, FR) carrying pOXAVL01-pOXAVL10 plasmids in LB medium supplemented with kanamycin 50 μg/mL (LB-kanamycin).

Antibiotics. Kanamycin was purchased from MP Biomedicals; cefotaxime, cephaloridine, and oxacillin from Sigma-Aldrich; cefazolin from Pharmacia & Upjohn SpA; imipenem from MSD; meropenem from Fresenius Kabi NV/SA; ampicillin from Fisher Scientific; amoxicillin from PanPharma; carbenicillin from Pfizer Italy; piperacillin from Lederle/AHP Pharma; temocillin from Eumedica N.V/S.A; and nitrocefin from Abcam.

Assessment of soluble enzymes expression levels. Six mL of LB-kanamycin was inoculated with single colonies of *E. coli* BL21(DE3) carrying the plasmids pOXAVL01 to pOXAVL10. The precultures were incubated overnight (O/N) at 37°C with orbital shaking at 250 rpm. Next, 2.5 mL of the different precultures were added to 100 mL of fresh LB-kanamycin. The bacteria were grown to an A_{600} of 0.7 and IPTG was added at a final concentration of 0.5 mM. The different cultures were divided in two, one incubated at 37°C and the other one at 18°C. Aliquots (1 mL) of the different cultures at 37°C were taken 0 h, 2 h, and 4 h after induction. In the case of the cultures incubated at 18°C, two aliquots (0 h and 24 h after induction) were analyzed. The different aliquots were centrifuged at 5,000 *g* for 10 min, the bacterial pellets were resuspended in 25 mM HEPES buffer (pH 7.0) and sonicated (three times for 30 seconds each time at 12 watts [W]). Cell debris was eliminated by centrifugation at 13,000 *g* for 30 min. 20 μL of the soluble fractions and pellets were loaded onto a sodium dodecyl sulfate polyacrylamide gel (SDS-PAGE) (4-20%). The run was performed at a constant voltage (120 V). The beta-lactamase activity of the different fractions was determined by measuring the initial rate of hydrolysis of 100 μM Nitrocefin, 1 mM oxacillin, 1 mM ampicillin, and 100 μM imipenem.

Production and purification of OXAVL02 and OXAVL06. One hundred mL of LB-kanamycin was inoculated with a single colony of *E. coli* BL21(DE3) pOXAVL02 or *E. coli* BL21(DE3) pOXAVL06. The preculture was incubated O/N at 37°C under agitation. Then, 40 mL of the preculture was added to 1 L of fresh LB-kanamycin. IPTG (100 μM final concentration) was added when the culture reached an A_{600} of 0.7. The cultures were incubated O/N at 18°C. Cells were harvested by centrifugation at 5,000 *g* for 10 min at 4°C. The pellets were resuspended in 15 mL 50 mM Sodium Phosphate, 0.5 M NaCl, 20 mM Imidazole pH 8.0 (buffer A) for pOXAVL02, and in 25 mM HEPES pH 7.0 (buffer B) for pOXAVL06. The bacteria were disrupted with a cell disrupter (Emulsiflex C3 Avestin GmbH, DE), which allows cell lysis at a pressure of 5,500 kPa. The lysates were isolated by centrifugation at 45,000 *g* for 30 min. The two supernatants were dialyzed O/N at 4°C against buffers A and B, respectively. The dialyzes samples were then filtered through a 0.45 μm filter.

For OXAVL02, the supernatant was loaded onto Ni Sepharose (24 mL) (GE Healthcare Europe GmbH, Freiburg) previously equilibrated with buffer A. The enzymes were eluted with a gradient using 50 mM Sodium Phosphate pH 8.0, 0.5 M NaCl, 0.5 M imidazole (buffer C). The fractions displaying a beta-lactamase activity were pooled, and then dialyzed O/N against buffer B and loaded onto a Source 15 Q column 20 mL (Pharmacia Biotech/BioSurplus Inc., San Diego, CA, USA) equilibrated with the same buffer. The enzyme was eluted with a salt gradient using buffer B with 1 M NaCl (buffer D). The fractions were pooled and loaded on a molecular sieve Superdex 75 GL 500 mL column (GE Healthcare Europe GmbH, Freiburg, DE) equilibrated in buffer B.

Because the production level of OXAVL06 was much lower, the first two purification steps were inverted compared with OXAVL02. This strategy avoided a poor efficiency of the Ni Sepharose column caused by an unspecific binding of the crude protein extract that saturates the matrix. Hence, the cleared supernatant was loaded onto a 10 mL Q Sepharose HP column (GE Healthcare Europe GmbH, Freiburg) equilibrated in buffer B. The enzyme was eluted with a salt gradient using buffer D. The fractions with a beta-lactamase activity were pooled, and dialyzed O/N in 50 mM Sodium Phosphate pH 7.5, 0.5 M NaCl, 20 mM imidazole (buffer E). The dialyzes sample was loaded onto Ni Sepharose (24 mL) (GE Healthcare Europe GmbH, Freiburg) previously equilibrated with buffer E. The enzymes were eluted with a gradient using 50 mM Sodium Phosphate pH 7.5, 0.5M NaCl, 0.5 M imidazole pH 7.5 . The active fractions were collected and concentrated by ultrafiltration on a YM-10 membrane (Amicon) to a final volume of 2 mL, then loaded onto a molecular sieve Superdex 75 GL (10/300) column (GE Healthcare Europe GmbH, Freiburg) equilibrated in buffer B.

Conformational characterization of OXAVL02 and OXAVL06. The oligomeric states of the DBL-homolog enzymes were analyzed by SEC-MALS (Treos II, WYATT Technology France) (49). The experiments were performed using a HPLC Bio-inert Shimadzu Prominence LC-20Ai (SHIMADZU Benelux B.V) coupled to a SPD-20A UV/VIS detector and a RID-20 refractive index detector. The different active fractions isolated by size exclusion chromatography were dialyzed against a "SECMALS-PBS buffer" (Na_2HPO_4 10 mM, KH_2PO_4 1.8 mM, NaCl 137 mM, KCl 2.7 mM pH 7.4). Samples (100 μL OXAVL02 or OXAVL06 at 0.5 to 1 mg/mL) were loaded onto a Superdex 200 Increase 10/300 G column (GE Healthcare Bio-Sciences AB Uppsala) pre-equilibrated with the "SECMALS-PBS buffer." The column was calibrated by using bovine serum albumin (BSA) (MM = 66,430 Da) as reference standard. The data acquisition of molecular mass, distribution of Monomer-Dimer equilibrium, and percentage of aggregates were estimated using the ASTRA software (49).

Kinetic constants determination. Steady-state kinetic constants (K_m and k_{cat}) were determined by measuring substrate hydrolysis under initial rate conditions and using the Hanes–Woolf linearization of the Michaelis–Menten equation (50). Kinetic experiments were performed by following the hydrolysis of each substrate at 30°C in 50 mM HEPES buffer pH 7.5, 50 mM Na_2CO_3. The reactions were performed in a total volume of 500 μL at 30°C. BSA (20 μg/mL) was added to diluted solutions of beta-lactamase in order to prevent enzyme denaturation. The data were collected with a Specord 50 PLUS spectrophotometer (Analytik Jena). Each kinetic value is the mean of three different measurements.

Data availability. Publicly available data sets analyzed in this study and the companion SQL database can be found here: https://doi.org/10.6084/m9.figshare.18544955.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.
SUPPLEMENTAL FILE 1, PDF file, 2.4 MB.

ACKNOWLEDGMENTS

V.L. is supported by a Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture (FRIA) fellowship of the FRS-FNRS. F.K. is a research associate of the FRS-FNRS. M.G. and P.S.M. were supported by the Belgian Federal Public Service Health, Food Chain Safety and Environment (Grant No. RF 17/6317 RU-BLA-ESBL-CPE). Computational resources were provided through two grants to DB (University of Liège "Crédit de démarrage 2012" SFRD-12/04; FRS-FNRS "Crédit de recherche 2014" CDR J.0080.15). We thank Mohammed Terrak and Adrien Boes for useful advice at the initial stage of the project and Hiba Jabri for her help with the OBO dictionary.

V.L. performed experiments, analyzed the data, drew the figures, wrote the manuscript and approved the final manuscript. D.B. and F.K. conceived the study and designed experiments, analyzed the data, wrote and reviewed the manuscript and approved the final manuscript. J-M.F. conceived the study, wrote and reviewed the manuscript ,and approved the final manuscript. P.S.M. performed experiments, wrote the manuscript, and approved the final manuscript. M.G. designed experiments, wrote the manuscript, and approved the final manuscript. B.J. analyzed the data, reviewed the manuscript, and approved the final manuscript.

REFERENCES

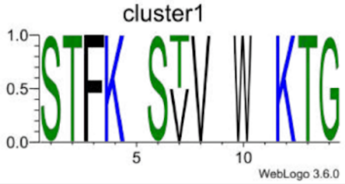
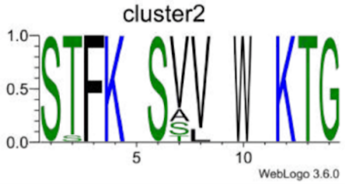
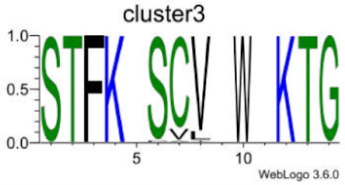
1. Abraham EP, Chain E. 1940. An enzyme from bacteria able to destroy penicillin. Nature 146:837–837. https://doi.org/10.1038/146837a0.
2. Livermore DM. 2000. Antibiotic resistance in staphylococci. Int J Antimicrob Agents 16:3–10. https://doi.org/10.1016/S0924-8579(00)00299-5.
3. Davies J, Davies D. 2010. Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev 74:417–433. https://doi.org/10.1128/MMBR.00016-10.
4. Waley SG. 1992. ß-Lactamase: mechanism of action, p 198–228. *In* Page MI (ed), The chemistry of β-lactams. Springer, Dordrecht, the Netherlands.
5. Bebrone C, Garau G, Garcia-Saez I, Chantalat L, Carfi A, Dideberg O. 2012. X-ray Structures and Mechanisms of Metallo-ß-Lactamase, p 41–77. *In* Frère J-M (ed), Beta-lactamases. Nova Science Publishers, New York, NY, USA.
6. Bush K. 2018. Past and Present Perspectives on β-Lactamases. Antimicrob Agents Chemother 62. https://doi.org/10.1128/AAC.01076-18.
7. Philippon A, Jacquier H, Ruppé E, Labia R. 2019. Structure-based classification of class A beta-lactamases, an update. Curr Res Transl Med 67:115–122. https://doi.org/10.1016/j.retram.2019.05.003.
8. Sougakoff W. 2012. Structure of class A beta-lactamases, p 21–39. *In* Frère J-M (ed), Beta-lactamases. Nova Science Publishers, New York, NY, USA.
9. Castanheira M, Simner PJ, Bradford PA. 2021. Extended-spectrum β-lactamases: an update on their characteristics, epidemiology and detection. JAC Antimicrob Resist 3:dlab092. https://doi.org/10.1093/jacamr/dlab092.
10. Evans BA, Amyes SGB. 2014. OXA β-lactamases. Clin Microbiol Rev 27:241–263. https://doi.org/10.1128/CMR.00117-13.
11. Naas T, Oueslati S, Bonnin RA, Dabos ML, Zavala A, Dortet L, Retailleau P, Iorga BI. 2017. Beta-lactamase database (BLDB) - structure and function. J Enzyme Inhib Med Chem 32:917–919. https://doi.org/10.1080/14756366.2017.1344235.
12. Danel F, Frère JM, Livermore DM. 2001. Evidence of dimerisation among class D beta-lactamases: kinetics of OXA-14 beta-lactamase. Biochim Biophys Acta 1546:132–142. https://doi.org/10.1016/S0167-4838(01)00133-9.
13. Golemi D, Maveyraud L, Vakulenko S, Samama JP, Mobashery S. 2001. Critical involvement of a carbamylated lysine in catalytic function of class D beta-lactamases. Proc Natl Acad Sci U S A 98:14280–14285. https://doi.org/10.1073/pnas.241442898.
14. Ledent P, Frère JM. 1993. Substrate-induced inactivation of the OXA2 beta-lactamase. Biochem J 295:871–878. https://doi.org/10.1042/bj2950871.
15. Poirel L, Naas T, Nordmann P. 2010. Diversity, epidemiology, and genetics of class D beta-lactamases. Antimicrob Agents Chemother 54:24–38. https://doi.org/10.1128/AAC.01512-08.
16. Zhu YF, Curran IH, Joris B, Ghuysen JM, Lampen JO. 1990. Identification of BlaR, the signal transducer for beta-lactamase production in Bacillus licheniformis, as a penicillin-binding protein with strong homology to the OXA-2 beta-lactamase (class D) of Salmonella typhimurium. J Bacteriol 172:1137–1141. https://doi.org/10.1128/jb.172.2.1137-1141.1990.
17. Colombo M-L, Hanique S, Baurin SL, Bauvois C, De Vriendt K, Van Beeumen JJ, Frère J-M, Joris B. 2004. The ybxI gene of Bacillus subtilis 168 encodes a class D beta-lactamase of low activity. Antimicrob Agents Chemother 48:484–490. https://doi.org/10.1128/AAC.48.2.484-490.2004.
18. Toth M, Antunes NT, Stewart NK, Frase H, Bhattacharya M, Smith CA, Vakulenko SB. 2016. Class D β-lactamases do exist in Gram-positive bacteria. Nat Chem Biol 12:9–14. https://doi.org/10.1038/nchembio.1950.
19. Hardt K, Joris B, Lepage S, Brasseur R, Lampen JO, Frère JM, Fink AL, Ghuysen JM. 1997. The penicillin sensory transducer, BlaR, involved in the inducibility of beta-lactamase synthesis in Bacillus licheniformis is embedded in the plasma membrane via a four-alpha-helix bundle. Mol Microbiol 23:935–944. https://doi.org/10.1046/j.1365-2958.1997.2761642.x.
20. Allen HK, Moe LA, Rodbumrer J, Gaarder A, Handelsman J. 2009. Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. ISME J 3:243–251. https://doi.org/10.1038/ismej.2008.86.
21. Keshri V, Panda A, Levasseur A, Rolain J-M, Pontarotti P, Raoult D. 2018. Phylogenomic analysis of β-lactamase in archaea and bacteria enables the identification of putative new members. Genome Biol Evol 10:1106–1114. https://doi.org/10.1093/gbe/evy028.
22. Sengupta S, Chattopadhyay MK, Grossart H-P. 2013. The multifaceted roles of antibiotics and antibiotic resistance in nature. Front Microbiol 4:47. https://doi.org/10.3389/fmicb.2013.00047.
23. Papp-Wallace KM, Endimiani A, Taracila MA, Bonomo RA. 2011. Carbapenems: past, present, and future. Antimicrob Agents Chemother 55:4943–4960. https://doi.org/10.1128/AAC.00296-11.
24. Dewi DAPR, Thomas T, Ahmad Mokhtar AM, Mat Nanyan NS, Zulfigar SB, Salikin NH. 2021. Carbapenem resistance among marine bacteria-an

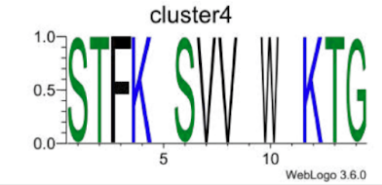
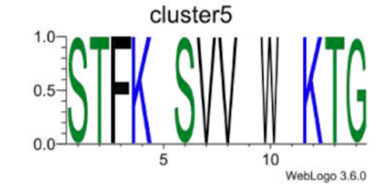
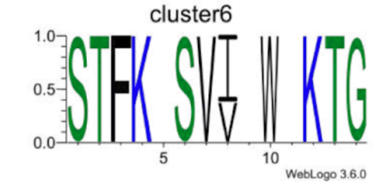
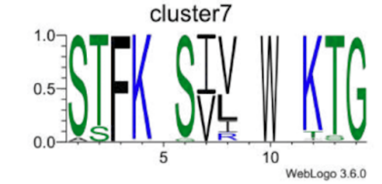
emerging threat to the global health sector. Microorganisms 9:2147. https://doi.org/10.3390/microorganisms9102147.

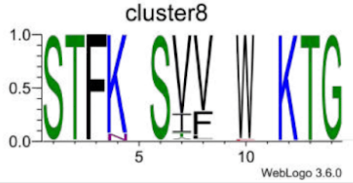
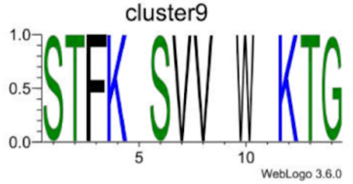
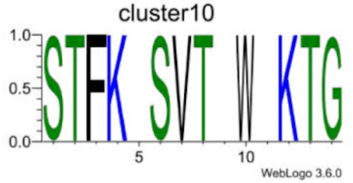
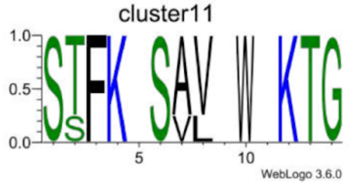
25. Cherak Z, Loucif L, Moussi A, Rolain J-M. 2021. Carbapenemase-producing Gram-negative bacteria in aquatic environments: a review. J Glob Antimicrob Resist 25:287–309. https://doi.org/10.1016/j.jgar.2021.03.024.
26. Lupo V, Van Vlierberghe M, Vanderschuren H, Kerff F, Baurain D, Cornet L. 2021. Contamination in reference sequence databases: time for divide-and-rule tactics. Front Microbiol 12:755101. https://doi.org/10.3389/fmicb.2021.755101.
27. Toth M, Stewart NK, Smith C, Vakulenko SB. 2018. Intrinsic class D β-lactamases of clostridium difficile. mBio 9. https://doi.org/10.1128/mBio.01803-18.
28. Diene SM, Pinault L, Armstrong N, Azza S, Keshri V, Khelaifia S, Chabrière E, Caetano-Anolles G, Rolain J-M, Pontarotti P, Raoult D. 2020. Dual RNase and β-lactamase activity of a single enzyme encoded in Archaea. Life Basel Switz 10:280. https://doi.org/10.3390/life10110280.
29. Sun Z, Zhou D, Zhang X, Li Q, Lin H, Lu W, Liu H, Lu J, Lin X, Li K, Xu T, Bao Q, Zhang H. 2020. Determining the genetic characteristics of resistance and virulence of the "epidermidis cluster group" through pan-genome analysis. Front Cell Infect Microbiol 10:274. https://doi.org/10.3389/fcimb.2020.00274.
30. Rodríguez MM, Power P, Radice M, Vay C, Famiglietti A, Galleni M, Ayala JA, Gutkind G. 2004. Chromosome-encoded CTX-M-3 from Kluyvera ascorbata: a possible origin of plasmid-borne CTX-M-1-derived cefotaximases. Antimicrob Agents Chemother 48:4895–4897. https://doi.org/10.1128/AAC.48.12.4895-4897.2004.
31. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2.
32. 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. https://doi.org/10.1093/bioinformatics/bts565.
33. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
34. Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
35. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37:420–423. https://doi.org/10.1038/s41587-019-0036-z.
36. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580. https://doi.org/10.1006/jmbi.2000.4315.
37. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031.
38. Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.
39. Garcia PS, Jauffrit F, Grangeasse C, Brochier-Armanet C. 2019. GeneSpy, a user-friendly and flexible genomic context visualizer. Bioinformatics 35:329–331. https://doi.org/10.1093/bioinformatics/bty459.
40. Krawczyk PS, Lipinski L, Dziembowski A. 2018. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res 46:e35. https://doi.org/10.1093/nar/gkx1321.
41. Philippe H. 1993. MUST, a computer package of Management Utilities for Sequences and Trees. Nucleic Acids Res 21:5264–5272. https://doi.org/10.1093/nar/21.22.5264.
42. Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
43. Lemoine F, Domelevo Entfellner J-B, Wilkinson E, Correia D, Dávila Felipe M, De Oliveira T, Gascuel O. 2018. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature 556:452–456. https://doi.org/10.1038/s41586-018-0043-0.
44. Hackathon R, Bolker B, Butler M, Cowan P V D de, Eddelbuettel D, Holder M, Jombart T, Kembel S, Michonneau F, Orme D, O'Meara B, Paradis E, Regetz J, Zwickl D. 2019. phylobase: base package for phylogenetic structures and comparative data. https://CRAN.R-project.org/package=phylobase.
45. Jombart T, Balloux F, Dray S. 2010. adephylo: new tools for investigating the phylogenetic signal in biological traits. Bioinformatics 26:1907–1909. https://doi.org/10.1093/bioinformatics/btq292.
46. Jäger ML. 2015. MCL: Markov cluster algorithm. https://CRAN.R-project.org/package=MCL.
47. Califice S, Baurain D, Hanikenne M, Motte P. 2012. A single ancient origin for prototypical serine/arginine-rich splicing factors. Plant Physiol 158:546–560. https://doi.org/10.1104/pp.111.189019.
48. Steiner D, Forrer P, Stumpp MT, Plückthun A. 2006. Signal sequences directing cotranslational translocation expand the range of proteins amenable to phage display. Nat Biotechnol 24:823–831. https://doi.org/10.1038/nbt1218.
49. Wyatt PJ. 2021. Differential light scattering and the measurement of molecules and nanoparticles: A review. Anal Chim Acta X 7–8:100070. https://doi.org/10.1016/j.acax.2021.100070.
50. Cornish-Bowden A. 1995. Fundamentals of enzyme kinetics pp 343. Portland Press, London, UK. https://books.google.be/books?id=_jZzQgAACAAJ.

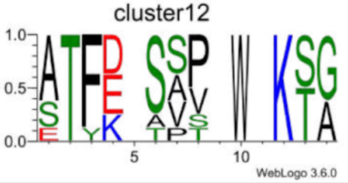
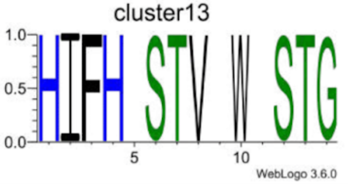
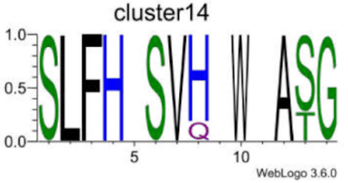
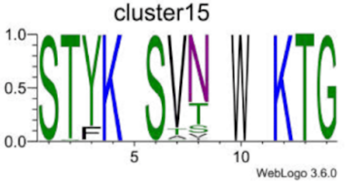
2.2.2. Supplementary Material

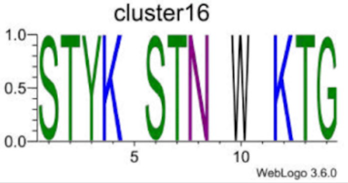
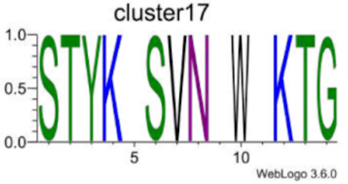
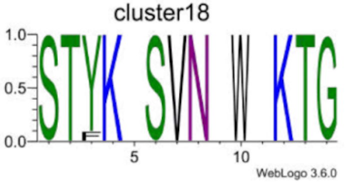
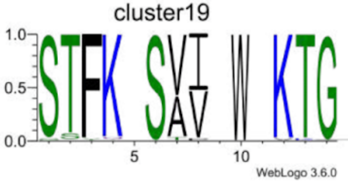
Table S1. **Details of the 64 phylogenetic clusters.** The retained clustering (x set at 0.20 and inflation at 1.5; see Materials and Methods) had a computed entropy of 0.762 and a score of 0.52. BLAST-based annotation (identity threshold of 90%) of the clusters was derived from sub-family annotations (if any, unless the protein name is used) of the class-D beta-lactamases from the Beta-lactamase Database (BLDB).

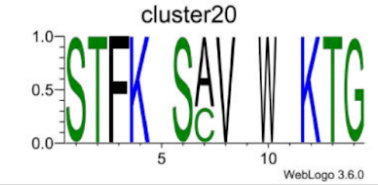
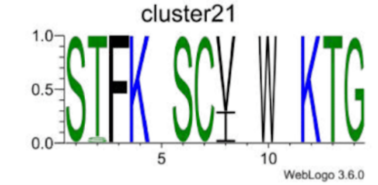
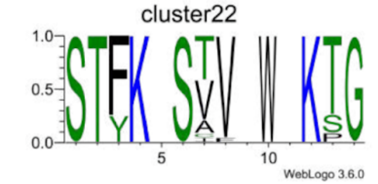
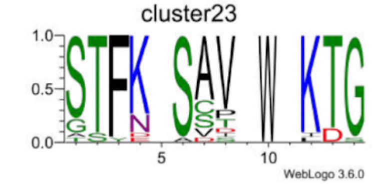
Cluster	# seqs (representative)	# seqs (unique)	# seqs (all)	# DBL-homologs (unique)	# BlaR-homologs (unique)	Taxonomy	BLDB sub-family	Active site logo
cluster1	4	5	5	5	0	Oligoflexia		
cluster2	16	31	860	31	0	Betaproteobacteria, Deltaproteobacteria, Gammaproteobacteria	OXA1	
cluster3	46	79	767	79	0	Chlamydiae, Alphaproteobacteria, Deltaproteobacteria, Gammaproteobacteria	OXA29	

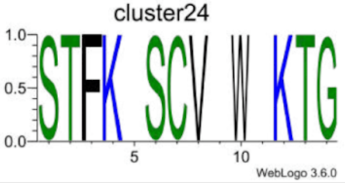
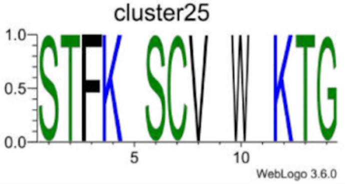
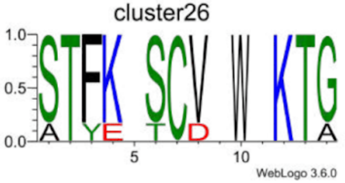
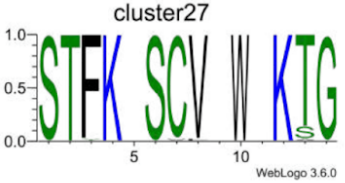
cluster4	1	1	2	1	0	Alphaproteobacteria		
cluster5	2	3	3	3	0	Alphaproteobacteria		
cluster6	16	98	164	98	0	Gammaproteobacteria	OXA12	
cluster7	98	170	719	170	0	Planctomycetes, Verrucomicrobia, Alphaproteobacteria, Gammaproteobacteria	OXA9	

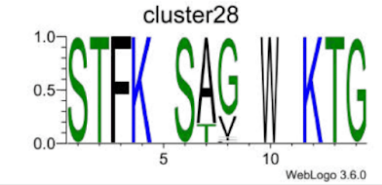
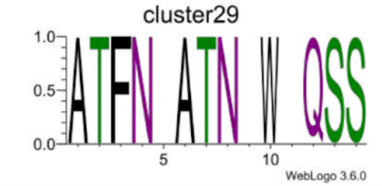
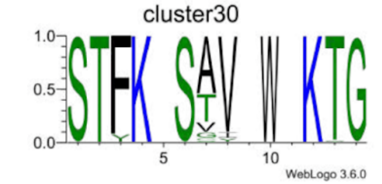
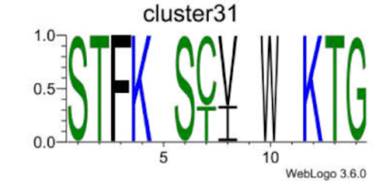
cluster8	195	561	1768	546	15	Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria	OXA114a, OXA22, OXA243, OXA258, OXA42, OXA457	
cluster9	1	1	1	1	0	Deltaproteobacteria		
cluster10	1	1	1	1	0	Firmicutes		
cluster11	4	6	6	6	0	Deltaproteobacteria		

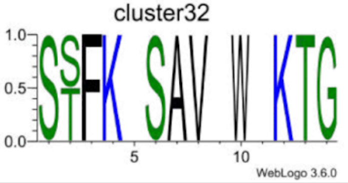
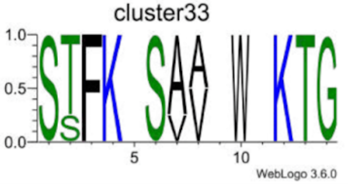
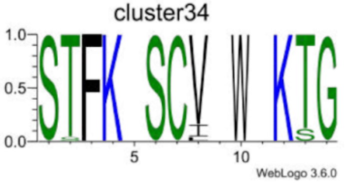
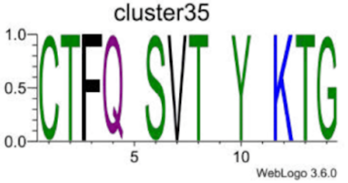
cluster12	8	12	16	12	0	Gammaproteobacteria		
cluster13	1	1	1	1	0	Spirochaetes		
cluster14	11	72	415	72	0	Spirochaetes		
cluster15	207	396	1306	28	368	Firmicutes, Fusobacteria, Deltaproteobacteria	BLAR1	

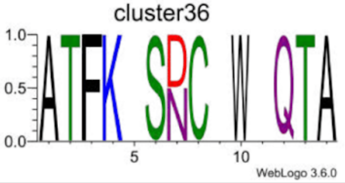
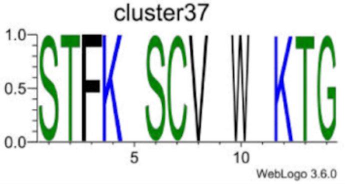
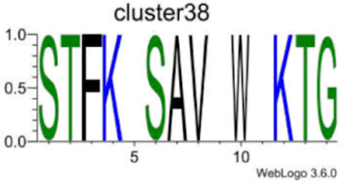
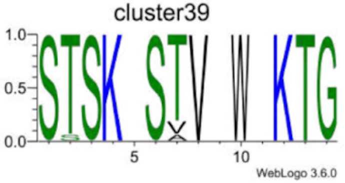
cluster16	4	5	13	0	5	Firmicutes		
cluster17	1	1	1	0	1	Firmicutes		
cluster18	11	478	9214	18	460	Firmicutes	BLAR1	
cluster19	83	209	374	209	0	Firmicutes	BAC1, BAD1, BAT1, BED1, BEN1, BOC1, BPU1, BSD1, BSU1, YBXI	

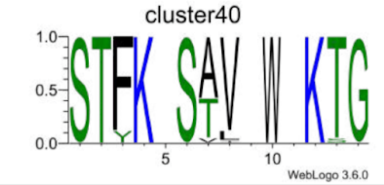
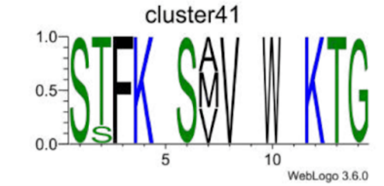
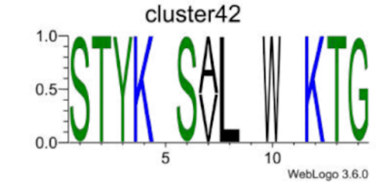
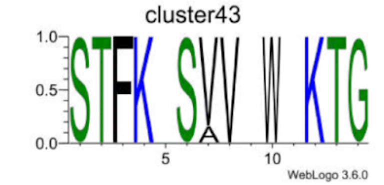
cluster20	9	9	9	9	0	Chlorobi, Verrucomicrobia, Deltaproteobacteria		
cluster21	38	87	956	87	0	Firmicutes	CDD	
cluster22	24	25	27	25	0	Bacteroidetes, Chloroflexi, Betaproteobacteria, Deltaproteobacteria, Gammaproteobacteria		
cluster23	23	24	25	24	0	Bacteroidetes		

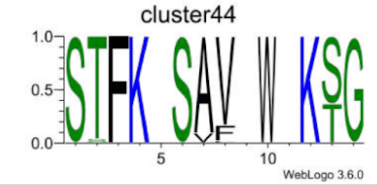
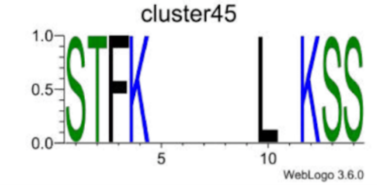
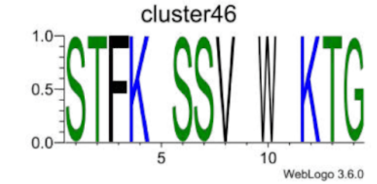
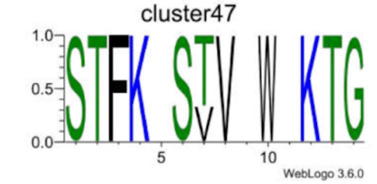
cluster24	1	1	1	1	0	Bacteroidetes		
cluster25	1	1	1	1	0	Bacteroidetes		
cluster26	6	6	6	6	0	Bacteroidetes		
cluster27	44	48	82	48	0	Bacteroidetes	OXA347	

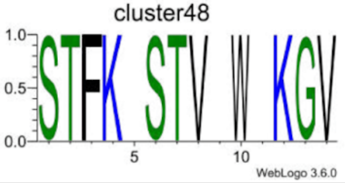
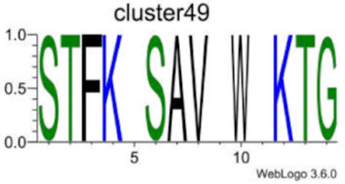
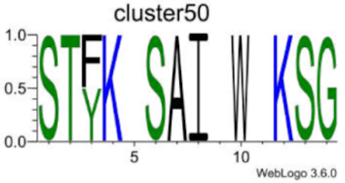
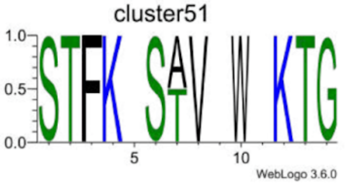
cluster28	55	62	70	62	0	Bacteroidetes, Fusobacteria, Balneolaeota		
cluster29	1	1	1	1	0	Bacteroidetes		
cluster30	53	60	81	60	0	Cyanobacteria, Bacteroidetes, Spirochaetes, Verrucomicrobia, Gammaproteobacteria		
cluster31	3	3	4	3	0	Bacteroidetes, Deltaproteobacteria		

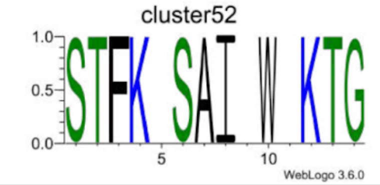
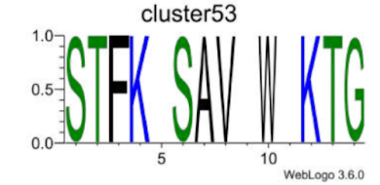
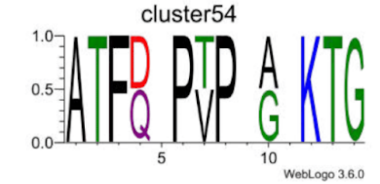
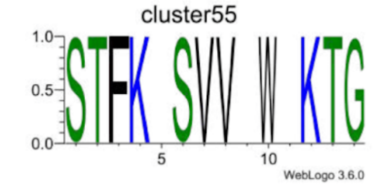
cluster32	4	4	4	4	0	Gammaproteobacteria		
cluster33	4	4	4	4	0	Alphaproteobacteria		
cluster34	26	36	39	36	0	Firmicutes, Alphaproteobacteria, Deltaproteobacteria, Epsilonproteobacteria, Gammaproteobacteria	OXA464	
cluster35	1	1	2	1	0	Gammaproteobacteria		

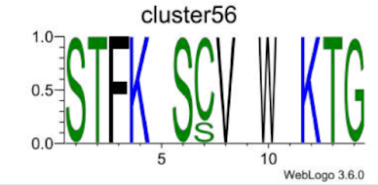
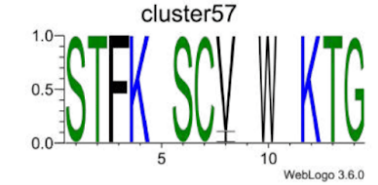
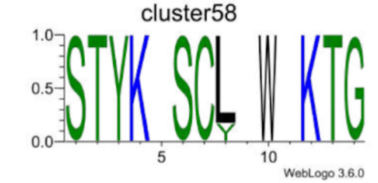
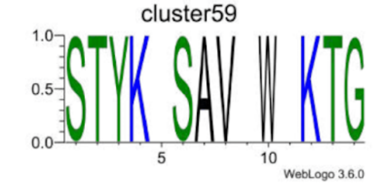
cluster36	2	2	2	2	0	Gammaproteobacteria		
cluster37	29	48	55	48	0	Chlorobi, Betaproteobacteria, Gammaproteobacteria	LCRNPS	
cluster38	2	2	2	2	0	Betaproteobacteria		
cluster39	16	16	18	16	0	Alphaproteobacteria		

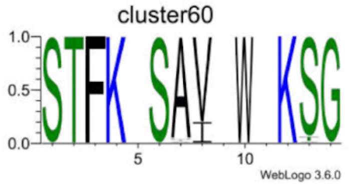
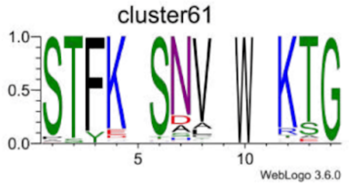
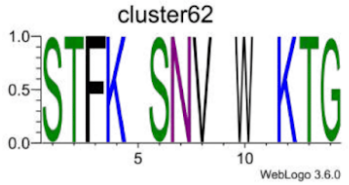
cluster40	43	81	217	81	0	Cyanobacteria, Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria	OXA2, OXA20, OXA46	
cluster41	6	7	7	7	0	Gammaproteobacteria		
cluster42	11	15	23	15	0	Gammaproteobacteria		
cluster43	19	57	263	57	0	Betaproteobacteria, Gammaproteobacteria	OXA10, OXA48, OXA5, OXA548, OXA55	

cluster44	55	106	141	106	0	Alphaproteobacteria		
cluster45	1	1	1	1	0	Alphaproteobacteria		
cluster46	1	1	1	1	0	Betaproteobacteria		
cluster47	3	3	3	3	0	Betaproteobacteria, Gammaproteobacteria		

cluster48	1	1	1	1	0	Gammaproteobacteria		
cluster49	1	1	1	1	0	Bacteroidetes		
cluster50	2	2	2	2	0	Deltaproteobacteria		
cluster51	2	2	2	2	0	Alphaproteobacteria		

cluster52	1	1	1	1	0	Deltaproteobacteria		
cluster53	1	1	1	1	0	Deltaproteobacteria		
cluster54	2	2	2	2	0	Alphaproteobacteria		
cluster55	1	1	1	1	0	Alphaproteobacteria		

cluster56	5	5	6	5	0	Deltaproteobacteria		
cluster57	9	15	21	15	0	Betaproteobacteria		
cluster58	12	17	22	17	0	Betaproteobacteria	OXA153, OXA154, OXA156, OXA157, OXA62	
cluster59	1	1	1	1	0	Alphaproteobacteria		

cluster60	66	313	3472	313	0	Firmicutes, Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria	OXA134, OXA143, OXA211, OXA213, OXA214, OXA228, OXA23, OXA24, OXA266, OXA274, OXA279, OXA286, OXA294, OXA296, OXA299, OXA308, OXA51, OXA58, OXA665	
cluster61	93	195	2361	195	0	Actinobacteria, Verrucomicrobia, Cyanobacteria, Fusobacteria, Alphaproteobacteria, Betaproteobacteria, Deltaproteobacteria, Epsilonproteobacteria, Gammaproteobacteria	OXA50, OXA60	
cluster62	1	1	1	1	0	Epsilonproteobacteria		

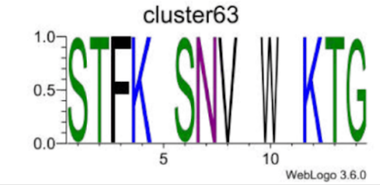
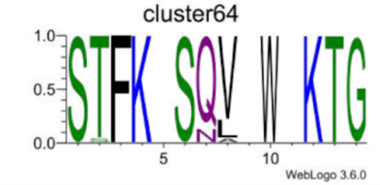
cluster63	1	1	1	1	0	Gammaproteobacteria		
cluster64	23	110	1339	110	0	Spirochaetes, Fusobacteria, Epsilonproteobacteria	OXA184, OXA493, OXA576, OXA61, OXA63, OXA85	

Table S2. **Details of the ten sequences selected for production.**

Cluster	Protein accession	Len. (AA)	Signal peptide	# TM	Encoding molecule	Strain status	Organism
cluster14	WP_003003775.1 (OXAVL01)	278	Lipo	0	chromosome	non-clinical	Leptospira weilii serovar Ranarum str. ICFT (GCF_000332415.1)
cluster22	WP_075071939.1 (OXAVL02)	289	Lipo	0	chromosome	non-clinical	Longilinea arvoryzae (GCF_001050235.1)
cluster23	WP_039139369.1 (OXAVL03)	272	Lipo	0	chromosome	non-clinical	Flavihumibacter solisilvae (GCF_000814475.1)
cluster28	WP_012859784.1 (OXAVL04)	264	Sec	0	chromosome	unclassified	Sebaldella termitidis ATCC 333866 (GCF_000024405.1)
cluster30	WP_017307273.1 (OXAVL05)	283	Sec	0	unclassified	non-clinical	Spirulina subsalsa PCC 9445 (GCF_000314005.1)
cluster39	WP_051601260.1 (OXAVL06)	274	Lipo	0	chromosome	non-clinical	Hyphomonas beringensis (GCF_000682755.1)
cluster41	WP_011041344.1 (OXAVL07)	303	Sec	0	chromosome	unclassified	Colwellia psychrerythraea 34H (GCF_000012325.1)
cluster42	WP_023398243.1 (OXAVL08)	262	Sec	0	chromosome	unclassified	Pseudoalteromonas luteoviolacea NCIMB 1944 (GCF_001625565.1)
cluster44	WP_012045014.1 (OXAVL09)	272	Tat	0	unclassified	unclassified	Bradyrhizobium sp. BTAi1 (GCF_000015165.1)
cluster57	WP_042878825.1 (OXAVL10)	278	Sec	0	unclassified	non-clinical	Cupriavidus necator A5-1 (GCF_000744095.1)

Table S3. **Quantile values of the patristic distances within each cluster.** The patristic distance is the sum of the branch lengths connecting two leaves in the phylogenetic tree. Here all intra-cluster pairwise combinations are considered.

Cluster	# seqs	Min	P25	Median	P75	Max
cluster1	4	0.00969	0.2059525	0.676575	0.7019525	0.71252
cluster2	16	0	0.415955	0.5568	0.7226125	1.0108
cluster3	46	0.01466	0.361375	0.81088	1.026605	1.67419
cluster5	2	0.27086	0.27086	0.27086	0.27086	0.27086
cluster6	16	0.01953	0.15301	0.21419	0.31377	0.54345
cluster7	98	0.01905	0.69301	0.86071	1.03713	1.92185
cluster8	195	0	0.770885	0.93889	1.110865	1.97655
cluster11	4	0.08396	0.30977	0.56715	0.6112075	0.69586
cluster12	8	0.04397	0.6758725	0.839215	0.95245	1.48239
cluster14	11	0.044	0.19852	0.24713	0.294545	0.41045
cluster15	207	0	0.92162	1.06751	1.22639	2.13709
cluster16	4	0.03267	0.2215675	0.40052	0.59001	0.60036
cluster18	11	0.00888	0.086785	0.90216	0.959735	1.07516
cluster19	83	0.00917	0.545395	0.72186	0.847135	1.48553
cluster20	9	0.04945	0.398155	0.715475	0.8981125	1.11107
cluster21	38	0.02332	0.77539	0.95729	1.162505	1.50658
cluster22	24	0.04456	0.67553	1.034565	1.400815	1.71683
cluster23	23	0.04849	0.5105	0.62438	0.78451	1.37281
cluster26	6	0.15873	0.49242	0.72263	0.83551	1.06572
cluster27	44	0.02671	0.5849225	0.738465	0.9223375	1.36627
cluster28	55	0.03899	0.62546	0.82497	1.00318	1.7674
cluster30	53	0.02332	0.4809275	0.62634	1.00227	1.61624
cluster31	3	0.30642	0.616195	0.92597	0.98358	1.04119
cluster32	4	0.25138	0.26632	0.30462	0.405275	0.441
cluster33	4	0.22863	0.6938975	0.77468	1.0105775	1.06262
cluster34	26	0.07511	0.80561	0.96124	1.10099	1.44135
cluster36	2	0.05568	0.05568	0.05568	0.05568	0.05568
cluster37	29	0.03376	0.6354725	0.80797	0.9243375	1.29984
cluster38	2	0.31643	0.31643	0.31643	0.31643	0.31643
cluster39	16	0.05838	0.419635	0.53467	0.958935	1.13721
cluster40	43	0.02916	0.463685	0.58767	0.694815	1.00944
cluster41	6	0.08239	0.678715	0.81042	0.91188	1.02917
cluster42	11	0.06239	0.267915	0.39539	0.66776	0.75012
cluster43	19	0.02369	0.57327	0.70553	0.865345	1.32908
cluster44	55	0.0192	0.20912	0.31989	0.40854	0.96414

cluster47	3	0.22092	0.556305	0.89169	0.93107	0.97045
cluster50	2	0.29347	0.29347	0.29347	0.29347	0.29347
cluster51	2	0.58105	0.58105	0.58105	0.58105	0.58105
cluster54	2	0.58559	0.58559	0.58559	0.58559	0.58559
cluster56	5	0.07076	0.40092	0.722035	0.7739	0.82261
cluster57	9	0.02085	0.0984025	0.298545	0.443115	0.53549
cluster58	12	0.01854	0.105465	0.14054	0.17294	0.23443
cluster60	66	0.01423	0.52437	0.61548	0.78495	1.57714
cluster61	93	0.0097	0.85578	1.085395	1.3369775	2.30433
cluster64	23	0	0.33179	0.56669	0.93299	1.25124

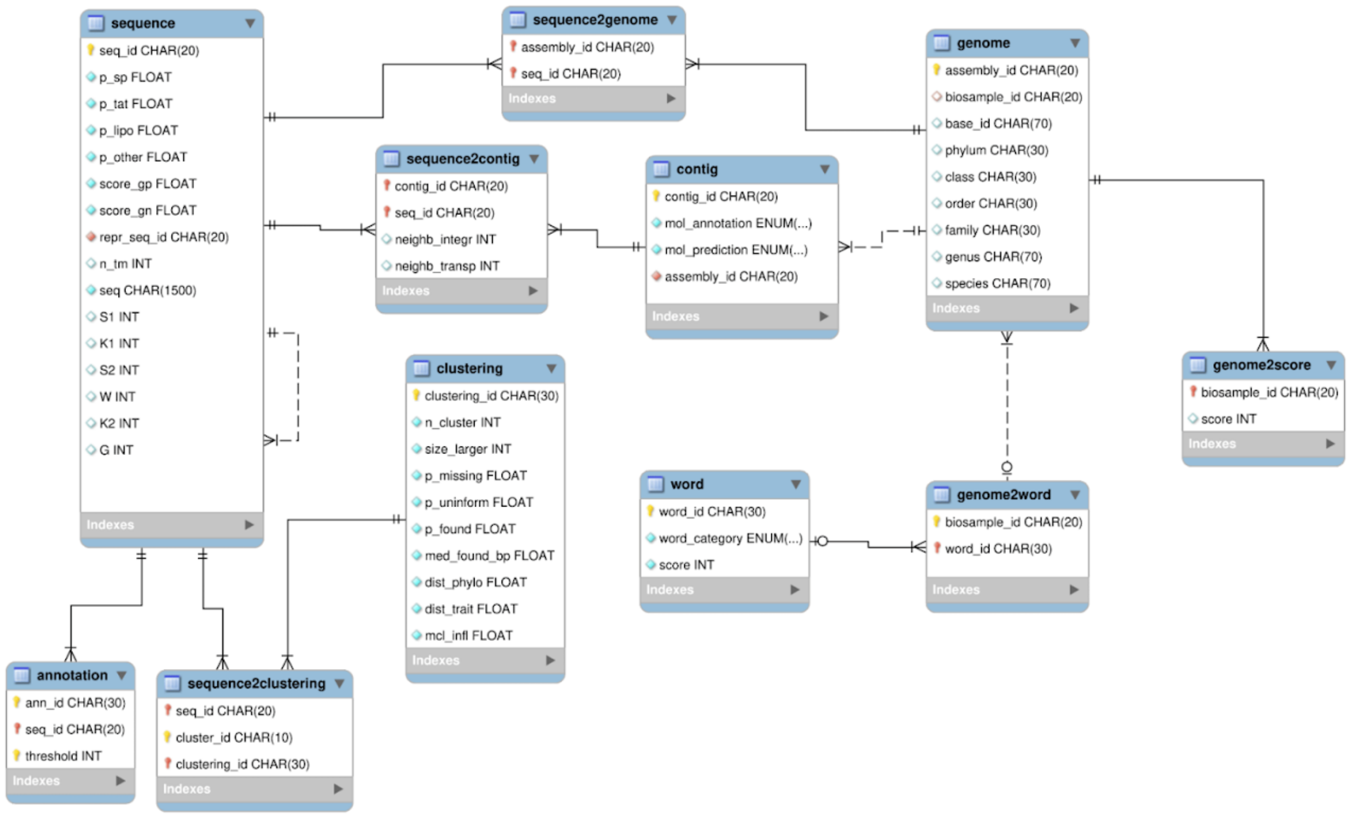
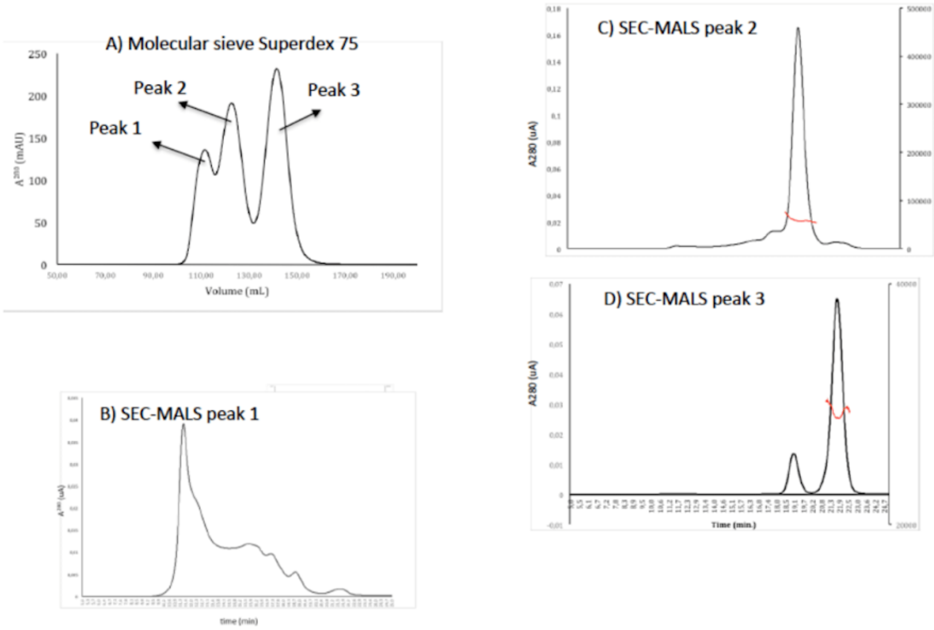


Figure S1. **Schematic representation of the SQL database.**

Figure S2. **SEC and SEC-MALS analysis performed on the purified OXAVL06.** (**A**) Size Exclusion Chromatography (SEC) analysis of the purified OXAVL06. (**B-D**) Determination of the multimeric state of OXAVL06 (peaks 1, 2 and 3) by SEC-MALS (Multi-Angle Light Scattering) analysis.



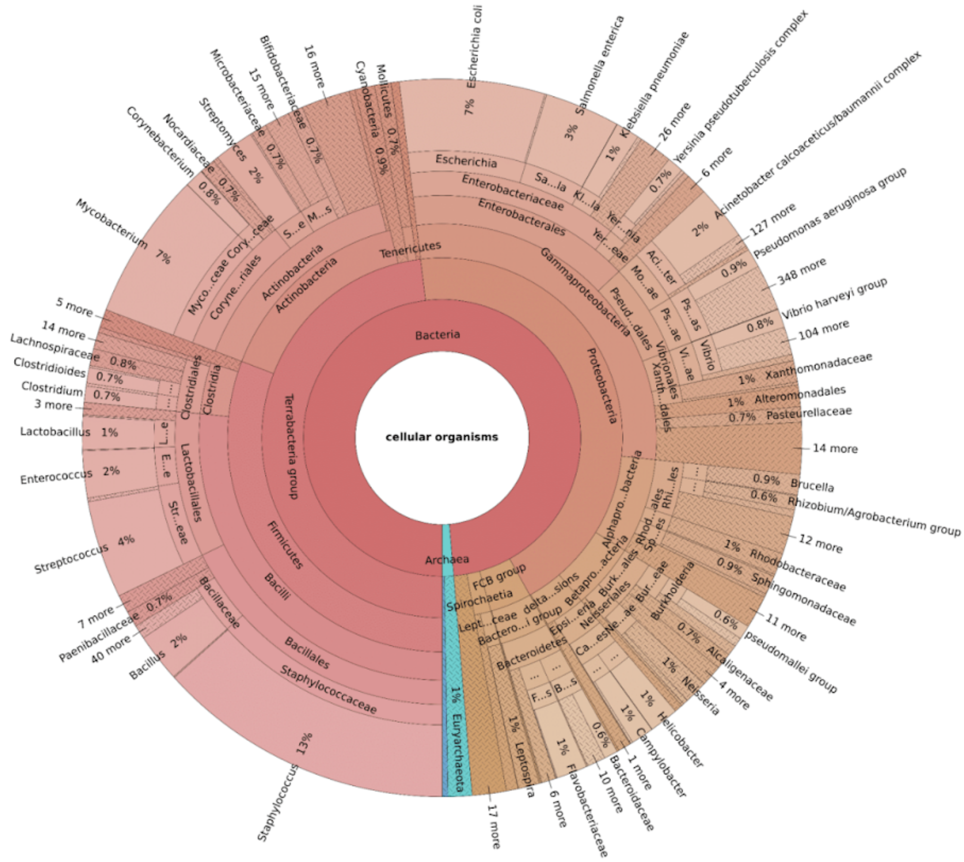


Figure S3. **Krona chart of the taxonomic diversity of the local NCBI RefSeq database.**

Supplemental Data 1

InterProScan raw results

```
GCF_001758545.1|WP_071849985.1619	ProSitePatternsPS00337 Beta-lactamase class-D active site.	49	59	-
GCF_001758545.1|WP_071849985.1619	Pfam	PF00144 Beta-lactamase 265	613	5.0E-65
GCF_001758545.1|WP_071849985.1619	Pfam	PF00905 Penicillin binding protein transpeptidase domain	39	230	8.2E-30
GCF_001758545.1|WP_071849985.1619	SUPERFAMILY	SSF56601		262	614	1.11E-84
GCF_001758545.1|WP_071849985.1619	ProSitePatternsPS00336 Beta-lactamase class-C active site.	313	320	-
GCF_001758545.1|WP_071849985.1619	SUPERFAMILY	SSF56601		26	235	6.91E-27
GCF_001758545.1|WP_071849985.1619	Gene3D	G3DSA:3.40.710.10		255	617	3.4E-116
GCF_001758545.1|WP_071849985.1619	Gene3D	G3DSA:3.40.710.10		24	254	3.3E-74
GCF_001423125.1|WP_056159201.1633	Pfam	PF00905 Penicillin binding protein transpeptidase domain	407	619	4.5E-33
GCF_001423125.1|WP_056159201.1633	Gene3D	G3DSA:3.40.710.10		391	632	2.4E-71
GCF_001423125.1|WP_056159201.1633	Pfam	PF05569 BlaR1 peptidase M56	27	291	6.5E-27
GCF_001423125.1|WP_056159201.1633	ProSitePatternsPS00337 Beta-lactamase class-D active site.	418	428	-
GCF_001423125.1|WP_056159201.1633	CDD	cd07341 M56_BlaR1_MecR1_like	134	324	6.12734E-30
GCF_001423125.1|WP_056159201.1633	SUPERFAMILY	SSF56601		384	624	3.42E-32
GCF_900129765.1|WP_073220251.1573	Pfam	PF00144 Beta-lactamase 220	562	9.1E-68
GCF_900129765.1|WP_073220251.1573	SUPERFAMILY	SSF56601		6	203	1.04E-25
GCF_900129765.1|WP_073220251.1573	ProSitePatternsPS00337 Beta-lactamase class-D active site.	12	22	-
GCF_900129765.1|WP_073220251.1573	ProSitePatternsPS00336 Beta-lactamase class-C active site.	268	275	-
GCF_900129765.1|WP_073220251.1573	Gene3D	G3DSA:3.40.710.10		1	217	4.1E-73
GCF_900129765.1|WP_073220251.1573	Pfam	PF00905 Penicillin binding protein transpeptidase domain	4	191	4.1E-31
GCF_900129765.1|WP_073220251.1573	Gene3D	G3DSA:3.40.710.10		218	571	2.8E-119
GCF_900129765.1|WP_073220251.1573	SUPERFAMILY	SSF56601		219	568	1.22E-86
GCF_000344615.1|WP_017879576.1613	Pfam	PF05569 BlaR1 peptidase M56	13	306	1.4E-28
GCF_000344615.1|WP_017879576.1613	SUPERFAMILY	SSF56601		372	596	4.13E-30
```

```
GCF_000344615.1|WP_017879576.1613    Pfam    PF00905 Penicillin binding protein transpeptidase domain    387    599    2.0E-41
GCF_000344615.1|WP_017879576.1613    Gene3D  G3DSA:3.40.710.10            369    613    2.7E-82
GCF_000344615.1|WP_017879576.1613    CDD     cd07341 M56_BlaR1_MecR1_like  131    306    1.94503E-32
GCF_000315425.1|WP_005669363.1571    Pfam    PF05569 BlaR1 peptidase M56   156    289    1.2E-26
GCF_000315425.1|WP_005669363.1571    SUPERFAMILY    SSF56601               328    557    6.91E-29
GCF_000315425.1|WP_005669363.1571    CDD     cd07341 M56_BlaR1_MecR1_like  119    245    1.62457E-28
GCF_000315425.1|WP_005669363.1571    Pfam    PF00905 Penicillin binding protein transpeptidase domain    354    557    1.5E-29
GCF_000315425.1|WP_005669363.1571    Gene3D  G3DSA:3.40.710.10            334    571    1.9E-72
GCF_001191005.1|WP_050407176.1598    Pfam    PF00905 Penicillin binding protein transpeptidase domain    373    584    6.9E-36
GCF_001191005.1|WP_050407176.1598    Gene3D  G3DSA:3.40.710.10            354    598    1.3E-77
GCF_001191005.1|WP_050407176.1598    SUPERFAMILY    SSF56601               354    575    6.87E-30
GCF_001191005.1|WP_050407176.1598    Pfam    PF05569 BlaR1 peptidase M56   11     307    2.8E-34
GCF_900129765.1|WP_073218605.1598    SUPERFAMILY    SSF56601               357    586    5.27E-32
GCF_900129765.1|WP_073218605.1598    Gene3D  G3DSA:3.40.710.10            354    598    1.8E-80
GCF_900129765.1|WP_073218605.1598    Pfam    PF05569 BlaR1 peptidase M56   117    285    1.8E-26
GCF_900129765.1|WP_073218605.1598    CDD     cd07341 M56_BlaR1_MecR1_like  169    306    4.50559E-20
GCF_900129765.1|WP_073218605.1598    Pfam    PF00905 Penicillin binding protein transpeptidase domain    366    585    2.4E-40
GCF_001425045.1|WP_057157847.1599    Pfam    PF05569 BlaR1 peptidase M56   19     306    8.6E-35
GCF_001425045.1|WP_057157847.1599    CDD     cd07341 M56_BlaR1_MecR1_like  134    307    4.81022E-28
GCF_001425045.1|WP_057157847.1599    MobiDBLite     mobidb-lite    consensus disorder prediction 83    106    -
GCF_001425045.1|WP_057157847.1599    SUPERFAMILY    SSF56601               356    585    1.82E-32
GCF_001425045.1|WP_057157847.1599    Pfam    PF00905 Penicillin binding protein transpeptidase domain    372    584    7.3E-38
GCF_001425045.1|WP_057157847.1599    Gene3D  G3DSA:3.40.710.10            356    599    5.0E-81
GCF_001758545.1|WP_071849986.1604    Pfam    PF00905 Penicillin binding protein transpeptidase domain    376    591    6.7E-35
GCF_001758545.1|WP_071849986.1604    Gene3D  G3DSA:3.40.710.10            360    604    3.0E-77
GCF_001758545.1|WP_071849986.1604    CDD     cd07341 M56_BlaR1_MecR1_like  123    299    1.76121E-32
GCF_001758545.1|WP_071849986.1604    Pfam    PF05569 BlaR1 peptidase M56   12     288    2.9E-32
GCF_001758545.1|WP_071849986.1604    SUPERFAMILY    SSF56601               363    587    8.54E-31
GCF_001425265.1|WP_056397199.1609    SUPERFAMILY    SSF56601               368    599    2.73E-34
GCF_001425265.1|WP_056397199.1609    CDD     cd07341 M56_BlaR1_MecR1_like  178    317    4.03119E-28
GCF_001425265.1|WP_056397199.1609    Pfam    PF05569 BlaR1 peptidase M56   20     316    5.3E-29
```

```
GCF_001425265.1|WP_056397199.1609    Gene3D G3DSA:3.40.710.10          366    609    6.5E-81
GCF_001425265.1|WP_056397199.1609    Pfam   PF00905 Penicillin binding protein transpeptidase domain    378    600    2.4E-40
GCF_001758795.1|WP_071658926.1613    CDD    cd07341 M56_BlaR1_MecR1_like  123    292    3.41763E-31
GCF_001758795.1|WP_071658926.1613    Pfam   PF00905 Penicillin binding protein transpeptidase domain    385    600    3.0E-36
GCF_001758795.1|WP_071658926.1613    Pfam   PF05569 BlaR1 peptidase M56   12     294    2.0E-30
GCF_001758795.1|WP_071658926.1613    Gene3D G3DSA:3.40.710.10          369    613    1.0E-78
GCF_001758795.1|WP_071658926.1613    SUPERFAMILY    SSF56601           372    598    1.11E-29
GCF_001425265.1|WP_056397197.1621    ProSitePatternsPS00336 Beta-lactamase class-C active site.  304    311    -
GCF_001425265.1|WP_056397197.1621    Pfam   PF00905 Penicillin binding protein transpeptidase domain    25     227    4.6E-28
GCF_001425265.1|WP_056397197.1621    ProSitePatternsPS00337 Beta-lactamase class-D active site.  48     58     -
GCF_001425265.1|WP_056397197.1621    Gene3D G3DSA:3.40.710.10          22     251    2.9E-74
GCF_001425265.1|WP_056397197.1621    SUPERFAMILY    SSF56601           256    594    1.24E-84
GCF_001425265.1|WP_056397197.1621    SUPERFAMILY    SSF56601           17     230    5.5E-27
GCF_001425265.1|WP_056397197.1621    Pfam   PF00144 Beta-lactamase 256     593    5.9E-67
GCF_001425265.1|WP_056397197.1621    Gene3D G3DSA:3.40.710.10          252    607    7.2E-118
GCF_001758785.1|WP_070245907.1632    ProSitePatternsPS00337 Beta-lactamase class-D active site.  417    427    -
GCF_001758785.1|WP_070245907.1632    CDD    cd07341 M56_BlaR1_MecR1_like  124    300    1.9486E-31
GCF_001758785.1|WP_070245907.1632    Pfam   PF00905 Penicillin binding protein transpeptidase domain    407    623    3.5E-35
GCF_001758785.1|WP_070245907.1632    SUPERFAMILY    SSF56601           393    616    1.09E-30
GCF_001758785.1|WP_070245907.1632    Pfam   PF05569 BlaR1 peptidase M56   20     284    1.2E-27
GCF_001758785.1|WP_070245907.1632    Gene3D G3DSA:3.40.710.10          390    631    1.4E-67
GCF_000335815.1|WP_008451277.1601    Gene3D G3DSA:3.40.710.10          357    601    3.6E-78
GCF_000335815.1|WP_008451277.1601    CDD    cd07341 M56_BlaR1_MecR1_like  123    308    4.69834E-32
GCF_000335815.1|WP_008451277.1601    Pfam   PF00905 Penicillin binding protein transpeptidase domain    373    588    1.5E-34
GCF_000335815.1|WP_008451277.1601    Pfam   PF05569 BlaR1 peptidase M56   12     288    2.3E-31
GCF_000335815.1|WP_008451277.1601    SUPERFAMILY    SSF56601           359    586    6.91E-30
GCF_000381565.1|WP_026334185.1605    Pfam   PF00905 Penicillin binding protein transpeptidase domain    375    588    3.2E-37
GCF_000381565.1|WP_026334185.1605    Pfam   PF05569 BlaR1 peptidase M56   12     293    5.9E-29
GCF_000381565.1|WP_026334185.1605    SUPERFAMILY    SSF56601           367    585    4.1E-32
GCF_000381565.1|WP_026334185.1605    CDD    cd07341 M56_BlaR1_MecR1_like  123    262    9.10106E-28
GCF_000381565.1|WP_026334185.1605    Gene3D G3DSA:3.40.710.10          362    605    2.6E-77
```

```
GCF_900143065.1|WP_072783046.1597    Pfam    PF05569 BlaR1 peptidase M56    18      271     2.0E-30
GCF_900143065.1|WP_072783046.1597    ProSitePatternsPS00337 Beta-lactamase class-D active site.  382     392     -
GCF_900143065.1|WP_072783046.1597    Gene3D  G3DSA:3.40.710.10              355     597     2.2E-71
GCF_900143065.1|WP_072783046.1597    CDD     cd07341 M56_BlaR1_MecR1_like   134     305     6.64842E-27
GCF_900143065.1|WP_072783046.1597    Pfam    PF00905 Penicillin binding protein transpeptidase domain    345     590     5.6E-34
GCF_900143065.1|WP_072783046.1597    SUPERFAMILY    SSF56601                358     589     1.09E-28
GCF_000383895.1|WP_019923896.1631    CDD     cd07341 M56_BlaR1_MecR1_like   143     322     7.635E-28
GCF_000383895.1|WP_019923896.1631    Pfam    PF05569 BlaR1 peptidase M56    21      287     1.1E-27
GCF_000383895.1|WP_019923896.1631    Pfam    PF00905 Penicillin binding protein transpeptidase domain    405     622     2.2E-33
GCF_000383895.1|WP_019923896.1631    Gene3D  G3DSA:3.40.710.10              389     630     2.6E-68
GCF_000383895.1|WP_019923896.1631    SUPERFAMILY    SSF56601                392     615     3.7E-29
GCF_000383895.1|WP_019923896.1631    ProSitePatternsPS00337 Beta-lactamase class-D active site.  416     426     -
```

EMBOSS pepwindowall

ClustalO alignment

GCF_001758545.1|WP_071849985.1, GCF_001425265.1|WP_056397197.1 and GCF_900129765.1|WP_073220251.1 are three fusion proteins between a class C and a class D beta-lactamase.

```
CLUSTAL O(1.2.4) multiple sequence alignment


GCF_900143065.1|WP_072783046.1      --------MTGFDIALVRLLLAAAGSLAAGGAVWGVALLCRRYLP---ALAQHRSLWLLG  49
GCF_001423125.1|WP_056159201.1      MIFGGIDTGIDLDLTLARLLLAAGGSLAVGCAVWAAALLCRRWLP---VLTQQRSLWLGA  57
GCF_000383895.1|WP_019923896.1      ------MMPIDLDIVLARLLLAAAGSLAAGGAVWAVAVLCRRTLP---ALAQQRSLWLSG  51
GCF_001758785.1|WP_070245907.1      ------MMPIDLDVVLARLLLAAAGSLAAGGAVWAVAVLCRRTLP---ALAQQRSLWLCG  51
GCF_001191005.1|WP_050407176.1      --------MSSFDLVLLRFLLASLGCLLAGLSVWGLSALLRRYLP---ALAAQRSIWLLG  49
GCF_000315425.1|WP_005669363.1      --------MNPLDQLVIRFLFAGAGCLAAGLAVWGATWLARRALP---ALGMQRSTWLLG  49
GCF_000381565.1|WP_026334185.1      --------MAFSTLFICRFLLASLGCMAAGLAVWALTAACRRHLP---QVAMQRSTWLLG  49
GCF_001758795.1|WP_071658926.1      --------MAISTLLLFRFLLASLGCMAAGLVVWALTAACRRCLP---QVAVQRSTWLLG  49
GCF_001758545.1|WP_071849986.1      --------MAISTLLIFRFLLASLGCMAAGLVVWALTAACRRYLP---QVALQRSTWLLG  49
GCF_000335815.1|WP_008451277.1      --------MAISTLLIFRFLLASLGCLAAGLVVWALTAACRRYLP---QVAVQRSTWLLG  49
GCF_000344615.1|WP_017879576.1      --------MSAFDLWLLRFLLASCGCVVAGLGVWALTALCRRCVA---EFSLQRSMWLLS  49
GCF_900129765.1|WP_073218605.1      --MTGL--SALLDQLPLRFLLASAACVAVGMGAWGVTAMFGRL-P---GVALRRSTWLLG  52
GCF_001425045.1|WP_057157847.1      --MTG---AVLADAWLLRFLLASAGCLAAGLAVWALTALCRRL-P---GVALHRSVWLLS  51
GCF_001425265.1|WP_056397199.1      --MSG---TAPLDAWLLRFLLASAGCLAAGLGVWALTALCRRL-P---GVALQRSTWLLG  51
GCF_001758545.1|WP_071849985.1      ------------------------------------------------------------  0
GCF_001425265.1|WP_056397197.1      ------------------------------------------------------------  0
GCF_900129765.1|WP_073220251.1      ------------------------------------------------------------  0
sp|P12287|BLAR_BACLI                --MS----SS----FFIPFLVSQIL----LSLFFSIIILIKKLLRTQITVGTHYYISVIS  46


GCF_900143065.1|WP_072783046.1      QIAVAVVFAAMLLPTTQRLRVVPVIEMNEEAPSPMPAASASPQASA----TVTPAADLTP  105
GCF_001423125.1|WP_056159201.1      QVAVAVVFLAMLLPPSDSLRVLPLIEIVEPEAAPASAVPAAAAATHA-AEATRPAPDLVA  116
GCF_000383895.1|WP_019923896.1      QVAVAAVFLAMLWPPAESLRVMPIIEIVEPAAAPMPATPASAEAHAAMAAAPAPGLGLTS  111
GCF_001758785.1|WP_070245907.1      QVAVAAVFLAMLLPPDEGLRVMPIIEIVEQEAAPAAPMPAVPPA----QAAAAPVLALAS  107
GCF_001191005.1|WP_050407176.1      QLTVIGTFVLILLPHSERVRLLLPIEGATETVSHYLVPAA-PAA------APS-QPVAPL  101
GCF_000315425.1|WP_005669363.1      QVAIAATFLLLLAPPAQQAPMHPVFEVDVGAAAA----SL-PAA------LAPPGLAAGV  98
GCF_000381565.1|WP_026334185.1      QLTIIATFLALLLPHSERLRLLPAIELPETMLAAPAAPDH-ATA------PAQDAAMAGG  102
GCF_001758795.1|WP_071658926.1      QMTIIATFLVILLPHSERLRLLPAIELPETMLAAAAAPDH-AAT------PAPAAAVADS  102
GCF_001758545.1|WP_071849986.1      QMTIIATFLVILLPHSERLRLLPAIELPETMLAAPAAPEH-AAT------PPHAAAASDD  102
GCF_000335815.1|WP_008451277.1      QMTIIATFLVILLPHSERLRLLPPIELPETVPAAPATPEH-AAT------PAPAAAVADS  102
GCF_000344615.1|WP_017879576.1      QITVVATFLVILLPHSERLRVIPPIDLADETAARLVAAGG-TTA------RAAASTAVSA  102
GCF_900129765.1|WP_073218605.1      QVTVMAAFLVVLLPHSEHLRLVPPIDLPEVALSHATATDG-GSA------LPGPGRPSAA  105
GCF_001425045.1|WP_057157847.1      QATVVAAFLVILLPHSERLRLVPPIELPEAAASRPAAQDG-PAA------RRAGEGSDAA  104
GCF_001425265.1|WP_056397199.1      QATVVAAFLVILLPHSERLRLVPPIDLPDAAQVRPAASGG-QAA------QPAAAGQDAL  104
GCF_001758545.1|WP_071849985.1      ------------------------------------------------------------  0
GCF_001425265.1|WP_056397197.1      ------------------------------------------------------------  0
GCF_900129765.1|WP_073220251.1      ------------------------------------------------------------  0
sp|P12287|BLAR_BACLI                LLALIAPFIPFHFLKSHHFDWILNLGGAQSALSQTHSTD-------------------KT  87
```

```
GCF_900143065.1|WP_072783046.1      AAPALSWHTWLRD------------AARAWLLLYLLGLGHAVWRWRRAQR----WLEALA  149
GCF_001423125.1|WP_056159201.1      VAADRPASAWMRD------------AGRAWLLLYLSGLAHALWRWWRAQR----LLDTLA  160
GCF_000383895.1|WP_019923896.1      SAADRPLSAWVRD------------AGRAWLVLYLLGLLHTLWRWQRAQR----LLEALA  155
GCF_001758785.1|WP_070245907.1      AATERPLSAWARD------------AGRAWLVLYLLGLVHALWRWRRAQR----LLDGLA  151
GCF_001191005.1|WP_050407176.1      VQAQREEHPWLNL------------AAYAWLAAYLLGLAYTVGRLLHGQR----MLNRLA  145
GCF_000315425.1|WP_005669363.1      PASIAGERSWLAW------------LAWAWASVYACGLAWTLGRLWRGQR----IVQRLL  142
GCF_000381565.1|WP_026334185.1      HSAAADHRAWLSH------------GAQAWLSLYLLGLAYTTGRVLQSQR----TLNGLA  146
GCF_001758795.1|WP_071658926.1      GDAPTDYRAWLTG------------GAQAWLSLYLLGLAYTTGRVLQAHR----TLNGLA  146
GCF_001758545.1|WP_071849986.1      GKAATDYRAWLTG------------GAQAWLSLYLLGLAYTTGRVLQAHR----TLNGLA  146
GCF_000335815.1|WP_008451277.1      GNAATDYRAWLTG------------GAQAWLSLYLLGLAYTTGRVLQAQR----TLNALA  146
GCF_000344615.1|WP_017879576.1      PVAGSIERPWLAY------------GAQAWLLVYLLGLGYAGARLLHARR----ILNGLC  146
GCF_900129765.1|WP_073218605.1      TGDALRPAIWLTR------------AAQVWLLVYLLGLGYTVLRLLYARR----LLDHLA  149
GCF_001425045.1|WP_057157847.1      STGQVQPASWLTR------------AAQAWLLAYLLGLGYTVFQLLRARR----MLNGLA  148
GCF_001425265.1|WP_056397199.1      AEAGTRPAAWLTH------------AAQAWLLAYLLGLGYAVFQLLRARR----ILNGLA  148
GCF_001758545.1|WP_071849985.1      ------------------------------------------------------------  0
GCF_001425265.1|WP_056397197.1      ------------------------------------------------------------  0
GCF_900129765.1|WP_073220251.1      ------------------------------------------------------------  0
sp|P12287|BLAR_BACLI                TEAIGQHVNWVQDFSLSIEQSSSKMIDSAFFAVWILGVAVMLLATLYSNLKIGKIKKNLQ  147

GCF_900143065.1|WP_072783046.1      ASGSPLDAAA-HVPL----------------PDVIEVAAPISPMLQGLRKPRLLLPRHLR  192
GCF_001423125.1|WP_056159201.1      ASGRALTGAD-HAGLAPD------Q--QALPLPVIEVEVPMSPMLLGLFRPRLLLPRHLR  211
GCF_000383895.1|WP_019923896.1      ASGRALGASE-HAGFAQH------TTAQVSPLAVVEVDVAMSPMLLGLFRPRLLLPRHLR  208
GCF_001758785.1|WP_070245907.1      ASGLPLAVSE-HKGFAQH------SRTQ-VPLAVVEVDVPMSPMLLGLFRPRLLLPRHLR  203
GCF_001191005.1|WP_050407176.1      GSGHGLPQDEAHAGFGSEL-------ARASRAQVIEVDAPISPMLLGPLRPRLLLPRHLR  198
GCF_000315425.1|WP_005669363.1      RCGAP--------------------QAHPAYPAVIEVDAPIPPMLVGPFKPRLLLPRGLR  182
GCF_000381565.1|WP_026334185.1      ATGERLVPPGRHQGLDAA--------PPPPSLAVIEVDAPISPMLFGWFRPRLMLPRHLR  198
GCF_001758795.1|WP_071658926.1      ATGERLILAGRHPGLDAAA-------TRPPSLTVIEVDAPISPMLFGWFRPRLLLPRHLR  199
GCF_001758545.1|WP_071849986.1      ATGERLATPGLHHGLDAAA-------TPAPSLAIIEVDAPISPMLFGWFRPRLLLPRHLR  199
GCF_000335815.1|WP_008451277.1      ATGERLILAGRHPGLDAAA-------TPPPALAIIEVDAPISPMLFGWFRPRLLLPRHLR  199
GCF_000344615.1|WP_017879576.1      AAGCRVDALNQHDGFA---------ATLARAPAVIEVDAPISPMLFGLFRPRLLLPRHLR  197
GCF_900129765.1|WP_073218605.1      ASGVRLPAQPGTE------------TAPTSLPSIIEVDAPISPMLFGLLNPRLLLPRHLR  197
GCF_001425045.1|WP_057157847.1      ASGSRLPAPIPAASQ----------PAPAAAPIVIEVDAPISPMLFGLRKPRLLLPRHLR  198
GCF_001425265.1|WP_056397199.1      ASGQHLPAPPSASASAAAAAAAPAAGRPVRTPAVIEVDAPISPMLFGLFQPRLLLPRHLR  208
GCF_001758545.1|WP_071849985.1      ------------------------------------------------------------  0
GCF_001425265.1|WP_056397197.1      ------------------------------------------------------------  0
GCF_900129765.1|WP_073220251.1      ------------------------------------------------------------  0
sp|P12287|BLAR_BACLI                IVNNKELLSLFHTCKE--------EIRFHQKVILSRSPLIKSPITFGVIRPYIILPKDIS  199
```

```
GCF_900143065.1|WP_072783046.1      TFDPLQQQLIVEHELTHWRRHDLCWSVAAFALQSLFWFNPFMRLLRARLGWAQEFGCDRD 252
GCF_001423125.1|WP_056159201.1      EVEVFQQRLIVAHELTHWRRRDLHWSAAALVLQSLFWFNPFMRLLGARLGWAQEFGCDRD 271
GCF_000383895.1|WP_019923896.1      GFDTLQQQLIVEHELTHWRRRDLHWSAAALLLQSLFWFNPFMRLLGARLGWAQEFGCDRD 268
GCF_001758785.1|WP_070245907.1      GFDTLQQQLIVEHELTHWRRRDLHWSAAALLLQSLFWFNPFMRLLGARLGWAQEFGCDRD 263
GCF_001191005.1|WP_050407176.1      EFDAMQQQMIVEHELTHLRRRDLQWMTLGLVLQTLLWFNPFMRLLRASLGWAQELGCDRD 258
GCF_000315425.1|WP_005669363.1      DIDPLQRELIVAHELTHWRRGDLWWLTVGAALQALCWFNPAMRLLRDKLAWAQELGCDRD 242
GCF_000381565.1|WP_026334185.1      SFEPLQQQMIVEHELMHLRRHDLQWMSAGIVLQTLLWFNPFMRLLRDKLAWAQELGCDRD 258
GCF_001758795.1|WP_071658926.1      SFDPGQQQMIVEHELTHLRRHDLQWMSAGIVLQTLLWFNPFMRLLRDNLAWAQELGCDRD 259
GCF_001758545.1|WP_071849986.1      SFDPAQQQMIVEHELTHLRRHDLQWMSAGIVLQTLLWFNPFMRLLRDNLAWAQELGCDRD 259
GCF_000335815.1|WP_008451277.1      SFDPEQQQMIVEHELTHLRRHDLQWMSAGIVLQTLLWFNPFMRLLRDNLAWAQELGCDRD 259
GCF_000344615.1|WP_017879576.1      GFDVLQQQMIVEHELTHLRRRDLHWMSAGVLLQTLLWFNPFMRLLRAKLSWAQELGCDRD 257
GCF_900129765.1|WP_073218605.1      GFDAKQQQLIIEHELMHWRRRDLHWMSVGIALQSLLWFNPFMRMLRNRLSWAQELGCDRD 257
GCF_001425045.1|WP_057157847.1      SFDAAQQQLIVEHELTHWRRRDLQWMSIGIVLQTLLWFNPFMRLLRSSLSWAQELGCDRD 258
GCF_001425265.1|WP_056397199.1      SFDPAQQQLIVEHELTHWRRRDLQWMSVGIALQTLLWFNPFMRLLRGSLSWAQELGCDRD 268
GCF_001758545.1|WP_071849985.1      ------------------------------------------------------------ 0
GCF_001425265.1|WP_056397197.1      ------------------------------------------------------------ 0
GCF_900129765.1|WP_073220251.1      ------------------------------------------------------------ 0
sp|P12287|BLAR_BACLI                MFSADEMKCVLLHELYHCKRKDMLINYFLCLLKIVYWFNPLVWYLSKEAKTEMEISCDFA 259

GCF_900143065.1|WP_072783046.1      VLRGRPPAERKAYAAALVAQFKLQLRP------------ADMALAFGASDAGA---HAPT 297
GCF_001423125.1|WP_056159201.1      VLRGRPPAERKAYAAALVAQLRWQHRP------------AGMALAFGAHDGGA---GTST 316
GCF_000383895.1|WP_019923896.1      VLRGRPPAERKAYAAALVAQLKLQYRP------------AGMALAFGASEANG--GHAPT 314
GCF_001758785.1|WP_070245907.1      VLRGRPSAERKAYAAALVAQLKLQCRP------------AGMALAFGASDAGSARTDAPT 311
GCF_001191005.1|WP_050407176.1      VLRGRPAAQRKVYAAALLAQLKLQVRP------------PEMALAFGSI-------DAST 299
GCF_000315425.1|WP_005669363.1      VLRGRPSFERRAYAAALLAQLRMQHRV------------VHGALAFGGV-------SPDT 283
GCF_000381565.1|WP_026334185.1      VLRHRPPALRRAYAAALVGQLRLQPHPATHAAT----HSATTALAFGGV-------SART 307
GCF_001758795.1|WP_071658926.1      VLRHRPSAQRKAYAAALVAQLRLQPHPATHPATHPATHLANTALAFGGV-------CART 312
GCF_001758545.1|WP_071849986.1      VLRNRPSAQRKAYAAALVAQLRLQPHPATHP--------ANTALAFGGV-------CART 304
GCF_000335815.1|WP_008451277.1      VLRNRPQAQRKAYAAALVAQLRLQPQSV------------KAALAFGGV-------SART 300
GCF_000344615.1|WP_017879576.1      VLRGRPQAQRKAYAAALVAQLRMQRGP------------TPAALAFGGV-------GAST 298
GCF_900129765.1|WP_073218605.1      VLRSRAPAQRKAYAAALVAQLRLQRDA------------RQTALAFGAV-------CAGT 298
GCF_001425045.1|WP_057157847.1      VLRGRPPAQRKAYAAALVAQLRLQHGT------------PKTALAFGGV-------CAST 299
GCF_001425265.1|WP_056397199.1      VLRGRPPAQRKAYAAALVAQLRLQHGA------------PKTALAFGGV-------CAST 309
GCF_001758545.1|WP_071849985.1      ------------------------------------------------------------ 0
GCF_001425265.1|WP_056397197.1      ------------------------------------------------------------ 0
GCF_900129765.1|WP_073220251.1      ------------------------------------------------------------ 0
sp|P12287|BLAR_BACLI                VLKTLDKKLHLKYGEVILKFTSIKQRTSSLL----------AASEFSSS--------YKH 301
```

6

```
GCF_900143065.1|WP_072783046.1      LAARISLIRTPTTERRR---WPRWLALASLTAVAVVSVALQPALGWRSAEVQ--------  346
GCF_001423125.1|WP_056159201.1      LAARIGLIRTPATARGA---WPRALALASLAAVAVANFALQPALAWQAAGPAIEPRPLLA  373
GCF_000383895.1|WP_019923896.1      LAARIGLIRTPATARGA---WSRWVALGSLAAVAIANFALQPALAWQAAEPAIEPVRLLA  371
GCF_001758785.1|WP_070245907.1      LAARIGLIRTPATAR-A---WPRWVALASLAAVALANFALQPALAWQAAEPVIEPARLLA  367
GCF_001191005.1|WP_050407176.1      LASRLALIRQPGSALRGR--WARWAGVAALAGLAAGNFALQSALAGNVAPDLE-------  350
GCF_000315425.1|WP_005669363.1      LAARVELIRKPGAARHAA--WARGAGLAILALAFGGNLALQPALAWSNPPA---------  332
GCF_000381565.1|WP_026334185.1      LATRISLIREPAAAPRGP--WARGAAVAGLAGVFAASLAFQPALADRA--PVQ-------  356
GCF_001758795.1|WP_071658926.1      VAARISLIREPGATPRGP--WARAATITGLAGVFATSLAFQPALADRAATPAH-------  363
GCF_001758545.1|WP_071849986.1      VAARISLIREPGATPRGP--WARAATITGLAGVFATSLAFQPALADRSA-PAL-------  354
GCF_000335815.1|WP_008451277.1      VAARISLIREPGATPRGA--WARAATITGLAGVFATSLAFQPALADRAATPAL-------  351
GCF_000344615.1|WP_017879576.1      LAQRIALIRQPGTASRRP--WGRCAALAGLAGIVGATLAFQPALAWRIDPVAAAGPALDK  356
GCF_900129765.1|WP_073218605.1      LAARIALIREPSNGRSRGADAARIFSIAALASLFGASLAFQPALAWRSDPSVA-------  351
GCF_001425045.1|WP_057157847.1      LASRIALIREPGRAQGRRARAARLLALAGLACVFAASLAFQPALAWRIAPAAS-------  352
GCF_001425265.1|WP_056397199.1      LASRIALIREPARAQGRRARAARLLALAGLAGVFAASLALQPALAWRIAPAAS-------  362
GCF_001758545.1|WP_071849985.1      -------------------MNFRHIILG--------------------------------  9
GCF_001425265.1|WP_056397197.1      --------------------MIAMAALG--------------------------------  8
GCF_900129765.1|WP_073220251.1      ------------------------------------------------------------  0
sp|P12287|BLAR_BACLI                IKRRIVTVVNFQTAS--PLLKAK--SALVFTLVLGAILAGTPSVSILA---MQKETRFLP  354


GCF_900143065.1|WP_072783046.1      --------------DASAGAALALDCTVMVDAANGASLVREGTCGERVTPASTFKIAISL  392
GCF_001423125.1|WP_056159201.1      SPAVND-----ARLAAAPAAPAVIDCTVMVDAASGATLVREGTCDARVTPASTFKIAISL  428
GCF_000383895.1|WP_019923896.1      RHSP-----DIAAAQPIATASATLDCTMMVDAASGAALVREGTCDASVTPASTFKIAISL  426
GCF_001758785.1|WP_070245907.1      GHNPAQHSPDTTAPPAAATTPASLDCTMLVDAASGVALVREGTCDASVTPASTFKIAISL  427
GCF_001191005.1|WP_050407176.1      -------------------ALAAIRCTQLMDAASGRVLQREGQCEARVTPASTFNIAVSL  391
GCF_000315425.1|WP_005669363.1      ----------------------SPDCTLMLDAASGARLVEEGDCDVRATPASTFNIAVSL  370
GCF_000381565.1|WP_026334185.1      -----------------AAAPATFSCTDMVDAASGAQLLRDGHCDERVTPASTFNIAVSL  399
GCF_001758795.1|WP_071658926.1      -----------------AATPATFSCTEMVDAASGKRLVHDGLCDERVTPASTFNIAVSL  406
GCF_001758545.1|WP_071849986.1      -----------------AATPATFSCTEMVDAASGKRLVHDGLCDERVTPASTFNIAVSL  397
GCF_000335815.1|WP_008451277.1      -----------------AATPATFSCTEMVDAASGKRLVHDGLCDERVTPASTFNIAVSL  394
GCF_000344615.1|WP_017879576.1      ----------ALWPFTPATPQGTISCTELVDAASGERLVHEGQCEQRVTPASTFNIPVSL  406
GCF_900129765.1|WP_073218605.1      --------------------QAAFSCTLLVDAASGAQLVRDGHCDEQVTPASTFNIPVAL  391
GCF_001425045.1|WP_057157847.1      --------------------QVPFTCTVIADAASGRQLAREGHCDERVTPASTFNIPVAL  392
GCF_001425265.1|WP_056397199.1      --------------------QAPFSCTVLADAASGQAPAREGHCDERVTPASTFNIVVAL  402
GCF_001758545.1|WP_071849985.1      ---------ALASLVPISAAHATEVCTALADS-NGPTLFQRGDCQRQVTAASTFKIAISL  59
GCF_001425265.1|WP_056397197.1      ---------CVAGLAG-VPAHGAEICTAIADAATGKVLMQRGDCQRQVTPASTFKIPLSL  58
GCF_900129765.1|WP_073220251.1      --------------------------------------MQRGDCQRQVTPASTFKIPLSL  22
sp|P12287|BLAR_BACLI                GTNVEYE--DYSTFFDKFSASG--GFVLFNSNRKKYTIYNRKESTSRFAPASTYKVFSAL  410
                                                                           .   .    : ***:::  :*
```

```
GCF_900143065.1|WP_072783046.1      MGFDSGVLRDEHAPYLPYQESYASSNPSWRHGTDPAGWLRESIVWYSQQVTSQLGAGSVR 452
GCF_001423125.1|WP_056159201.1      MGFDSGVLRDDHAPYLPYKASYASSNPSWRHGTDPAGWLRESIVWYSQQVTRRLGAASVR 488
GCF_000383895.1|WP_019923896.1      MGFDSGVLRDDHAPYLPYKASYASPNPSWRHGTDPAGWLRESIVWYSQQVTKRLGPASVR 486
GCF_001758785.1|WP_070245907.1      MGFDSGVLRDDHAPYLPYKASYASSNPGWRHGTDPAGWLRESIVWYSQQVTKRLGAASVR 487
GCF_001191005.1|WP_050407176.1      MGYDSGFLRDEHTPVLPFKEGYPAWIPEWRQDLDPSGWIKYSSVWYAQQVTRQLGAARFQ 451
GCF_000315425.1|WP_005669363.1      LGYDAGILVDAHTPALPFKPGYIDWLPAWRATTDPTSWIRSSTVWYAQQVTARLGLDGLQ 430
GCF_000381565.1|WP_026334185.1      MGYDSGILRDAHSPSLPFKPGYADWNPDWRATTDPASWIRNSTVWYAQQVTASLGAQRFR 459
GCF_001758795.1|WP_071658926.1      MGYDSGILRDAHAPSLPFKPGYIDWNPDWRATTDPTSWIRNSTVWYAQQVTAGLGARRFQ 466
GCF_001758545.1|WP_071849986.1      MGYDSGILRDAHSPSLPFKPGYIDWNPDWRATTDPTSWIRNSTVWYAQQVTAGLGARRFQ 457
GCF_000335815.1|WP_008451277.1      MGYDSGILRDAHSPSLPFKPGYIDWNPDWRATTDPTSWIRNSTVWYAQQVTAGLGARRFQ 454
GCF_000344615.1|WP_017879576.1      MGYDSGILRDEHTPKLPYRAGYVNWNPSWRAATDPTSWLKNSVLWYAQQVTLQLGAARFQ 466
GCF_900129765.1|WP_073218605.1      MGFDSGILQDEHAPMLPFKTGYPAYIPSWQADTDPSTWLQNSVLWYAQQITTRLGAKRFQ 451
GCF_001425045.1|WP_057157847.1      MGYDSGILQHEHAPLMPFKTGYPAYVPSWRADTDPSGWLHNSVLWYAQQVTATLGAARFQ 452
GCF_001425265.1|WP_056397199.1      MGYDSGILRDQHAPVLPFKAGYPAYIPSWRAATDPAGWLQNSVLWYAQQVTRQLGAARFQ 462
GCF_001758545.1|WP_071849985.1      MGYDAGILKDQRTPKLPFREGYVDWRADWRQDTDPTMWMTNSVVWYSQQVTQQLGMQRFA 119
GCF_001425265.1|WP_056397197.1      MGYDAGFLTDEHAPQLPFRRGDPDWRPSWRSATDPAKWMSESVVWYSQRITVALGQARFA 118
GCF_900129765.1|WP_073220251.1      MGYDAGFLKDTQTPELPFRQGYVDWRPSWRSATAPAKWMSESVVWYSQQITQSLGKKRFA 82
sp|P12287|BLAR_BACLI                LALESGIITKNDSHMTWDGTQY--PYKEWNQDQDLFSAMSSSTTWYFQKLDRQIGEDHLR 468
                                    :. ::*.: .  :               *.        :  *  ** *::   :*   .

GCF_900143065.1|WP_072783046.1      NYVQSFEYGNRDIASVAGVDDAVAFSELSPTLRISALEQAAFLRKVVNRSLPLSAHAYDM 512
GCF_001423125.1|WP_056159201.1      GYVQAFDYGNRDLSSVAGVTEAVAVSELSPTLRISPQEQTVFLRKVVNRKLPLSPHAYEA 548
GCF_000383895.1|WP_019923896.1      GYVQAFDYGNRNLASVAGVDDAVAVSELSPTLRITPQQQTDFLRKVVNRELAVSQQAYDV 546
GCF_001758785.1|WP_070245907.1      NYVRAFDYGNRTLASVAGVADAVAVSELSPTLRITPQQQTEFLRKVVNRELALSPQAYDV 547
GCF_001191005.1|WP_050407176.1      RYIADFGYGNRDVAGDAGADNGLGYAWINSSLKISGDEQVAFLGRMARRELPLQPQAYEM 511
GCF_000315425.1|WP_005669363.1      SYVRRFDYGNQDLSG------GVADAWIGSSLQISAQEQAAFLRKVVNRELGLNPHAYDM 484
GCF_000381565.1|WP_026334185.1      QYVRGFGYGNLDVSGDPGKDNGLAMSWIASSLKISPAEQTAFLRKIVNRQLPLSAHAYDM 519
GCF_001758795.1|WP_071658926.1      HYVNSFGYGNRDVSGDAGKDNGLAMAWIESSLKISATEQTAFLRKIVNRQLPLSAHAYDM 526
GCF_001758545.1|WP_071849986.1      QYLNSFGYGNLDVSGDAGKDNGLAMSWIASSLKISAAEQTAFLRKVVNRQLPLSAHAYDM 517
GCF_000335815.1|WP_008451277.1      QYLNSFDYGNLDVSGDAGKDNGLAMSWIASSLKISAAEQTAFLRKVVNRQLPLSAHAYDM 514
GCF_000344615.1|WP_017879576.1      RYVKDFHYGNHDVAGDAGKDNGLTLSWVSSSLKISPVEQVAFLRNVVNRELPLTAKAYDM 526
GCF_900129765.1|WP_073218605.1      DYVQGFSYGNQDLGGDPGKDNGLVQSWVSSSLRISPSEQVNFLRKVANRELPLSPQAYAK 511
GCF_001425045.1|WP_057157847.1      HYVQRFGYGNQDLSGDPGKDNGLSLAWVGSSLRISPLEQVAFLRKVANRELPLSAHAYAM 512
GCF_001425265.1|WP_056397199.1      QYVQRFGYGNQDLAGEPGQDNGLTQSWVGSSLRISPLEQVAFLRKVARRELPLSAHAYAM 522
GCF_001758545.1|WP_071849985.1      AYTSQFKYGNANVAGDAEHD-GLTLSWISSSLKISPLEQLDFLNKVVNRQLGVSAHAYDM 178
GCF_001425265.1|WP_056397197.1      AYTRRFEYGNADVAGDARND-GLTASWLGSSLRISPLGQLSFLGRVVNRQLGVSEKAYEM 177
GCF_900129765.1|WP_073220251.1      EYTTRFNYGNADVSGDAGHE-GLTAAWLESSLRISPLEQLSFLGKVVNRQLGVSEHAYAM 141
sp|P12287|BLAR_BACLI                HYLKSIHYGNEDFSVP-------ADYWLDGSLQISPLEQVNILKKFYDNEFDFKQSNIET 521
                                     *   : ***  ..             :  :*:*:   *  :* ..  ..: .
```

```
GCF_900143065.1|WP_072783046.1      TARLLKLDQPVNGWEVYGKTGTAAVRLPDGSEDQAQDIGWFVGWAVKDGRTVVFARLLQH 572
GCF_001423125.1|WP_056159201.1      TARLLKLDAMPAGWEVHGKTGTAPVQLADGRTDRDNNIGWFVGWTIRDGRTLVFARLMQY 608
GCF_000383895.1|WP_019923896.1      TARLLKVDAAPNGWEVHGKTGTAPVRLANGRADRDNNIGWFVGWTVRDGRKLVFARLMQH 606
GCF_001758785.1|WP_070245907.1      TARLLKVEETPNGWEVHGKTGTAPVRLADGSADKDNYIGWFVGWTIKDGRKLVFARLMQY 607
GCF_001191005.1|WP_050407176.1      SARLFKLASFANGWEVYGKTGTGYPVKADGKEDKTRAYGWFVGWAAKGGRTIVFAYLVQD 571
GCF_000315425.1|WP_005669363.1      TETLLRLPALPNGWDVYAKTGTAVLEQPKGAQDPPRSYGWFVGWARRDGRTIVFARLILD 544
GCF_000381565.1|WP_026334185.1      TARLTALGALPNGWQIHGKTGTASPVLADGGDDRRHSYGWFVGWASKGGRTVVFSRLVLE 579
GCF_001758795.1|WP_071658926.1      TARLTALGTLPNGWQLHGKTGTASPVLADGSDDPRHSYGWFVGWAAKDGRTVVFSRLVLA 586
GCF_001758545.1|WP_071849986.1      TARLTALGTLPNGWQIHGKTGTASPVLADGSDDPRHSYGWFVGWATKDGRTVVFSRLVLA 577
GCF_000335815.1|WP_008451277.1      TARLTALGALPNGWQLHGKTGTASPVLADGSDDPQHSYGWFVGWATKDGRTVVFSRLVLA 574
GCF_000344615.1|WP_017879576.1      TLRIMQSDTLANGWEVHGKTGTASPVLPDGRDDEAHQYGWFVGWAKKDGRTIVFARLAQD 586
GCF_900129765.1|WP_073218605.1      TERILPQQTLGNGWHVVGKTGTASALLPDGGDDATRQYGWYVGWAKKGRRTVVFARLVLD 571
GCF_001425045.1|WP_057157847.1      TARIMPQQTLANGWQVTGKTGTASALLPDGSEDGTRQYGWYVGWATKGQRTVVFARLAMD 572
GCF_001425265.1|WP_056397199.1      TESIMPRQTLANGWEVHGKTGTASALLPDGSEDGTRQYGWYVGWASKGQRTVVFARLVLD 582
GCF_001758545.1|WP_071849985.1      TARLTQRDQPLAGWRIHGKTGAAS------------GYGWYVGWATKGKRSFSFAHLMQR 226
GCF_001425265.1|WP_056397197.1      TARLTRYGQPVEGWSVNGKTGSGS------------GFGWYVGWAEKGGRKYVFARLIEK 225
GCF_900129765.1|WP_073220251.1      TAQLTQWGQSPDGWRIHGKTGSGD------------GYGWYVGWASKGARAYVFARLIQK 189
sp|P12287|BLAR_BACLI                VKDSIRL-EESNGRVLSGKTGTSVI-------NGELHAGWFIGYVETADNTFFFAVHIQG 573
                                                *  : .***:.               **::*:.    .   *:

GCF_900143065.1|WP_072783046.1      PVE--ADLYAGRQTRDAFLRELAQRVL--------------------------------- 597
GCF_001423125.1|WP_056159201.1      PVQ--SAGYAGPKTRAAFLDELAQRSL--------------------------------- 633
GCF_000383895.1|WP_019923896.1      PVT--SESYAGLKTREAFLGELAQRSL--------------------------------- 631
GCF_001758785.1|WP_070245907.1      PAD--SNSYAGLKTRQAFLGELAQRTL--------------------------------- 632
GCF_001191005.1|WP_050407176.1      QKE--EEGAAGPRLRAAVLNQLPAQLETL------------------------------- 598
GCF_000315425.1|WP_005669363.1      RQH--PDRAAGPRLKEAFLRELPSRLDAL------------------------------- 571
GCF_000381565.1|WP_026334185.1      DTQ--A-DAAGPRTRDAFLRELPAQLDTL------------------------------- 605
GCF_001758795.1|WP_071658926.1      DKQ--AGSAAGPRTRDAFLRDLPAQLDAL------------------------------- 613
GCF_001758545.1|WP_071849986.1      DKQ--AGSAAGPRTKDAFLHDLPAQLDAL------------------------------- 604
GCF_000335815.1|WP_008451277.1      DKQ--AGSAAGPRTRDAFLRDLPALLDAL------------------------------- 601
GCF_000344615.1|WP_017879576.1      PQR--QTGAAGPRAKAAFLRDLPARLDAL------------------------------- 613
GCF_900129765.1|WP_073218605.1      AKQ--ADAMGGARARAAMLRDLPPQLDRL------------------------------- 598
GCF_001425045.1|WP_057157847.1      AKQ--EGAMGGPRSREALLRELPARLDAF------------------------------- 599
GCF_001425265.1|WP_056397199.1      ARQ--EVAMGGARAREALLRELPARLDAL------------------------------- 609
GCF_001758545.1|WP_071849985.1      DDTQPKEVSTGVLAREALLKELPLLLGSVEQEALLRETVDQTILPLMKKYDVPGMALALT 286
GCF_001425265.1|WP_056397197.1      EQGEPQDVPAGVLARDGLVAEFPALANAIE--------VDQAFKPLLEKHGLPGMAVALS 277
GCF_900129765.1|WP_073220251.1      DKSDAADVPGGMLARDSLMEEFPALVNGIA--------VDQTMRPLMQENDIPGMAVAVS 241
sp|P12287|BLAR_BACLI                EKRAAG-----SSAAEIALSIL-------DKKG---------IYPSVSR----------- 601
                                                      :  :
```

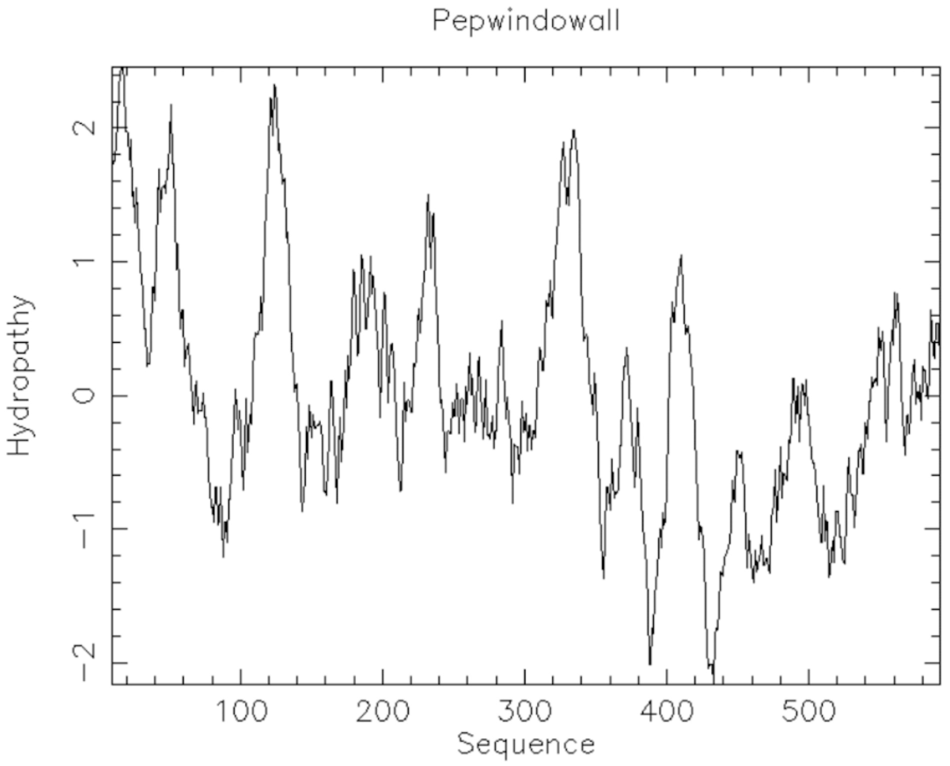
```
GCF_900143065.1|WP_072783046.1      ------------------------------------------------------------ 597
GCF_001423125.1|WP_056159201.1      ------------------------------------------------------------ 633
GCF_000383895.1|WP_019923896.1      ------------------------------------------------------------ 631
GCF_001758785.1|WP_070245907.1      ------------------------------------------------------------ 632
GCF_001191005.1|WP_050407176.1      ------------------------------------------------------------ 598
GCF_000315425.1|WP_005669363.1      ------------------------------------------------------------ 571
GCF_000381565.1|WP_026334185.1      ------------------------------------------------------------ 605
GCF_001758795.1|WP_071658926.1      ------------------------------------------------------------ 613
GCF_001758545.1|WP_071849986.1      ------------------------------------------------------------ 604
GCF_000335815.1|WP_008451277.1      ------------------------------------------------------------ 601
GCF_000344615.1|WP_017879576.1      ------------------------------------------------------------ 613
GCF_900129765.1|WP_073218605.1      ------------------------------------------------------------ 598
GCF_001425045.1|WP_057157847.1      ------------------------------------------------------------ 599
GCF_001425265.1|WP_056397199.1      ------------------------------------------------------------ 609
GCF_001758545.1|WP_071849985.1      DHGKNYVFNYGLASRETRQPVDRDTLFEVGSVSKTLVATLATYAQAQGRLALSDKVSQHM 346
GCF_001425265.1|WP_056397197.1      VNGKHYFYNYGVASQETGQPVSEATLFELGSVSKTFTVTLAAYAQAQGRLALTDPVSRHL 337
GCF_900129765.1|WP_073220251.1      VNGKHYFYHYGVASKETGQPVTNATLFEIGSLSKTFTATLATYAQAQGKLAMTDAVSQHV 301
sp|P12287|BLAR_BACLI                ------------------------------------------------------------ 601

GCF_900143065.1|WP_072783046.1      ------------------------------------------------------------ 597
GCF_001423125.1|WP_056159201.1      ------------------------------------------------------------ 633
GCF_000383895.1|WP_019923896.1      ------------------------------------------------------------ 631
GCF_001758785.1|WP_070245907.1      ------------------------------------------------------------ 632
GCF_001191005.1|WP_050407176.1      ------------------------------------------------------------ 598
GCF_000315425.1|WP_005669363.1      ------------------------------------------------------------ 571
GCF_000381565.1|WP_026334185.1      ------------------------------------------------------------ 605
GCF_001758795.1|WP_071658926.1      ------------------------------------------------------------ 613
GCF_001758545.1|WP_071849986.1      ------------------------------------------------------------ 604
GCF_000335815.1|WP_008451277.1      ------------------------------------------------------------ 601
GCF_000344615.1|WP_017879576.1      ------------------------------------------------------------ 613
GCF_900129765.1|WP_073218605.1      ------------------------------------------------------------ 598
GCF_001425045.1|WP_057157847.1      ------------------------------------------------------------ 599
GCF_001425265.1|WP_056397199.1      ------------------------------------------------------------ 609
GCF_001758545.1|WP_071849985.1      PALRGSSFDHIKLIHLGTHTAGEFPMQVPGNIKNYDQLMDYYRSWQQPASAAGASRTYSN 406
GCF_001425265.1|WP_056397197.1      PALRGSVFDRVSLVHLGTHTAGDFPLQLPQEITTHAQLMAYYKGWQ-PGHAPGSHRTYSN 396
GCF_900129765.1|WP_073220251.1      PQLRGSNFDHIQLLHLGTHTVGDFPMQVPLDIKTYDQLMDYYKRWQ-PGHGAGTHRTYSN 360
sp|P12287|BLAR_BACLI                ------------------------------------------------------------ 601
```

```
GCF_900143065.1|WP_072783046.1      ------------------------------------------------------------ 597
GCF_001423125.1|WP_056159201.1      ------------------------------------------------------------ 633
GCF_000383895.1|WP_019923896.1      ------------------------------------------------------------ 631
GCF_001758785.1|WP_070245907.1      ------------------------------------------------------------ 632
GCF_001191005.1|WP_050407176.1      ------------------------------------------------------------ 598
GCF_000315425.1|WP_005669363.1      ------------------------------------------------------------ 571
GCF_000381565.1|WP_026334185.1      ------------------------------------------------------------ 605
GCF_001758795.1|WP_071658926.1      ------------------------------------------------------------ 613
GCF_001758545.1|WP_071849986.1      ------------------------------------------------------------ 604
GCF_000335815.1|WP_008451277.1      ------------------------------------------------------------ 601
GCF_000344615.1|WP_017879576.1      ------------------------------------------------------------ 613
GCF_900129765.1|WP_073218605.1      ------------------------------------------------------------ 598
GCF_001425045.1|WP_057157847.1      ------------------------------------------------------------ 599
GCF_001425265.1|WP_056397199.1      ------------------------------------------------------------ 609
GCF_001758545.1|WP_071849985.1      LTIGLLGMISAQSMGLPIADAMEKQLLPALGMRQTYIKVPADQMTHYAQGYNDANAPVRV 466
GCF_001425265.1|WP_056397197.1      PGIGLLSLATAASLGVPYADAVEQTLFPALGLAHSYLRVPAGQMAQYAQGYNSKGAPVRM 456
GCF_900129765.1|WP_073220251.1      LGIGLLSIATAHSLGMPYVDAVEQTLLPALGLKHTWIKVPADEMAQYAQGYNSKGAPVRV 420
sp|P12287|BLAR_BACLI                ------------------------------------------------------------ 601


GCF_900143065.1|WP_072783046.1      ------------------------------------------------------------ 597
GCF_001423125.1|WP_056159201.1      ------------------------------------------------------------ 633
GCF_000383895.1|WP_019923896.1      ------------------------------------------------------------ 631
GCF_001758785.1|WP_070245907.1      ------------------------------------------------------------ 632
GCF_001191005.1|WP_050407176.1      ------------------------------------------------------------ 598
GCF_000315425.1|WP_005669363.1      ------------------------------------------------------------ 571
GCF_000381565.1|WP_026334185.1      ------------------------------------------------------------ 605
GCF_001758795.1|WP_071658926.1      ------------------------------------------------------------ 613
GCF_001758545.1|WP_071849986.1      ------------------------------------------------------------ 604
GCF_000335815.1|WP_008451277.1      ------------------------------------------------------------ 601
GCF_000344615.1|WP_017879576.1      ------------------------------------------------------------ 613
GCF_900129765.1|WP_073218605.1      ------------------------------------------------------------ 598
GCF_001425045.1|WP_057157847.1      ------------------------------------------------------------ 599
GCF_001425265.1|WP_056397199.1      ------------------------------------------------------------ 609
GCF_001758545.1|WP_071849985.1      HPAVLEPEAYGIKTTAADLIRFVDANLGQAALDEALRQAVEATHIGYFKVGKMTQDLIWE 526
GCF_001425265.1|WP_056397197.1      NPGVLAEEAYGVKSTTRDLIRFVDANMGLLPLEDKLARAVAATHTGYFKTGAMTQDLVWE 516
GCF_900129765.1|WP_073220251.1      NPGVLADEAYGVKSTAADLIHFLDANMGLITLDANLARAIRDTHAGYFKAGPMTQDLVWE 480
sp|P12287|BLAR_BACLI                ------------------------------------------------------------ 601
```

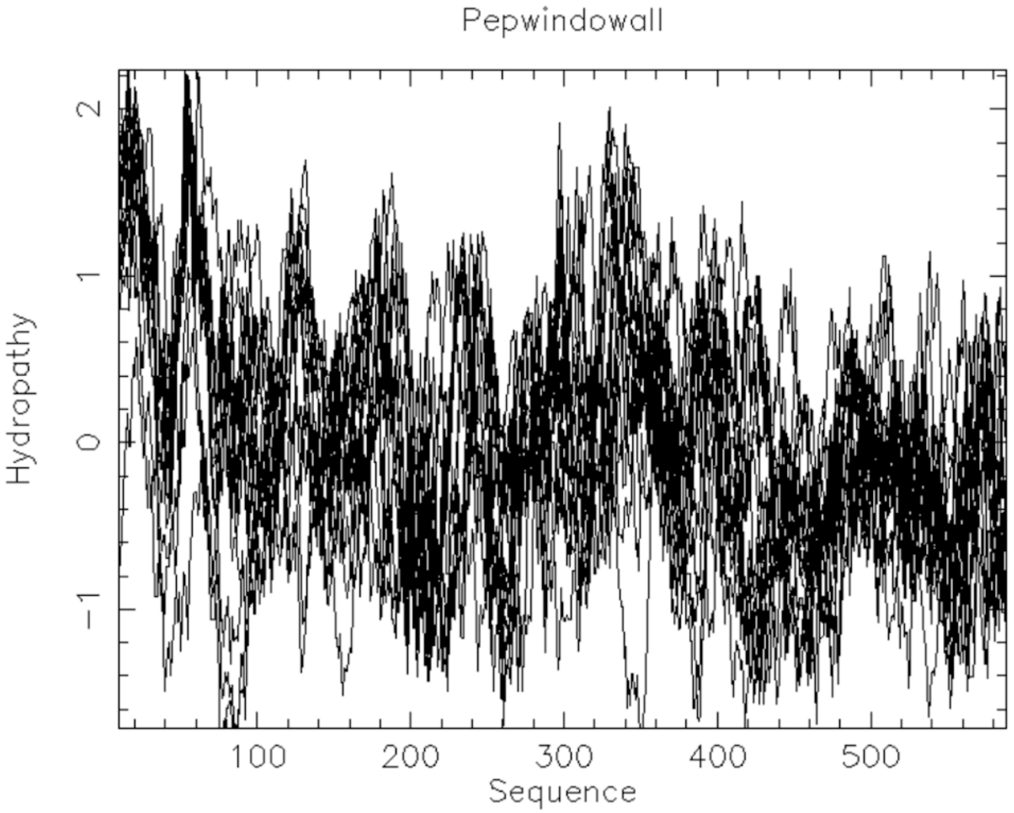
```
GCF_900143065.1|WP_072783046.1      ------------------------------------------------------------ 597
GCF_001423125.1|WP_056159201.1      ------------------------------------------------------------ 633
GCF_000383895.1|WP_019923896.1      ------------------------------------------------------------ 631
GCF_001758785.1|WP_070245907.1      ------------------------------------------------------------ 632
GCF_001191005.1|WP_050407176.1      ------------------------------------------------------------ 598
GCF_000315425.1|WP_005669363.1      ------------------------------------------------------------ 571
GCF_000381565.1|WP_026334185.1      ------------------------------------------------------------ 605
GCF_001758795.1|WP_071658926.1      ------------------------------------------------------------ 613
GCF_001758545.1|WP_071849986.1      ------------------------------------------------------------ 604
GCF_000335815.1|WP_008451277.1      ------------------------------------------------------------ 601
GCF_000344615.1|WP_017879576.1      ------------------------------------------------------------ 613
GCF_900129765.1|WP_073218605.1      ------------------------------------------------------------ 598
GCF_001425045.1|WP_057157847.1      ------------------------------------------------------------ 599
GCF_001425265.1|WP_056397199.1      ------------------------------------------------------------ 609
GCF_001758545.1|WP_071849985.1      QYPAAAGLPGLLVSASEQVTWKSNPATPLTPPLAPQADALLHKTGSTGGFGAYVLFSPGR 586
GCF_001425265.1|WP_056397197.1      QYPGHAGLDQLLVSTAEKVVFEPNPATEITPPLPPQADAWLHKTGSTGGFSAYVLFNPAR 576
GCF_900129765.1|WP_073220251.1      QYPSQAALEQLLVSTSEKMTRESNPVSTIAPPLPPQAHAWLHKTGSTGGFSAYALFNPAR 540
sp|P12287|BLAR_BACLI                ------------------------------------------------------------ 601

GCF_900143065.1|WP_072783046.1      --------------------------------------------- 597
GCF_001423125.1|WP_056159201.1      --------------------------------------------- 633
GCF_000383895.1|WP_019923896.1      --------------------------------------------- 631
GCF_001758785.1|WP_070245907.1      --------------------------------------------- 632
GCF_001191005.1|WP_050407176.1      --------------------------------------------- 598
GCF_000315425.1|WP_005669363.1      --------------------------------------------- 571
GCF_000381565.1|WP_026334185.1      --------------------------------------------- 605
GCF_001758795.1|WP_071658926.1      --------------------------------------------- 613
GCF_001758545.1|WP_071849986.1      --------------------------------------------- 604
GCF_000335815.1|WP_008451277.1      --------------------------------------------- 601
GCF_000344615.1|WP_017879576.1      --------------------------------------------- 613
GCF_900129765.1|WP_073218605.1      --------------------------------------------- 598
GCF_001425045.1|WP_057157847.1      --------------------------------------------- 599
GCF_001425265.1|WP_056397199.1      --------------------------------------------- 609
GCF_001758545.1|WP_071849985.1      KTGIVMLANKFYPGAARIEAAYSILSQLEQRRQ------------ 619
GCF_001425265.1|WP_056397197.1      KAGIVMLSNRSFSGAQRVSAGFEVLSRVAPAGPAVAPAAQSAAAN 621
GCF_900129765.1|WP_073220251.1      KVGIVILANRVLPGDQRVRAAYGLLNQLGPDAP------------ 573
sp|P12287|BLAR_BACLI                --------------------------------------------- 601
```

pepwindowall figures

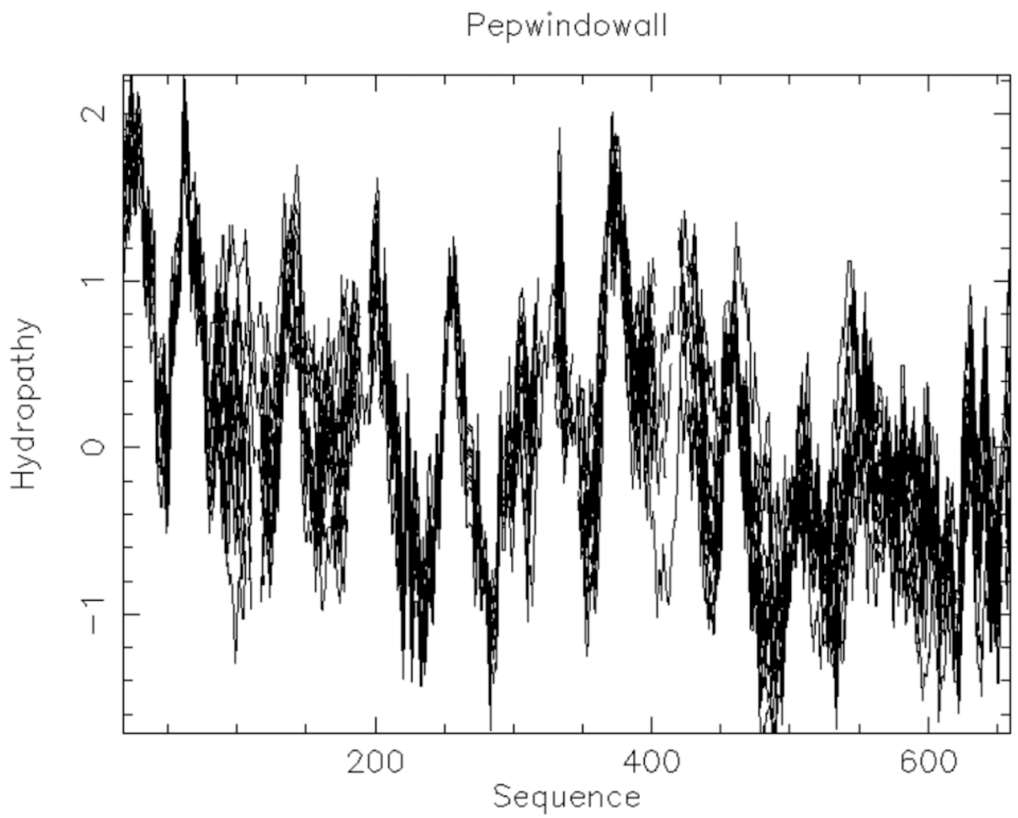
Hydrophobicity profile of BlaR from *Bacillus licheniformis* (sp|P12287|BLAR_BACLI).



Hydrophobicity profiles of the 14 unaligned long sequences (without the three class C/class D fusion proteins).



Hydrophobicity profiles of the 14 long sequences (without the three class C/class D fusion proteins) aligned with ClustalO.



2.3. Chapter 3 - Origin and Evolution of Pseudomurein Biosynthetic Gene Clusters

2.3.1. Manuscript

Origin and Evolution of Pseudomurein Biosynthetic Gene Clusters

Valérian Lupo [1,2], Célyne Roomans [1], Edmée Royen [1], Loïc Ongena [1], Olivier Jacquemin [1], Frédéric Kerff [2], Denis Baurain [1,#]

[1] InBioS-PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, Liège, Belgium
[2] InBioS, Center for Protein Engineering, University of Liège, Liège, Belgium

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

Keywords

pseudomurein, peptidoglycan, cell wall, synteny, Mur ligases, ATP-grasp, MraY-like, inter-domain HGT, phylogenetic inference

Abstract

The peptidoglycan (PG; or murein) is a mesh-like structure, which is made of glycan polymers connected by short peptides and surrounds the cell membrane of nearly all bacterial species. In contrast, there is no PG counterpart that would be universally found in Archaea, but rather various polymers that are specific to some lineages. Methanopyrales and Methanobacteriales are two orders of Euryarchaeota that harbor pseudomurein (PM) in their cell-wall, a structural analogue of the bacterial PG. Owing to the differences between PG and PM biosynthesis, some have argued that the origin of both polymers is not connected. However, recents studies have revealed that the genomes of PM-containing Archaea encode homologues of the bacterial genes involved in PG biosynthesis, even though neither their specific functions nor the relationships within the corresponding inter-domain phylogenies have been investigated so far. In this work, we devised a bioinformatic pipeline to identify all potential proteins for PM biosynthesis in Archaea without relying on a candidate gene approach. After an *in silico* characterization of their functional domains, the taxonomic distribution and evolutionary relationships of the collected proteins were studied in detail in Archaea and Bacteria through HMM similarity searches and phylogenetic inference of the Mur domain-containing family, the ATP-grasp superfamily and the MraY-like family. Our results notably show that the extant archaeal muramyl ligases are ultimately of bacterial origin, but likely diversified through a mixture of horizontal gene transfer and gene duplication. Moreover, structural modeling of these enzymes allowed us to propose a tentative function for each of them in pentapeptide elongation. While our work clarifies the genetic determinants behind PM biosynthesis in Archaea, it also raises the question of the architecture of the cell wall in the last universal common ancestor.

Introduction

The cell wall is a complex structure that surrounds most prokaryotic cells, protects them against the environment and maintains their internal turgor pressure (Pazos and Peters 2019; Meyer and Albers 2020). It also constitutes one of the striking phenotypic differences between Archaea and Bacteria. Indeed, while most archaeal species possess a paracrystalline protein surface layer (S-layer; Rodrigues-Oliveira et al. 2017), other species harbor a large variety of cell-wall polymers (e.g., sulfated heteropolysaccharides, glutaminylglycan, methanochondroitin) (Albers and Meyer 2011; Meyer and Albers 2020), whereas nearly all bacterial cell walls contain a single common polymer termed peptidoglycan (PG; also known as murein) (Vollmer et al. 2008; Pazos and Peters 2019). PG is a net-like polymer (Fig. 1) formed by long glycosidic chains of alternating N-acetylglucosamine (GlcNAc) and N-acetylmuramic acid (MurNAc) units linked by a β-(1→4) bond. To MurNAc is attached a short peptide, from three to five amino acids (AA) long, usually composed of L-alanine (L-Ala), D-glutamic acid (D-Glu), meso-diaminopimelic acid (meso-DAP) or L-lysine (L-Lys), and two D-alanines (D-Ala). This short peptide serves as a bridge between two glycosidic chains and is built at the final stage of PG biosynthesis (Vollmer et al. 2008; Pazos and Peters 2019). Interestingly, there exists an archaeal cell wall polymer that shows a three-dimensional structure similar to PG, hence named pseudopeptidoglycan or pseudomurein (PM). Compared to PG, PM (Fig. 1) contains N-acetyl-L-talosaminuronic acid (NAT) units linked to GlcNAc through a β-(1→3) bond, instead of MurNAc, and only has L-amino acids attached to NAT (König et al.

1982; König et al. 1993; Meyer and Albers 2020). Depending on the species, both PG and PM can show variation in their amino acids and glucidic composition (König et al. 1982; Vollmer et al. 2008; Pazos and Peters 2019; Meyer and Albers 2020). In the early 1990s, a PM biosynthesis pathway was proposed (Hartmann and König 1990; König et al. 1993; Hartmann and König 1994) and, due to differences between PG and PM biosynthesis, it was concluded that both polymers had evolved independently (Kandler and Konig 1993; Scheffers and Pinho 2005; Albers and Meyer 2011). In contrast to the ubiquity of PG, PM is found only in two orders of Euryarchaeota: Methanopyrales and Methanobacteriales. In recent phylogenomic reconstructions, Methanopyrales and Methanobacteriales are both monophyletic and further form a clade with Methanococcales as an outgroup, all three orders being collectively termed class I methanogens (CIM) (Bapteste et al. 2005; Williams et al. 2020). Unlike Methanopyrales and Methanobacteriales, the cell wall of Methanococcales is composed of an S-layer and does not contain PM. This restricted taxonomic distribution suggests that PM has appeared in the last common ancestor (LCA) of these two orders of methanogens, after their separation from the Methanococcales lineage, and thus that PM was not a feature of a more ancient archaeal ancestor. In other studies, Methanopyrales are basal to the whole clade of CIM (Williams et al. 2020; Aouad et al. 2022), which would point to a loss of PM in Methanococcales. However, it has been proposed that the latter topology might be caused by a long-branch attraction (LBA) artifact (Gribaldo et al. 2006; Da Cunha et al. 2018).

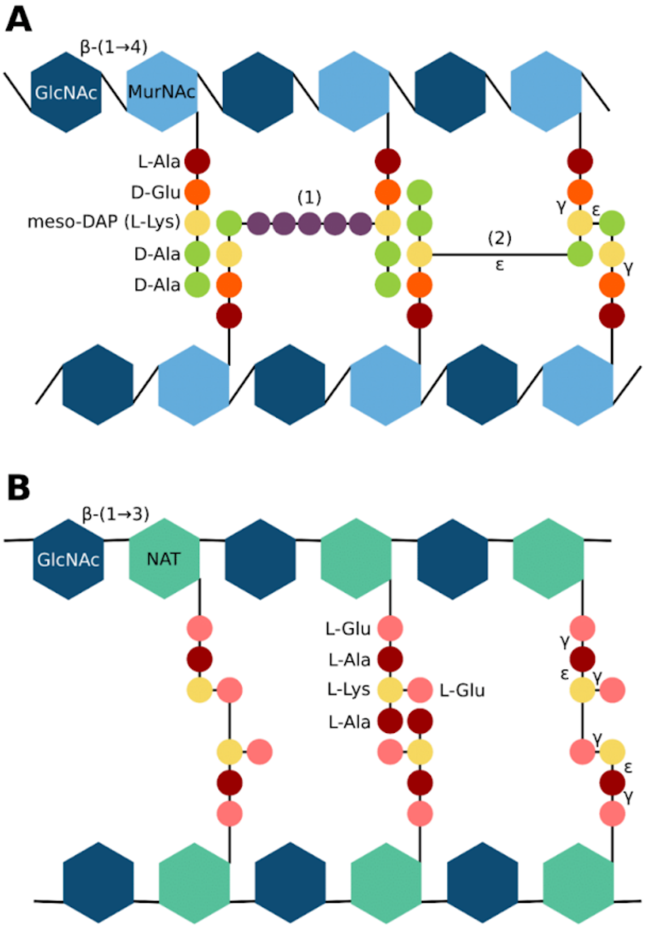


Figure 1. Structure comparison of the bacterial peptidoglycan (PG) and the archaeal pseudomurein (PM). (A) The glycosidic chain of PG is composed of alternating N-acetylglucosamine (GlcNAc) and N-acetylmuramic acid (MurNAc) units linked by a β-(1→4) bond. In most bacterial species, the pentapeptide attached to MurNAc is composed of L-alanine (L-Ala), D-glutamic acid (D-Glu), meso-diaminopimelic acid (meso-DAP; in *Escherichia coli*) or L-lysine (L-Lys; in *Staphylococcus aureus*), and two D-alanines (D-Ala). Interchain cross-linking usually occurs between the third amino acid (AA) of the first chain and the fourth AA of the second chain, accompanied by the loss of the D-Ala in position five. This cross-linking is either (1) indirect, through a pentaglycine bridge in *S. aureus*, or (2) direct in *E. coli*. **(B)** Instead of MurNAc, PM contains N-acetyl-L-talosaminuronic acid

(NAT) units linked through β-(1→3) bonds to GlcNAc units. To NAT is attached a pentapeptide composed of L-Glu, L-Ala and L-Lys. Beyond the lack of D-AA, the archaeal pentapeptide bears more ε- and γ-peptide bonds than its bacterial counterpart.

Regarding PG, it is so crucial for cell survival and growth that even bacteria once thought to lack PG, like Planctomycetes or Chlamydiae, were actually shown to synthesize a thin layer of PG, notably during septal division (Liechti et al. 2014; Jeske et al. 2015; Packiam et al. 2015; van Teeseling et al. 2015; Liechti et al. 2016). Therefore, the proteins involved in PG biosynthesis have been extensively studied over the last years, in particular as potential targets for antimicrobial agents (Bhattacharjee 2016). Usually, many genes involved in PG biosynthesis lie in the *dcw* (division and cell-wall synthesis) gene cluster. The order of the genes within this cluster is relatively well conserved across the different bacterial lineages (Tamames 2001; Mingorance and Tamames 2004; Real and Henriques 2006), even if some species lack one or more PG biosynthesis genes in their genome (Pilhofer et al. 2008; Martínez-Torró et al. 2021). A recent reconstruction of the ancestral state of the *dcw* cluster showed that the last bacterial common ancestor (LBCA) had a complete *dcw* cluster, composed of 17 genes (Léonard et al. 2022).

Among the proteins encoded by *dcw* cluster genes, the four muramyl ligase enzymes, MurC, MurD, MurE, MurF, and the D-alanine--D-alanine ligase, Ddl, are critical for PG biosynthesis. The four muramyl ligase add, respectively and successively, L-Ala, D-Glu, meso-DAP (or L-Lys) and D-Ala-D-Ala to UDP-MurNAc, while Ddl binds two D-Ala to yield the D-Ala-D-Ala dipeptide (Pazos and Peters 2019; Egan et al. 2020). Inhibiting one of those genes leads to lysis of the bacterial cell (Zawadzke et al. 2008; Kouidmi et al. 2014). The muramyl ligases belong to the ATP-dependent Mur domain-containing family, which further includes four other enzymes: 1) MurT, which forms a complex with GatD to catalyze the amidation of D-Glu to D-glutamine (D-Gln) in *Staphylococcus* species (Münch et al. 2012; Nöldeke et al. 2018), 2) CapB, which plays a role in the formation of the poly-γ-glutamic acid capsule in *Bacillus* (Makino et al. 1989; Ashiuchi 2013; Hsueh et al. 2017), 3) cyanophycin synthetase (CphA), which catalyzes the polymerisation of L-arginine (L-Arg) and L-aspartate (L-Asp) into cyanophycin, a polymer that constitutes a nitrogen reserve in Cyanobacteria (Aboulmagd et al. 2001; Sharon et al. 2021), 4) folylpolyglutamate synthase (FPGS), which is responsible for the addition of polyglutamate to folate. The FPGS enzyme is found in the three domains of life: Archaea, Bacteria and Eukarya, but not in methanogenic archaea (Levin et al. 2004; Gorelova et al. 2019; Kordus and Baughn 2019; Kordus and Baughn 2019). Ddl is part of the ATP-grasp superfamily, including at least 21 groups of enzymes (Fawaz et al. 2011). Among those, the synthetase domain of carbamoylphosphate synthetase (CPS; Shi et al. 2018), CarB, is a well-studied enzyme that has been used to root the tree of life because it results from an internal gene duplication that occurred before the Last Universal Common Ancestor (LUCA) (Lawson et al. 1996; Philippe and Forterre 1999; Cammarano et al. 2002).

With the advances in genome sequencing, homologues of genes involved in PG biosynthesis, including muramyl ligases, have been identified in Methanopyrales and Methanobacteriales (Smith et al. 1997; Slesarev et al. 2002; Samuel et al. 2007; Leahy et al. 2010). Consequently, it was suggested that, despite the difference between the two biosynthetic pathways, the evolution of PG and PM are connected. More precisely, archaeal PM could have arisen from horizontal transfers (HGTs) of PG genes from Bacteria (Graham and Huse 2008; Subedi et al. 2021; Ithurbide et al. 2022). Last year, Subedi et al. 2021 re-investigated the PM biosynthetic pathway proposed by (Leahy et al. 2010) and resolved the first structure of an archaeal muramyl ligase, which they named pMurC, after its supposed homology with bacterial MurC. These recent studies have thus led to an increase in the number of candidate genes for PM biosynthesis. However, their function and exact role in the different steps of PM biosynthesis have still to be experimentally validated.

In the present work, we used a *de novo in-silico* approach to identify candidate genes for PM biosynthesis, characterized their functional domains using various prediction software and assessed the taxonomic distribution of their homologs in both bacterial and archaeal domains. We also investigated the evolutionary origins of PM by performing phylogenetic analyses of the Mur domain-containing family, the ATP-grasp superfamily and the MraY-like family using multiple variations of the taxon sampling and different AA substitution models. Our results reveal a bacterial origin of the four main archaeal muramyl ligases, which probably traces back to two HGT events in an ancestor of Methanopyrales and Methanobacteriales, followed by one or two rounds of gene duplication, depending on the considered gene. Moreover, *in silico* structural characterization of the muramyl ligases from two model archaea allowed us to tease apart their potential functions in PM biosynthesis.

Results

Collection of potential proteins for pseudomurein biosynthesis

For the identification of candidate genes for pseudomurein (PM) biosynthesis following an approach independent of already identified genes, we used the whole proteomes of ten archaeal organisms, corresponding to five PM-containing archaea (i.e., four Methanobacteriales and one Methanopyrales) and five non-PM Euryarchaeota (i.e., one Methanococcales, two representatives from different orders of Methanomicrobia, one Archaeoglobales and one Thermoplasmatales). The protein sequences of the ten archaeal assemblies were first clustered into 6,321 orthologous groups (OGs; clusters named from OG0000001 to OG0006321). A taxonomic filter allowed us to select 82 OGs specific to the PM-containing archaea, among which 26 OGs contained sequences of all five PM-containing archaea, whereas 56 OGs contained sequences of the only Methanopyrales and three Methanobacteriales (retained to maximize the sensitivity of our search). No OG was specific to the four Methanobacteriales. The paralogue-targeting approach (see Material and Methods) allowed us to identify 20 additional OGs. In parallel, eight OGs were selected using three pseudomurein-related HMM profiles downloaded from the NCBI CDD (Conserved Domain Database) (see Material and Methods). In total, 110 OGs were thus identified as candidates for PM biosynthesis (Fig. S1).

Genetic environment of candidate proteins

Synteny analysis revealed that 22 out of 110 OGs are encoded by genes clustered in five regions of the genomes of PM-containing archaea, which we termed clusters A to E (Fig. S2). *In silico* functional analysis indicates (Table S1; sheet 1 to 3) that proteins of cluster A and B may be involved in PM biosynthesis while proteins of clusters C, D and E are probably not. Cluster C is a bidirectional cluster, where annotated proteins belong to different pathways. Indeed, OG0001177 and OG0001178 proteins are associated with pilus assembly proteins and/or surface proteins, while OG0001176 and OG0000094 can be associated with cell shape or gene regulation (the latter is not identified in our pipeline but its gene is always located downstream of the OG0001176 gene). Cluster D is related to nucleic acid metabolism or cellular signal transduction (Braun et al. 2021), whereas cluster E code for the four proteins that compose the methyl-coenzyme M reductase, which is implied in methane formation (Chen et al. 2020). Very recently, two potential clusters for PM biosynthesis were identified using bacterial proteins from PG biosynthesis as BLAST queries (Subedi et al. 2021). Those clusters correspond to our clusters A and B. Cluster A is composed of five genes: 1) OG0001014, which was experimentally characterized as the smallest CPS (Popa et al. 2012), 2) OG0001163, a type 4 glycosyltransferase homologue to MraY, 3) OG0001473, a Mur domain-containing protein, 4) OG0001162 and 5) OG0001472, two hypothetical proteins. Regarding cluster B, it is composed of three genes: 1) OG0001150, a Mur domain-containing protein, 2) OG0001147, a hypothetical protein and 3) OG0001146, a MobA-like NTP transferase domain-containing protein. In addition, two genes of Mur domain-containing proteins (i.e., OG0001148 and OG0001149) can be located either in cluster A or cluster B, and even outside any cluster, depending on the PM-containing species considered. Furthermore, another PM-specific gene (OG0000796, coding for a hypothetical protein) is located just downstream of the OG0001472 gene in the genome of *Methanopyrus sp. KOL6*, while a second one (OG0000169, coding for a Zn peptidase) is only three genes away from the OG0001146 gene in *Methanothermobacter thermautotrophicus str. Delta*. Based on the genetic environment of clusters A and B, we attempted to identify a conserved regulon for PM biosynthesis by phylogenetic footprinting (Cristianini and Hahn 2006; Anderssen et al. 2022). However, unlike in Bacteria (Anderssen et al. 2022), such analyses were unsuccessful on our archaeal dataset (Supplementary data).

Taking into account OG0000094, identified by its conserved localisation within cluster C, our pipeline recovered 23 syntenic genes (out of 111 OGs), of which half are likely to be involved in PM biosynthesis (Table 1). For clarity, in the following, the four Mur domain-containing proteins OG0001148, OG0001149, OG0001150 and OG0001473 will be arbitrary called Murα, Murß, Murγ and Murδ, respectively, without considering any specific homology with bacterial MurCDEF. Most of the proteins encoded in clusters A and B have no predicted signal peptide (SP) and are either cytoplasmic or transmembrane (TM) proteins. TM segment prediction was used as a complement to SP prediction. It allowed us to distinguish between cytoplasmic and transmembrane proteins, and revealed that only OG0000796, OG0001163 and OG0001472 are TM proteins. OG0000169 and OG0001162 feature a Sec SP and are thus the only exported proteins of these gene clusters. In PM-containing archaea, the synteny of the two genes of OG0001472 and murδ is highly conserved. However, in *Methanothermobacter thermautotrophicus str. Delta*, both genes were annotated as pseudogenes and thus not predicted as proteins.

Table 1. Overview of the proteins identified in our search for genes involved in PM biosynthesis. Orthologous Groups (OGs) composing the identified gene clusters, named clusters A to E are listed. For each OG, there is the functional prediction of InterProScan (if any), the predicted signal peptide type (SP) and the number of predicted transmembrane (TM) segments (0 = cytoplasmic, 1 = monotopic, >1 = polytopic).

Cluster	Orthologous Groups	InterProScan Prediction	Signal peptide	# TM
A	OG0001014	CPS	Other	0
	OG0001163	MraY-like	Other	>1
	OG0001473	Muramyl ligase (= Murδ)	Other	0
	OG0001162	/	Sec	0
	OG0001472	/	Other	1
	OG0000796	/	Other	>1
B	OG0001150	Muramyl ligase (= Murγ)	Other	0
	OG0001147	/	Other	0
	OG0001146	MobA-like NTP transferase domain	Other	0
	OG0000169	Zn peptidase	Sec	0
A-B	OG0001148	Muramyl ligase (= Murα)	Other	0
	OG0001149	Muramyl ligase (= Murß)	Other	0
C	OG0001210	Aminotransferases class-I	Other	0

		pyridoxal-phosphate attachment site		
	OG0000094	MreB/DnaK-like	Other	0
	OG0001176	Coiled coil protein	Other	0
	OG0001177	Flp pilus assembly protein RcpC/CpaB	Other	1
	OG0001178	Sortase E	Other	>1
D	OG0001213	Zc3h12a-like Ribonuclease NYN domain	Other	0
	OG0001214	Nucleotide cyclase	Other	>1
E	OG0000266	Methyl-coenzyme M reductase, beta subunit	Other	0
	OG0000231	Methyl-coenzyme M reductase operon protein D	Other	0
	OG0000230	Methyl-coenzyme M reductase, gamma subunit	Other	0
	OG0000229	Methyl-coenzyme M reductase, alpha subunit	Other	0

Taxonomic distribution of candidate proteins and their homologues

To ensure the completeness of the selected OGs, we looked for corresponding pseudogenes or mispredicted proteins in the genomes of the five PM-containing archaea (see Material and Methods). After completing the OGs, we retained only those containing protein sequences from all five PM-containing archaea, decreasing the number of OGs from 111 to 49. Interestingly, no OG from the five syntenic regions was discarded. Similarity searches in three local databases showed that 15 OGs are widespread (though not universal) among Bacteria and Archaea, 9 OGs have homologues only in bacteria, while 25 OGs are exclusive to archaea, among which 15 to PM-containing archaea (Fig. 2; for details see Table S2). In clusters A and B, which likely encode proteins involved in PM biosynthesis, 6 OGs are exclusive to Methanopyrales and Methanobacteriales whereas 7 OGs share homology with bacterial proteins. We also noticed that our HMM profiles of the four muramyl ligases (i.e., Murα, Murß, Murγ and Murδ) recovered a common set of sequences, indicating that Murαßγδ are specifically related. According to this taxonomic distribution, we further investigated the origin of CPS, the MraY-like and the four muramyl ligases Murαßγδ. The MobA-like NTP transferase, OG0001146, was not considered for phylogenetic analysis because, compared to the aforementioned proteins, no homologous protein was identified in the representative bacterial database (nor for OG0001215 and OG0000138). However, some bacterial homologues were identified when we determined the taxonomic distribution of the 49 OGs using the (much larger) prokaryotic database.

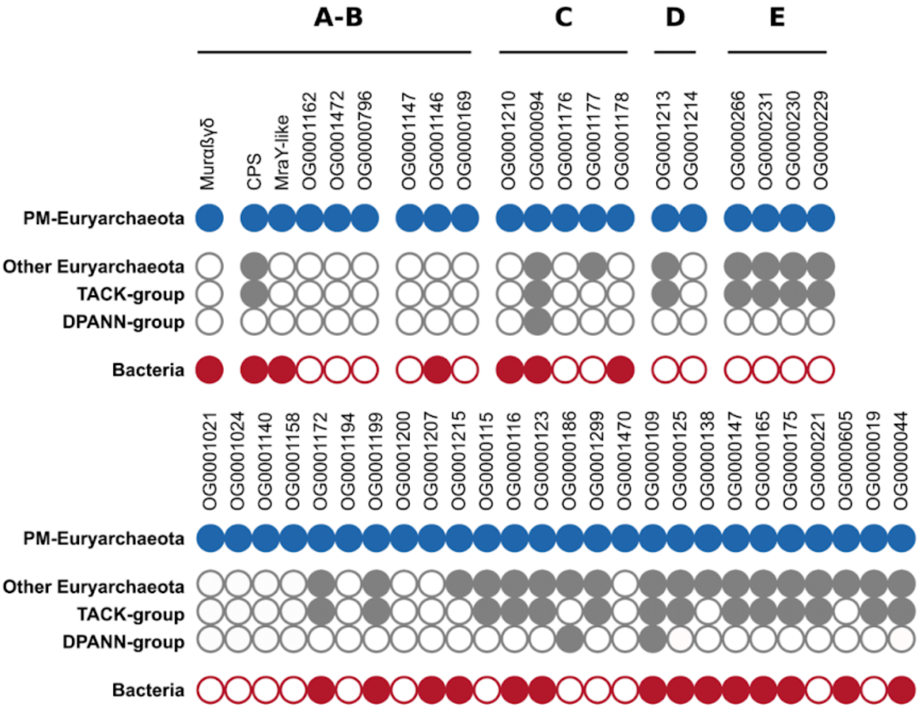


Figure 2. Taxonomic distribution patterns of the 49 retained orthologous groups (OGs). The four OGs OG0001148, OG0001149, OG0001150 and OG0001473 are considered together and referred to as Murαßγδ, OG0001014 is referred to as CPS and OG0001163 as MraY-like. Black lines delineate gene clusters in the genomes of PM-containing archaea (clusters A to E). Full circle = gene present in the taxonomic group; empty circle = gene absent from the taxonomic group.

Phylogenetic trees

ATP-grasp superfamily

The CPS from the cluster A of PM-containing archaea, as well as the Ddl from the *dcw* cluster of bacteria, are member proteins of the ATP-grasp superfamily. Due to the large number of protein functions and architectures within the ATP-grasp superfamily (Fawaz et al. 2011), we focused our phylogenetic analyses on the ATP-grasp domain. Furthermore, we wanted to investigate whether CPS is closely related to Ddl (through HGT for instance). Thus, we excluded eukaryotic ATP-grasp proteins from our analyses. In the local databases, we identified 8,013 unique protein sequences containing at least one ATP-grasp domain, which are distributed across 1387 prokaryotic organisms. ATP-grasp domains were spliced out of full-length proteins, yielding a total of 12,074 domain sequences, then sequence deduplication led to 2344 sequences from which 149 highly divergent sequences were removed. Annotation showed that 1788 domain sequences correspond to 17 members of the ATP-grasp superfamily, while 406 sequences have no similarity with reference ATP-grasp sequences (see Material and Methods). We also observed that PyC, PccA and AccC reference sequences annotate sequences belonging to the same monophyletic group. These three enzymes use hydrogenocarbonate as a substrate (Diesterhaft and Freese 1973; Shen et al. 2006; Hou et al. 2015), which could explain the phylogenetic proximity of their ATP-grasp domain sequences. Accordingly, we decided to indistinctly tag the whole group with the three annotations. A similar observation and decision were made for PurK and PurT proteins, though the former uses hydrogenocarbonate as its substrate, while the latter uses formate (Mueller et al. 1994; Marolewski et al. 1997).

Due to an internal gene duplication that occurred before LUCA (Lawson et al. 1996; Philippe and Forterre 1999; Cammarano et al. 2002), the seven phylogenetic trees (see Material and Methods) were rooted on CarB, the monophyly of which is supported by high statistical values. Despite a low topology conservation between the different evolutionary models and number of tree search iterations, some recurring patterns can be observed (Fig. 3 and Fig S3 to S8). RimK ATP-grasp domain sequences are always paraphyletic, due to the inclusion of GshB, GshAB and CphA, the latter two clustering into a smaller clan. The monophyly of Acetate--CoA ligases AcD (Musfeldt and Schönheit 2002) is maximally supported and a long branch is present at the base of the group. Except for the C40 model (Fig S5 and S6), AcD forms a clan with the Succinate--CoA ligase SucC (Joyce et al. 1999). The position of the other members of the ATP-grasp superfamily is much more elusive. For example, Pur2 (Cheng et al. 1990) emerges somewhat alone in the LG4X tree (Fig. 3), whereas it forms a clan with either AcD and SucC in the four C20 and C60 trees (Fig S3-4 and S7-8) or only with SucC in the two C40 trees (Fig S5 and S6). Similarly, albeit Ddl and CPS branch together in one C20 tree with a branch support of 63 (Fig S3), their respective positions within the ATP-grasp superfamily are unstable (Fig 3 and Fig S3 to S8). Therefore, there is no strong phylogenetic evidence for a specific relationship between the Ddl and CPS proteins. In contrast, CPS is never close to CarB, which is at odds with the less extensive phylogenetic analyses of Popa et al. 2012.

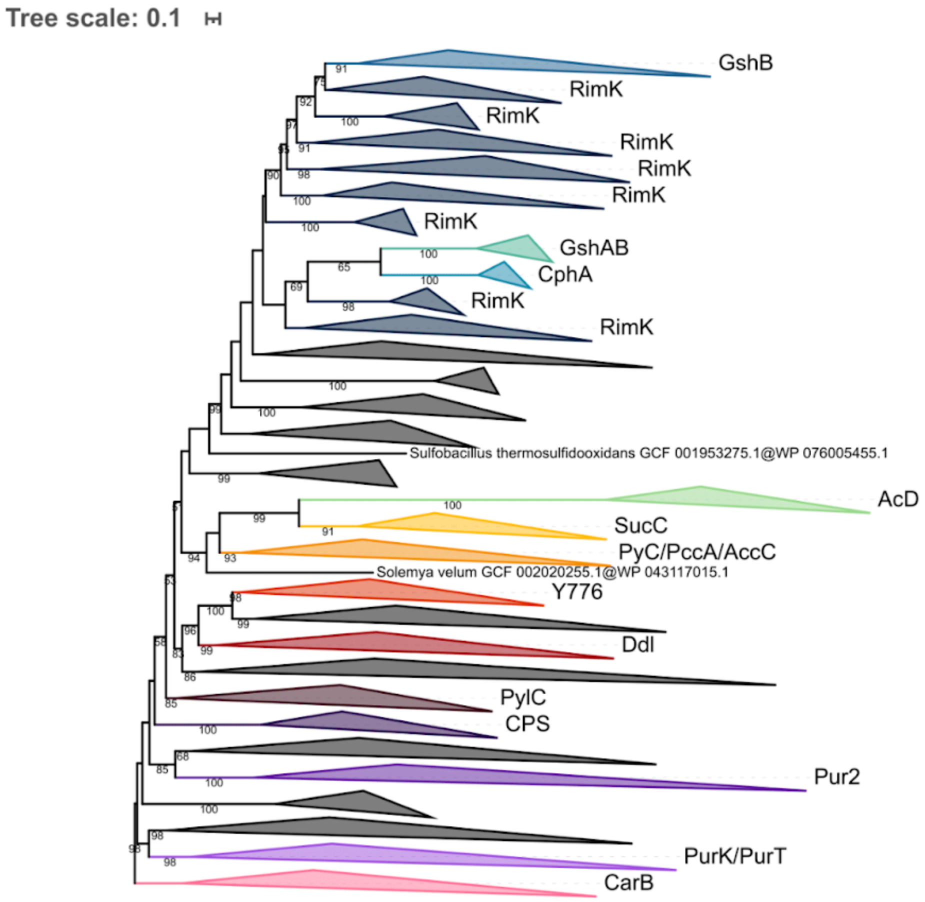


Figure 3. **Phylogenetic tree of the ATP-grasp superfamily rooted on CarB.** The tree was inferred from a matrix of 2,194 sequences x 180 unambiguously aligned AAs using IQ-TREE under the LG4X+R4 model. Tree visualization was performed using iTOL. Bootstrap support values are shown if greater or equal to 50. Branches were collapsed on sequence annotation based on reference sequences. Black collapsed branches correspond to unannotated sequences.

MraY-like family

Homology searches revealed that the bacterial homologue of OG0001163 is the glycosyltransferase 4 (GT4) MraY. According to the NCBI CDD (Lu et al. 2020), MraY is part of the MraY-like family, which further includes WecA (Amer and Valvano

2001), WbpL (Campbell et al. 1997; Price and Momany 2005), and eukaryotic and archaeal GPT (Dal Nogare et al. 1998). In addition to the MraY-like OG0001163, our pipeline has highlighted transmembrane proteins in OG0001207 (Fig. 2), for which the only bacterial homologue also has a MraY/WecA-like GT4 domain. Therefore, we decided to add the sequences of OG0001207 to the phylogenetic analysis of the MraY-like family. Although only one sequence similar to OG0001207 had been identified in the bacterial database, 62 additional bacterial OG0001207 homologues were identified in the (larger) prokaryotic database. According to the study of Lupo et al. 2021, none of the genomes coding for those protein sequences are considered as contaminated, which suggests that OG0001207 homologues genuinely exist in these bacteria. Overall, a total of 1267 sequences from the MraY-like family were identified in our databases, corresponding to 1071 unique sequences. Interestingly, 773 sequences among 1267 were identified by two or more HMM profiles of the individual members of the MraY-like family. During the annotation pipeline, six bacterial sequences remained unannotated due to their ambiguous position within the preliminary guide tree (see Material and Methods). Moreover, reference sequences of WecA and WbpL annotated putative sequences from the same monophyletic group and thus, the whole group was considered as WecA/WbpL.

Due to this non-universal taxonomic distribution and lack of an ancestral gene that could be present in the genome of LUCA, the three MraY-like family trees (see Material and Methods) were left unrooted. Phylogenetic analysis showed that each of the five members of the MraY-like family are monophyletic and all supported by high bootstrap values. Moreover, MraY formed a clan with WecA/WbpL while GPT formed a clan with OG0001163 and OG0001207 (Fig. 4). Those results are similar for the three evolutionary models LG4X, C20 and C40. Regarding the six unannotated sequences, the sequence of *Syntrophaceticus schinkii* is always basal to OG0001207, whereas the group composed of two sequences of *Ruminococcaceae* sp. and two sequences of *Treponema* sp. is always basal to MraY. The last sequence from *Ruminococcus* sp. is basal to MraY in the LG4X tree, while it is basal to WecA/WbpL in the C20 and C40 trees (Fig. S9 and S10). Taxonomic analysis revealed that MraY and WecA/WbpL are exclusive to bacteria, while GPT is only found in archaea. Regarding OG0001163, it is exclusive to

PM-containing archaea, as would be OG0001207, ignoring the few exceptions discussed above.

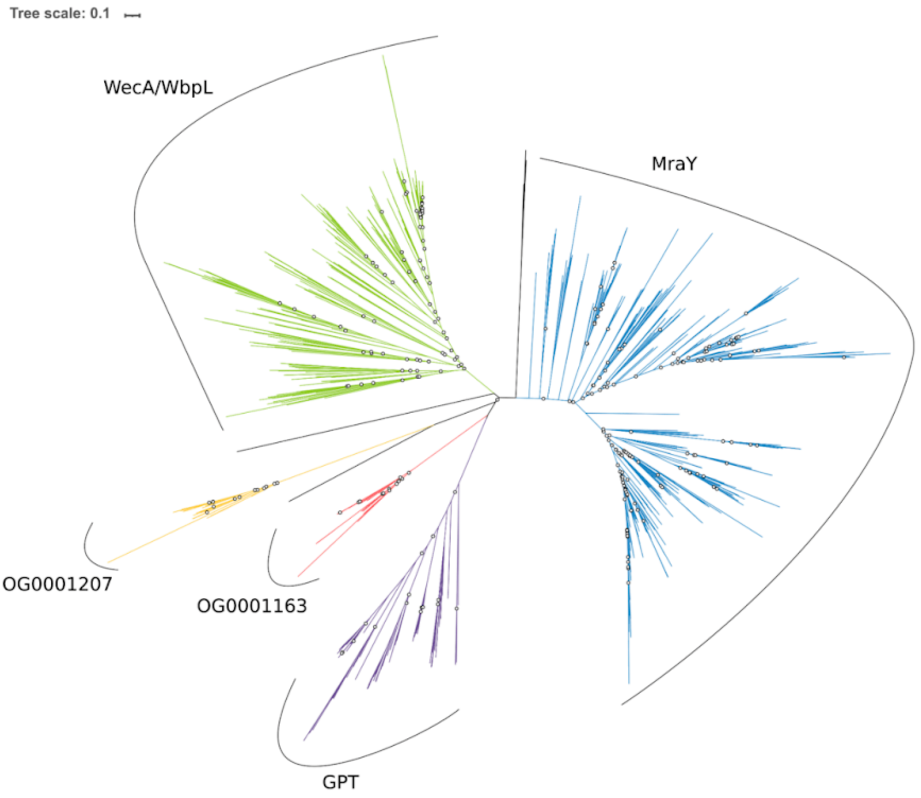


Figure 4. Unrooted phylogenetic tree of the MraY-like family. The tree was constructed from a matrix of 1,070 sequences x 408 unambiguously aligned AAs using IQ-TREE under the C40+G4 model. Open circles correspond to bootstrap support values under 90. Blue sequences correspond to a MraY annotation, green to WecA/WbpL, red to OG0001163 (MraY-like), yellow to OG0001207, purple to GPT, and black to unannotated bacterial sequences.

Mur domain-containing family

Homology searches allowed us to identify 3398 unique sequences distributed across 755 prokaryotic organisms. These sequences correspond to 12 members of the Mur domain-containing family, which are the four bacterial MurCDEF, the four archaeal

Murαßγδ, MurT, CapB, CphA and FPGS. Taxonomic distribution within each member protein group revealed that MurCDEF and CphA are specific to bacteria, Murαßγδ are specific to PM-containing archaea, while MurT, CapB and FPGS are found both in Bacteria and Archaea, albeit not universally. According to the function and ubiquity of FPGS, we assumed that a FPGS protein was already present in LUCA, and trees were rooted on the corresponding clan. The phylogenetic trees, inferred with three models from a matrix of 3407 sequences x 550 AAs including the 12 members of the Mur domain-containing family, showed that each member group is monophyletic and supported by high statistical values (bootstrap values around 100; Fig S11 to S13), except for the long-branched sequence of *Francisella noatunensis*, tagged as MurE, which is positioned basal to the CphA clan (except in the C20 tree). In spite of the solid monophyly of each Mur domain-containing family member, the recovered relationships between these members (i.e., the topology of the family tree) depend on the evolutionary model (LG4X, C20 or C40). We made the same observation for phylogenetic reconstructions based on a smaller matrix restricted to the most conserved AAs over the full-length sequence (3386 sequences x 228 AAs) (Fig. S14 to S16).

In order to investigate the orthology relationships between the four bacterial muramyl ligases MurCDEF and their uncharacterized archaeal homologues Murαßγδ, we performed phylogenetic analyses using only one out of four potential outgroups among MurT, CapB, CphA and FPGS, under the three models (Fig. 5a and Fig S17 to S27). In these trees, Murα and Murß always group together, and further form a clan with Murγ and MurD in 11 trees out of 12. Murδ groups with MurC in eight of the single-outgroup trees. Furthermore, MurαßγD and MurδC form a clan in five trees, while this larger clan further includes MurT in the three trees where the latter is present. Interestingly, the sequence of *Francisella noatunensis*, tagged as MurE, groups with CphA instead of MurE when CphA is considered during phylogenetic inference. In the CapB and CphA outgroup trees computed with the C40 model, Murδ branches inside the MurE clan, within Firmicutes. Even though such an alternative relationship would fit the structure of *Methanothermus fervidus* Murδ (PDB codes 6VR8 and 7JT8), described as a 'type E peptide ligase' (Subedi et al. 2022), the analysis of the two matrices under the more sophisticated PMSF LG+C60+G4 model (Fig S28 and S29) did not return that topology, and instead supported the first solution. Besides, two phylogenetic trees focusing on indels, with FPGS as the only outgroup (see Material and Methods), tend to confirm the first topology too (Fig 5b and S30). Indeed, when using a binary encoding, we also observe a clan formed by MurαßγD and MurδC, which is supported by a bootstrap value of 100, while MurE and MurF are paraphyletic. However, in those indel trees, Murß forms a clan with Murγ rather than Murα.

In parallel, jackknife support values from species resampling analyses (Table 2; see Table S3 for complete results and Material and Methods for details) confirmed the monophyly of each of MurC, MurD, Murα, Murß, Murγ, Murδ, CapB and FolC with jackknife support ranging between 99.7% and 100% under the three evolutionary models. Support for MurE, MurF and CphA is slightly lower and lies between 89.5 and 95.7%, whereas support for MurT is really low, with values ranging from 37.5 to 51.1% (Table 2 and Table S3). ASTRAL trees (Fig. S32 to S34) showed that the sequences of *Francisella noatunensis* (tagged as MurE) and *Solemya velum* gill symbiont (tagged as MurF) both group with CphA, which explains the lower jackknife support for the latter. When these two sequences are instead considered as belonging to CphA, support increases to 100% under the three evolutionary models. Support for MurE and MurF also increases (Table 2), which suggests that both sequences were mistagged by the annotation pipeline and rather are (divergent) CphA proteins. Furthermore, LG4X and C40 species trees revealed that MurT is polyphyletic and split into two distinct clans: 1) a large one composed of bacterial and Methanobacteriales sequences, and 2) a smaller one composed of sequences of Methanopyrales and Methanobacteriales, which we termed MurT-like. Indeed, support for MurT increases to 99.7% when MurT-like sequences are considered as a separate clan (Table S3). ASTRAL trees (Fig. S32 to S34) also confirmed the relationships between the eight muramyl ligases observed in the single-outgroup trees, even if those are blurred by the unstable positions of MurT and MurT-like. Murα and Murß are clustered in the three trees with a jackknife support of 86%, 71.3% and 66.7%, under LG4X, C20 and C40 models, respectively (Table 2). Regarding Murγ, it groups with MurT-like in LG4X (jackknife support of 39.0%) and C40 (38.9%) trees, which further form a clan with MurD (27.9% and 31.3%), whereas Murγ forms a clan with only MurD in the C20 tree (37.3%). Moreover, Murδ and MurC form a clan in the LG4X ASTRAL tree (47.5%), but are paraphyletic in the C20

(30.9%) and C40 (29.6%) trees. Murαßγ, MurD and MurT-like are grouped in the LG4X (27.5%) and C40 (15.9%) ASTRAL trees. In addition, MurαßγD, MurδC, MurT and MurT-like are grouped in the C20 (22.6%) and C40 (29.7%) trees. Symmetrically, these analyses revealed that MurE forms a clan either with MurF (30.7%, 16.4% and 14.7%) or CphA (36.8%, 34.1% and 30.6%) (Tables 2 and S3). Moreover, CapB appears to be closely related to FPGS in C20 and C40 ASTRAL trees, with a jackknife support of 62.1% and 65.5%, respectively. As expected, the clan formed by MurE, MurF, CphA, FPGS and CapB has the same jackknife support as its counterpart (MurαßγDδCTT-like) in C20 (22.6%) and C40 (29.7%) trees (Tables 2 and S3). Therefore, it appears that the primary sequences of MurEF proteins are quite distinct from the six other muramyl ligases MurαßγδCD.

Table 2. Jackknife support values computed from the 1000 replicates of species resampling under three phylogenetic models: LG4X+R4, C20+G4 and C40+G4. Specific clans are shown if the support value reaches 200‰ in at least one of the three models. Here, the two misclassified sequences of MurE and MurF are considered as CphA sequences. For complete results, see Table S3.

	Support value (‰)		
Clan	**LG4X**	**C20**	**C40**
CapB	1000	1000	1000
CphA	1000	1000	1000
FPGS	1000	998	997
MurC	1000	1000	1000
MurD	1000	1000	1000
MurE	993	930	935
MurF	990	998	994
MurT	357	511	510
Murα	1000	1000	1000
Murß	1000	1000	1000
Murγ	1000	1000	999
Murδ	1000	1000	1000
α-ß	860	713	667
D-γ	379	373	356
C-δ	475	309	296

E-F	307	164	147
CapB-FPGS	269	621	655
CapB-δ	208	108	163
CphA-FPGS	322	109	110
CphA-E	368	341	306
T-γ	87	211	261
CphA-E-F	242	141	145
CphA-FPGS-E	258	72	67
CphA-FPGS-E-F	243	84	70
D-T-α-ß-γ	136	159	208
CapB-CphA-FPGS-C-E-F-δ	136	159	208
CapB-CphA-FPGS-E-F-δ	188	324	404
C-D-T-α-ß-γ	188	324	404
C-D-T-α-ß-γ-δ	94	226	297
CapB-CphA-FPGS-E-F	94	226	297
C-D-E-F-T-α-ß-γ-δ	58	189	258
CapB-CphA-FPGS	58	189	258
CapB-FPGS-δ	143	204	270
T-α-ß	115	200	143

Overall, our analyses showed that neither MurT nor CphA should be considered as an outgroup for the Mur domain-containing family. Indeed, we observe that MurT sequences form either one or two (MurT + MurT-like) clans, which emerge from within the larger clan formed by the six muramyl ligases MurαßγδCD. In spite of the difficulty to determine the exact positions of MurT and MurT-like, topology and jackknife support tend to indicate that MurT sequences derive from the same ancestral gene as MurαßγδCD. In contrast to the other members of the Mur domain-containing family, CphA originates from the fusion of two functional domains: 1) an ATP-grasp domain at the N-terminal region (see ATP-grasp superfamily) and 2) the Mur ligase domain at the C-terminal region. This C-terminal region appears to be closely related to MurE and MurF in our phylogenetic inferences. Regarding CapB, species resampling showed that it is not related to the four bacterial muramyl ligases MurCDEF nor to the four archaeal muramyl ligases Murαßγδ, but more likely to FPGS (Table 2), thus indicating that it can be used as an outgroup to study the relationships between MurCDEF and Murαßγδ. However, unlike FPGS, CapB distribution is more restricted, the gene being found only in Gammaproteobacteria, Bacilli, Synergistetes, Halobacteria and a few Methanosarcinales and Korarchaota, according to our taxonomic analyses.

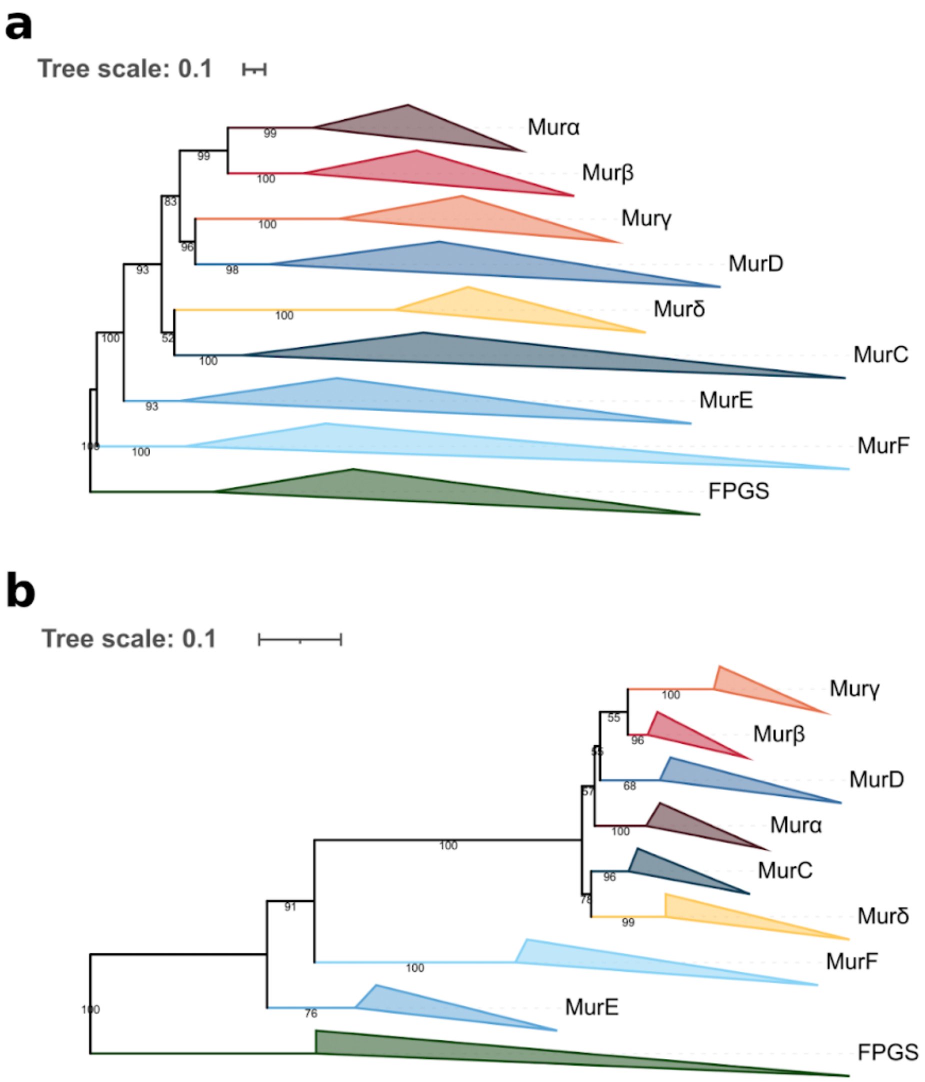


Figure 5. **Phylogenetic trees of the Mur domain-containing family rooted on FPGS.** (**a**) The tree was inferred from a matrix of 3,046 sequences x 543 unambiguously aligned AAs using IQ-TREE under the C40+G4 model. (**b**) “Indels” tree inferred from a matrix of 2997 sequences x 2243 unambiguously aligned AAs

using RAxML under the BINGAMMAX model. Tree visualization was performed using iTOL. Bootstrap support values are shown if greater or equal to 50. Branches were collapsed on sequence annotation.

3D models of the archaeal Mur ligases

As the four archaeal muramyl ligases do not have straightforward orthology relationships with their four bacterial counterparts, phylogeny alone cannot help determining the origin of those enzymes. However, the 3D structure of proteins can be used as a complement to unravel the evolution of muramyl ligases (Chang et al. 2004; Illergård et al. 2009). The structures of “Murα” (PDB code 6VR7) and “Murδ” (PDB codes 6VR8 and 7JT8) from *Methanothermus fervidus* are available in the Protein Data Bank (PDB). This data was complemented by the 3D models of the Murαßγδ ligases from *M. fervidus* and *Methanothermus smithii* obtained with the AlphaFold software (Jumper et al. 2021). Importantly, the Murα and Murδ models were obtained using a version of the PDB reference database predating the release of the corresponding structures to assess the accuracy of AlphaFold on this type of protein. The overall quality of all the models is very good, with average pLDDT (predicted local-distance difference test) values of the best model superior to 90% and only a few loops with significantly lower pLDDT values (Fig. S35). For Murα from *M. fervidus*, the rms (root-mean-square) deviation between the crystallographic structure and the AlphaFold model calculated for the Cα is 2.1Å, while it is below 0.7Å when calculated separately for each of the three domains. For Murδ, these values are 2.43Å and below 1.0Å, respectively. This shows that the AlphaFold models are of very high accuracy for the individual domains but with some slight movements observed between the domains.

As the nature of the AAs transferred to the pseudomurein precursors depends on the structural features of the C-terminal domain of the various Mur ligases, the 3D structures of Murαßγδ were compared with those of MurCDEF to identify their respective role in PM biosynthesis. For Murδ, a clear homology was observed with the structure of the C-terminal domain of MurC (Mol Clifford D. et al. 2003), which adds L-Ala to MurNAc in Bacteria (Fig. 6a). The residues surrounding the L-Ala moiety are either strictly conserved (H198, R377, A459, H348 in MurC from

Haemophilus influenzae) or substituted by an identical AA from a different structural element (R380 in *H. influenzae*) or substituted by residues with similar properties (H376 by a glutamine and Y346 by a phenylalanine). The AA added by Murδ to the archaeal PM peptide will therefore likely be an L-Ala as well, further strengthening the phylogenetic link identified between Murδ and MurC. However, as recently reported, the N-terminal domain of Murδ is more closely related to the corresponding MurE domain (both the primary and secondary structures) than to the MurC domain (Subedi et al. 2022).

A second significant match was observed between the structure of C-terminal domains of Murγ and MurD (Bertrand et al. 1999), which is responsible for the addition of D-Glu in Bacteria (Fig. 6b). The conservation is less strict in this case (only I416 of MurD from *E. coli* is conserved in Murγ), but the functionality of other AAs surrounding the D-Glu substrate is maintained. S415 and F422, which stabilize the γ-carboxylic acid through their backbone nitrogen and serine hydroxyl, are replaced by the backbone nitrogen of a glycine and a subsequent glutamine. In PM, the only AA with a carboxylic group away from the reaction center is the L-Glu added at the fifth position through its γ-carboxylic acid. This reaction must however involve a significant modification in the vicinity of the reaction center, as the functional groups of the stem peptide and AA added are inverted (bond between the γ carboxylic acid of L-Glu and ε amine of L-Lys at the third position of the peptide). In this context, it is therefore difficult to interpret the replacement of K348 and T321, which stabilize the α-carboxylic acid in MurD, by an arginine and a lysine, respectively, as well as the presence of an arginine and an aspartic acid (R312 and D289 in *M. fervidus*) close to the reaction center. While the ligation of L-Glu to the L-Lys in third position by Murγ is not fully validated by the comparison with MurD, it remains the most likely role of this enzyme.

For Murα and Murß, the comparison with the structure of the C-terminal domain of bacterial Mur enzymes did not reveal obvious similarities. However, in the Murα structure from *M. fervidus* and the model from *M. smithii*, two glutamic acids are conserved in the cavity usually accommodating the substrate, suggesting a role in the ligation of the L-Lys rather than the second L-Ala. This would leave Murß for the addition of the other L-Ala of the PM stem peptide, but it is difficult to verify because the two AlphaFold models of Murß analyzed are not congruent in this region.

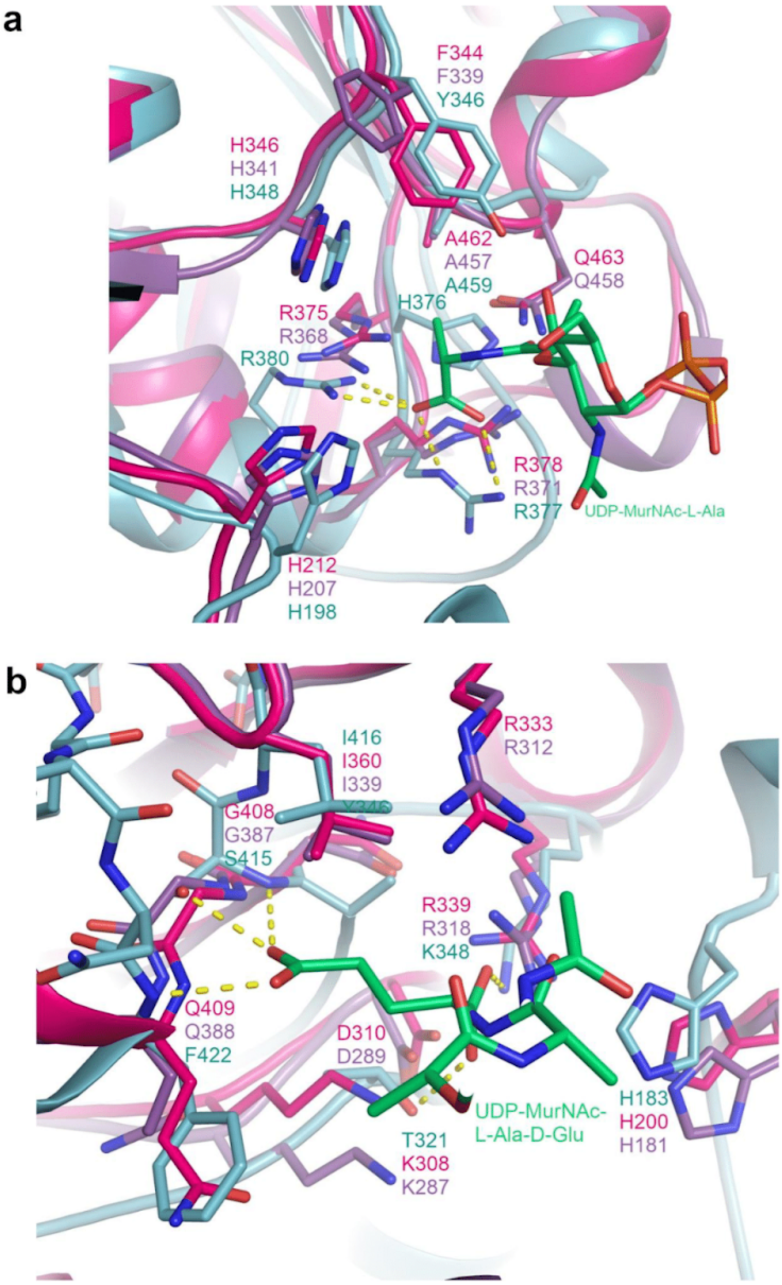


Figure 6. **Identification of the amino acid recognized by the C-terminal domain of Murδ and Murγ**. (**a**) Superimposition of the Murδ structure from *M. fervidus* (PDB

code 6VR8; purple) with the AlphaFold model of Murδ from *M. smithii* (pink) and the MurC structure from *Haemophilus influenzae* (PDB code 1P3D; light cyan) in complex with UDP-MurNAc-L-Ala (green). H-bonds between the L-Ala moiety and MurC are shown as yellow dashed lines. (**b**) Superimposition of the AlphaFold models of Murγ from *M. fervidus* (purple) and *M. smithii* (pink) and the MurD structure from *E. coli* (PDB code 4UAG; light cyan) in complex with UDP-MurNAc-L-Ala-D-Glu (green). H-bonds between the D-Glu moiety and MurD are shown as yellow dashed lines.

Discussion

Our phylogenetic analyses of the Mur domain-containing family show that each member of the Mur family is monophyletic. However, the relationships between those members are hard to establish owing to the low phylogenetic signal within the family and because phylogenetic artifacts, such as LBA (Gouy et al. 2015), probably affect phylogenetic reconstruction, especially for the trees including all non-Mur “outgroups”. Indeed, compared to MurCDEF, archaeal muramyl ligases (here termed Murαßγδ) are characterized by very long branches, and particularly Murδ, which has experienced more than one substitution per site since its probable separation from MurC. When focussing on Mur trees with only one outgroup, the topology is quite robust to different evolutionary models and species resampling within each member of the Mur domain-containing family. In this topology, MurD forms a clan with Murα+Murß+Murγ, MurC a clan with Murδ, and MurE a clan with MurF, a result that is also compatible with unrooted trees devoid of any outgroup (Fig S36 to S38). Moreover, structural analyses of the C-terminal domain of the four archaeal muramyl ligases allowed us to assign them a putative function in PM biosynthesis (Fig. 7). Indeed, due to some similarities between MurC and Murδ and between MurD and Murγ, we assume that Murδ adds one of the two L-Ala and Murγ adds L-Glu to the stem peptide. Although there are no obvious similarities between Murα and Murß and bacterial muramyl ligases, some clues suggest that Murα is responsible for the addition of L-Lys. Therefore, the second L-Ala of the stem peptide is probably added by Murß.

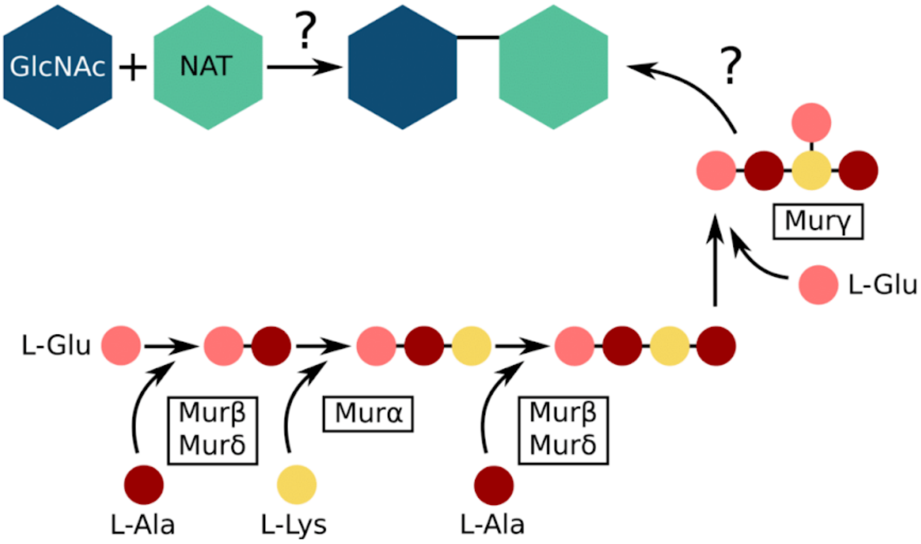


Figure 7. **Putative functions proposed for the four archaeal muramyl ligases (Murαßγδ) based on 3D structure comparisons.** The pathway presented here is the scenario proposed by Evamarie Hartmann, Helmut König and Uwe Kärcher (Hartmann and König 1990; König et al. 1993; Hartmann and König 1994). Although a specific function has been attributed to each archaeal muramyl ligase, we could not determine which one between Murβ and Murδ adds the L-Ala in position 2 and the L-Ala in position 4 of the stem peptide.

As previously stated, early analyses of their biosynthetic pathways have suggested that neither PG nor PM were a feature of LUCA (Scheffers and Pinho 2005; Albers and Meyer 2011; Subedi et al. 2021; Ithurbide et al. 2022). Therefore, LUCA probably did not possess the various muramyl ligases presently involved in cell-wall biosynthesis. However, FPGS is found in the three domains of life (Levin et al. 2004; Gorelova et al. 2019; Kordus and Baughn 2019), indicating that the gene was already part of the genome of LUCA. Thus, muramyl ligases emerged in Bacteria from a duplication of an ancestral version of FPGS and then were transferred to the other domain. In our phylogenetic trees, archaeal muramyl ligases (Murαßγδ) never branch within bacterial muramyl ligases (MurCDEF), and those trees do not give clues about the direction of the transfers. However, this topology could also be an artifact due to fast-evolving sequences in archaeal species. This kind of artifact has already been reported, e.g., with plastidial genes in eukaryotes, which rarely branch within (and rather sister to) Cyanobacteria (Sato 2021), although the endosymbiotic origin of the plastid is widely accepted (Ponce-Toledo et al. 2019). Because the LBCA already possessed a complete *dcw* gene cluster (Léonard et al. 2022), and given that PM is restricted to Methanopyrales and Methanobacteriales (Meyer and Albers 2020), we propose a scenario for the evolution of archaeal muramyl ligases through HGT (Fig. 8).

In this scenario, the ancestral gene of *murCDEF* was duplicated a first time in the pre-LBCA lineage to yield the ancestral genes of *murCD* and *murEF*, followed by a second round of duplications, which led to the current four bacterial muramyl ligases. Some evidence indicates that the duplication of the *murEF* ancestral gene to yield *murE* and *murF* could have occurred later than the duplication of the *murCD* ancestral gene. In fact, *murE* and *murF* genes are always in tandem in the *dcw* cluster of most bacterial species, as well as in the reconstruction of the LBCA *dcw* cluster (Léonard et al. 2022), and can even be expressed as a single fusion protein MurE-MurF (Laddomada et al. 2019). Moreover, in the majority of our Mur domain-containing family trees, MurE and MurF have slightly shorter branches than those of MurC and MurD. Early after the diversification of the LBCA, the *murD* gene was transferred to the common ancestor of Methanopyrales and Methanobacteriales, then *murD* experienced two duplications that yielded murα, murß and murγ (our nomenclature). In addition, *M*urα, Murß and Murγ exhibit a 3D fold similar to MurC/MurD for each of its three domains except for the presence of insertions in some loops (Fig. S39). In contrast, there is strong evidence that *murα* and *murß* arose from a gene duplication. These two muramyl ligases group together in almost all phylogenetic reconstructions (in both rooted and unrooted trees) and, as for MurF and MurE, their genes are in tandem in the genome of the majority of PM-containing archaea. Moreover, some Methanobrevibacter and Methanothermobacter genomes (two genera of Methanobacteriales) code for a Murα-Murß fusion protein (Subedi et al. 2021). As for the first, older, duplication of *murD*, leading to *murγ* and the *murαß* ancestor, it is visible in unrooted trees, where *murα*, *murß* and *murγ* form a clan.

However, the origin of the *murδ* gene remains unclear: while most of the phylogenetic trees and conserved residues in the C-terminal domain associate Murδ with MurC, the 3D structure of the N-terminal domain suggests that Murδ is rather related to MurE (Subedi et al. 2022). This inconsistency between phylogeny and structure can be due to different phenomena that are still to be untangled. First, the phylogenetic models struggle to exactly position the Murδ clan, probably due to its long basal branch. In most of the cases, Murδ forms a clan with MurC, while two phylogenetic trees using a C40 model (Fig S28 and S29) show Murδ emerging from within the MurE clan. Second, one cannot exclude evolutionary convergence, where a *murC* gene was first transferred and then its 3D structure gradually shifted to a MurE-like fold, or conversely, a *murE* gene was transferred and its key AAs converged to a MurC-like sequence. Finally, a more complex scenario would be the transfer of both *murC* and *murE* genes, followed by their recombination at the domain level, leading to the current Murδ.

Species resampling allowed us to complete this scenario with the three remaining proteins from the Mur domain-containing family: MurT, CphA and CapB. Hence, our analyses showed that MurT is clearly related to the clan formed by MurαßγDδC, CphA related to the MurEF clan, while CapB appears close to the outgroup, FPGS. In contrast to FPGS and MurCDEF, which are ubiquitous in Bacteria, MurT, CphA and CapB have a patchy distribution. Thus, they have probably arisen in a specific lineage, followed by HGT, instead of being a feature of the LBCA. In such a context, we assume that MurT could be derived from MurC or MurD, while CphA would originate from the fusion of an ATP-grasp containing gene, similar to the Glutathione biosynthesis GshAB, and a MurE or MurF gene. Regarding CapB, its origin is less clear but, like FPGS, CapB uses L-Glu as a substrate (Hsueh et al. 2017; Gorelova et al. 2019). Therefore, CapB could have been recruited from a duplicated FPGS gene, which suggests that it was indeed a suitable outgroup to study the relationships among the eight muramyl ligases.

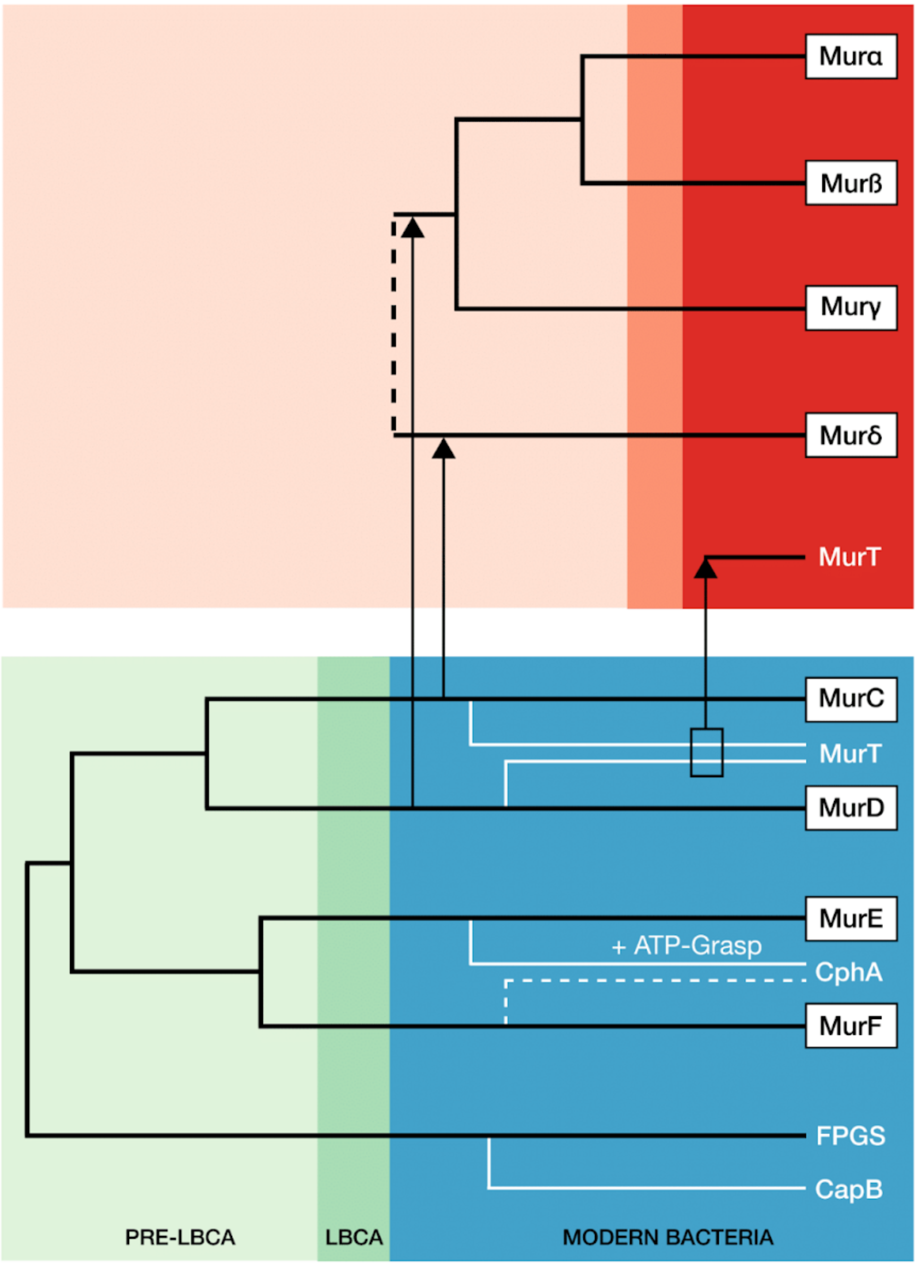


Figure 8. **Proposed scenario for the duplication events and horizontal gene transfers from Bacteria to Archaea having led to the extant organization of the Mur domain-containing family.** In this figure, only one possible origin is

represented for Murδ, the hypothesis where it stems from MurC.

Further insight about the transfers between Bacteria and PM-containing archaea can be obtained from the phylogeny of MurT. Previously, MurT has been described in *Staphylococcus* spp*, Streptococcus pneumoniae* and *Mycobacterium tuberculosis* (Münch et al. 2012; Morlot et al. 2018; Nöldeke et al. 2018; Maitra et al. 2021). Our analyses revealed that MurT is not ubiquitous in Bacteria, being only found in Firmicutes, Actinobacteria, *Caldisericum exile* (Caldiserica) and *Thermobaculum terrenum* (Chloroflexi). We also identified homologues in Archaea, specifically in Methanopyrales and Methanobacteriales. Interestingly, almost all bacteria have one copy of the *murT* gene while some PM-containing archaea have two copies, which we named *murT* and *murT-like*. Surprisingly, Methanobacteriales can possess only MurT or only MurT-like or both, while the few available Methanopyrales solely have one MurT-like gene. Moreover, archaeal MurT sequences are monophyletic and emerge from within Firmicutes (as sometimes observed for Murδ; Fig S28 and S29), while bacterial MurT sequences are consequently paraphyletic. Regarding MurT-like, the clan is monophyletic and basal to the MurT clan. In genomes of *Staphylococcus* species, *murT* and *gatD* genes are clustered in an operon (Münch et al. 2012; Morlot et al. 2018). Methanobacteriales and bacterial species that harbor a MurT homolog also have a GatD homolog while no GatD homologs are found in archaeal species bearing only MurT-like. This pattern suggests that MurT and GatD genes were transferred together to Methanobacteriales from a Terrabacteria lineage, probably Firmicutes. According to the taxonomic distribution of archaeal MurT/GatD and Murαßγδ, we can assume that the gene transfers of *murT*/*gatD* and the two ancestor genes of murαßγδ both occurred before the diversification of PM-containing archaea. In contrast, the origin of the *murT-like* gene is enigmatic, even though one possible explanation would be a duplication of *murT* in the LCA of Methanopyrales and Methanobacteriales, followed by differential loss of either *murT*/*gatD* or murT-like in some recent lineages.

In any case, those scenarios assume that the LBCA is older than the LCA of Methanopyrales and Methanobacteriales. However, molecular dating of prokaryotes is challenging since there are only a few microbial fossils or traces for which a meaningful taxonomy was proposed. The oldest evidence for microbial life has been identified in the Nuvvuagittuq belt in Quebec, Canada, which is between 3.75 and 4.28 billion years old (Gy) (Dodd et al. 2017; Papineau et al. 2022). There are also Archean rocks from up to 3.5 Gy containing chemical traces of microbial methanogenesis and sulfate reduction (Shen et al. 2001; Ueno et al. 2006; Aoyama and Ueno 2018; Catling and Zahnle 2020; Mißbach et al. 2021), thereby indicating that methanogenesis could be one of the most ancient biochemical pathways. Moreover, methanogenesis is a metabolism specific to the archaeal lineage (Gribaldo et al. 2006; Sorokin et al. 2017; Spang and Ettema 2017; Drake and Reiners 2021). Regarding bacterial microfossils, only three are unambiguously identified, all affiliated with the cyanobacterial lineage, of which *Eoentophysalis*, the oldest one, has been described from 1.9 Gy stromatolites (Hofmann 1976). Most scientists agree on the idea that the Great Oxidation Event (GOE) that occurred 2.4 Gy ago was due to the rise of oxygenic photosynthesis by Cyanobacteria. Using the GOE and the cyanobacterial fossil record as constraints for molecular clocks, it has been estimated that the cyanobacterial lineage appeared slightly before the GOE, as reviewed in Demoulin et al. 2019. Two recent molecular clock studies used horizontal gene transfers between archaeal methanogens and the LCA of Cyanobacteria, along with the cyanobacterial fossil record and the GOE, to date the origin of euryarchaeotal methanogens. They estimate the divergence between Euryarchaeota and the TACK group to have occurred around 4.1 and 3.8 Gy ago. Within Euryarchaeota, the LCA of class I methanogens (CIM) and class II methanogens (CIIM; Bapteste et al. 2005) (i.e., Methanomicrobiales and Methanosarcinales) originated 3.66 Gy ago (Gribaldo et al. 2006; Wolfe and Fournier 2018). As we hypothesize above, murT/gatD and murαßγδ genes were transferred from one or more bacterial lineages to the ancestor of Methanopyrales and Methanobacteriales. In present-day microbial communities, methanogens and sulfate-reducing bacteria (e.g., Deltaproteobacteria or Firmicutes) share the same ecological niche and can live in syntrophy under certain conditions (Lin et al. 2006; Muyzer and Stams 2008; Ozuolmez et al. 2015; Zouch et al. 2017). Evidence of such associations between sulfur-reducing and methanogens organisms were also identified in geological fluid inclusions from 3.5 Gy ago (Mißbach et al. 2021). Therefore, gene transfers between sulfur-reducing bacteria and methanogenic archaea could have occurred in that kind of environment and led to the origin of PM-containing archaea.

In one of the two putative gene clusters for PM biosynthesis, there is an ATP-grasp domain-containing gene that is always located upstream of the *mraY-like* and *murδ* genes. Moreover, this ATP-grasp domain-containing gene is exclusive to Methanopyrales and Methanobacteriales species. Thus, it has been proposed that it is probably involved in PM biosynthesis (Subedi et al. 2021). However, this gene was previously experimentally characterized by Popa et al. 2012, who concluded that it was a small (actually the smallest) carbamoyl phosphate synthetase (CPS) closely related to the "true" CPS, CarB. Given its putative function in cell-wall biosynthesis and its restricted taxonomic distribution, we hypothesized that this small CPS was not related to CarB but to Ddl instead and, as the muramyl ligases, had been transferred from a bacterial lineage to PM-containing archaea. However, our extensive phylogenetic analyses of the ATP-grasp superfamily remained inconclusive about the origin of the small CPS. Indeed, in our seven trees, it never clusters with CarB, nor with Ddl (except in the C20 tree). Moreover, the whole group is supported by a long branch, which can explain the difficulty to position the small CPS (i.e., LBA artifact). Although our phylogeny of the small CPS is inconclusive, its genetic environment suggests that it is indeed involved in PM biosynthesis. Accordingly, we postulate that the reported CPS function of this enzyme might be non-specific. If so, its real function in PM biosynthesis still has to be experimentally determined.

Located right downstream of the ATP-grasp domain-containing gene, the *mraY-like* (OG0001163) gene codes for a transmembrane protein that shows homology with the bacterial MraY. However, this archaeal MraY-like does not appear to have evolved from the bacterial MraY (i.e., through HGT). Indeed, bacterial and archeal proteins are clearly separated in all unrooted phylogenetic trees, although archaeal monophyletic groups are characterized by long branches, especially OG0001207, which could lead to strong phylogenetic artifacts (i.e, LBA). In contrast to Mur domain-containing family and ATP-grasp superfamily trees, MraY-like family trees were left unrooted. In fact, none of the MraY-like family members is found in both Bacteria and Archaea. As shown in the Results section, MraY and WecA/WbpL are only present in bacterial species, GTP is ubiquitous to archaea while MraY-like and OG0001207 are exclusive to PM-containing archaea. In addition, WecA/WbpL is the only monophyletic group where some organisms bear two sequences, which could indicate that WecA and WbpL are two paralogues. The position of the monophyletic group composed of the four bacterial unannotated sequences revealed that they could be divergent MraY sequences. According to the taxonomic distribution of MraY, WecA/WbpL and GPT, we propose a scenario where an ancestral GT4 gene found in LUCA was vertically transmitted to both Archaea (GPT) and Bacteria (the ancestral gene of MraY and WecA/WbpL). The bacterial gene was then duplicated once to yield *mraY* and *wecA/WbpL*, and the latter experienced a second duplication in some bacterial species. Thus, GPT would be the orthologue of MraY and WecA/WbpL, while MraY and WecA/WbpL would be paralogous. For this phylogenetic analysis of the MraY-like family, we followed the family as defined in the NCBI CDD (https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=264002) to increase the sequence sampling. In theory, it is possible that we undersampled the family. Indeed, the GT4 domain is also present in other proteins, like MurG (Mengin-Lecreulx et al. 1991; Laddomada et al. 2019), which are not part of the MraY-like family. A proper way to study the origin of the MraY-like family would be to infer a phylogenetic tree of the GT4 domain. However, such an analysis would be very time-consuming due to the large number of GT sequences (Lombard et al. 2014). For now, overlapping HMM search results starting from the different family members do not suggest any undersampling issue. Furthermore, although bacterial homologues of OG0001207 have a MraY/WecA-like GT4 domain, the long branch of the monophyletic group could indicate that OG0001207 and homologues are probably not part of the MraY-like family.

The current architecture of PG and PM are well-known, but it is clear that both polymers were different in their early evolutionary state, i.e., before acquisition and diversification of their respective muramyl ligases. However, inferring the ancestral states of PG and PM is almost impossible because those evolved in the stem branch of Bacteria or CIM Archaea, respectively, before the LCAs of extant organisms. As other Mur-ligase family proteins, like CapB, FPGS, MurT or CphA, bind AAs with an α-carboxylic acid group (i.e., aspartic acid and glutamic acid), we can speculate that the first muramyl ligase proteins were also associated with those AAs. Moreover, glutamic acid is one of the most abundant AAs in many organisms, and it participates in a wide array of metabolisms (Walker and van der Donk 2016).

Therefore, glutamic acid could be one of the first AAs to have been selected by muramyl ligases. In *Bacillus*, the complex formed by CapB, CapC, CapA and CapE recruits L-Glu or D-Glu to synthesize the poly-γ-glutamic acid capsule. This kind of cell wall has been suggested to occur in *Haloquadratum walsbyi*, based on genomic analyses. *H. walsbyi* is classified in Halobacteria, a class of Euryarchaeota characterized by a diverse variety of cell walls: S-layer, sulfated heteropolysaccharides, halomucin and a glutaminylglycan. The latter is composed of poly-γ-L-glutamate, to which are linked two types of oligosaccharides (Meyer and Albers 2020). Analyses showed that CapB is ubiquitous in Halobacteria, indicating that CapB could be involved in glutaminylglycan biosynthesis. Consequently, we suggest that this simpler cell wall could resemble the ancient forms of PG and/or PM.

Material and Methods

Data availability

Publicly available datasets, including all detailed YAML configuration files used with Forty-Two (Irisarri et al. 2017; Simion et al. 2017) and classify-ali.pl (D. Baurain; https://metacpan.org/dist/Bio-MUST-Core), and a detailed command line log file can be found here: https://doi.org/10.6084/m9.figshare.21641612.

Protein sequence databases

Three local mirrors of NCBI RefSeq were used during this study: 1) an archaeal database composed of the 819 whole genomes that were available on March 7, 2019, 2) a bacterial database of 598 representative genomes selected by the ToRQuEMaDA pipeline (Léonard et al. 2021) and 3) a prokaryotic database of 80,490 genomes, already used in (Lupo et al. 2022). To assemble the bacterial database, ToRQuEMaDA was run in June 2018, according to a 'direct' strategy and using the following parameters: dist-metric set to JI (Jaccard Index), dist-threshold set to 0.86, clustering-mode set to 'loose', and pack size set to 200.

Identification of candidate proteins for pseudomurein biosynthesis

Protein orthologous groups (OGs) were built from the conceptual translations of ten archaeal whole genomes using OrthoFinder v2.2.1 (Emms and Kelly 2015) with default parameters. These archaeal genomes correspond to five organisms having pseudomurein (PM) (GCF_000008645.1, GCF_000016525.1, GCF_000166095.1, GCF_002201915.1, GCF_900095295.1) and five without PM (GCF_000011185.1, GCF_000013445.1, GCF_000017165.1, GCF_000025285.1, GCF_000251105.1) and were downloaded from the NCBI RefSeq database on March 7th, 2019. Then, taxonomic filters were applied to the OGs using classify-ali.pl v0.212670 in order to select candidate proteins for PM biosynthesis. Hence, we first looked for OGs with protein sequences from all five PM-containing archaea or from one Methanopyrales and three Methanobacteriales or from four Methanobacteriales. To identify OGs corresponding to a widespread gene that would also include a paralogue potentially specific to PM-containing archaea, we used the same taxonomic criteria but set the ‘min_copy_mean’ option to 1.75 for PM-containing archaea and to 1.25 for other species (see YAML configuration files for details). In addition, three HMM profiles from NCBI CDD (Conserved Domain Database) (Lu et al. 2020) featuring ‘pseudomurein’ in their annotation were downloaded on December 18th, 2020. Then the profiles were used to identify homologues in the conceptual translations of the five PM-containing archaea with hmmsearch from the HMMER package v3.3 (Mistry et al. 2013) with default parameters. Matching protein sequences were graphically selected using the Ompa-Pa v0.211430 interactive software package (A. Bertrand and D. Baurain; https://metacpan.org/dist/Bio-MUST-Apps-OmpaPa) with the ‘max_copy’ option set to 20 and ‘min_cov’ to 0.7. Finally, the corresponding OGs were added to the selection.

Genetic environment analysis of candidate proteins and *in-silico* characterization of their domains

Genetic environment databases were built for the genes of the selected OGs using the “3 in 1” module of GeneSpy (Garcia et al. 2019). Functional domains were predicted using InterProScan v5.37-76.0 (Jones et al. 2014), along with SignalP v5.0b (Almagro Armenteros et al. 2019) and TMHMM v2.0c (Krogh et al. 2001). InterProScan was used with default parameters and we disabled the precalculated match lookup, while the SignalP organism option was set to 'arch'. To avoid misprediction by TMHMM, the signal peptide was first removed from the original sequences when the cleavage site prediction probability was greater than or equal to 0.1.

Filtering of candidate proteins

To rescue potential pseudogenes or mistranslated proteins missing in selected OGs with protein sequences from only four (out of five) PM-containing archaea, Forty-Two v0.213470 was run in TBLASTN mode on the whole genomic sequences of the five PM-containing archaea. Then, classify-ali.pl was used again to retain only the OGs having sequences from all five PM-containing archaea. To enrich OGs with further archaeal orthologues, a second round of forty-two.pl in BLASTP mode was performed using the archaeal database of 819 whole genomes (see YAML configuration files for details). Each enriched OG was aligned using MAFFT L-INS-i v7.273 (Katoh and Standley 2013). From those alignments, HMM profiles were built using the HMMER package and bacterial homologues were identified separately in the bacterial and the prokaryotic databases. Protein sequences were graphically selected using Ompa-Pa with 'max_copy' and 'min_cov' options set to 20 and 0.7, respectively. For each OG, identical length and e-value thresholds were used for both databases when selecting homologous proteins.

Phylogenetic analyses

ATP-grasp superfamily

In order to select a set of representative sequences containing the ATP-grasp domain, we first built a HMM profile from the alignment of the OG containing archaeal ATP-grasp domain proteins using the HMMER package. This profile was uploaded to the HMMER website (https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch) from which we retrieved homologous sequences (excluding eukaryotes) from the Swiss-Prot database (Poux et al. 2017). From those sequences, homologues were identified in our local bacterial databases using the HMM profile and Ompa-Pa. In parallel, the archaeal OGs (see Identification of candidate proteins for pseudomurein biosynthesis) homologous to the Swiss-Prot proteins were identified using NCBI BLASTp v2.2.28+ (Camacho et al. 2009) and enriched using Forty-Two with the archaeal database as 'bank'. Finally, all archaeal and bacterial homologous sequences were merged into one single file.

To identify most ATP-grasp-containing domain proteins in our local databases, the merged file was aligned using MAFFT L-INS-i and the alignment was masked using the mask-ali.pl perl script (D. Baurain; https://metacpan.org/dist/Bio-MUST-Core) to isolate the ATP-grasp domain. From this domain alignment, an HMM profile was built using the HMMER package to identify ATP-grasp domain-containing homologues in our archaeal and bacterial databases, and homologous sequences were selected using Ompa-Pa. Protein sequences with two ATP-grasp domains (i.e., CarB) were cut at half-length, then both complete and half-sequences were aligned using MAFFT and their ATP-grasp domain again isolated using mask-ali.pl. Protein sequences were deduplicated using cdhit-tax-filter.pl perl script (V. Lupo and D. Baurain; https://metacpan.org/dist/Bio-MUST-Drivers) with the 'keep-all' option enabled and the identity threshold set to 0.65, then tagged using a BLAST-based annotation script (part of Bio-MUST-Drivers) and highly divergent sequences were removed using prune-outliers.pl v0.213470 with the 'evalue' option set to 1e-3, 'min-hits' to 1, 'min_ident' to 0.01 and 'max_ident' to 0.2. Finally, sequences were realigned with MAFFT L-INS-i. Conserved sites were selected using ali2phylip.pl v0.212670 (D. Baurain; https://metacpan.org/dist/Bio-MUST-Core) with the 'min' and 'max' options set to 0.3. The resulting matrix of 2,194 sequences x 180 AAs was used to infer phylogenetic domain trees using IQ-TREE v1.6.12 (Nguyen et al. 2015) with 1000 ultrafast bootstrap (UFBoot) replicates (Hoang et al. 2018) and under four models: LG4X+R4, C20+G4, C40+G4 and PMSF LG+C60+G4. In total, seven trees were computed because we tested the effect of increasing the number of iterations from 1000 to 3000 for the C20 and C40 models, and from 3000 to 5000 for the PMSF model.

MraY-like family

The two OGs (see Identification of candidate proteins for pseudomurein biosynthesis) containing proteins predicted with a domain glycosyltransferase 4 were enriched in bacterial homologues using Forty-Two in BLASTP mode. In parallel, representative sequences from other members of the MraY-like family (https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=264002) were downloaded from the UniProtKB (The UniProt Consortium 2021) database: WecA (P0AC78, P0AC80, Q8Z38), GPT (P96000, B5IDH8) and WbpL (G3XD50, A0A379IBB8). The three files were then enriched in bacterial and archaeal (if any) homologues using Forty-Two. Finally, the five files were aligned using MAFFT L-INS-i.

To better explore the diversity of the MraY-like family, HMM profiles were built from those alignments and homologous sequences were selected from HMMER hits on the bacterial database using Ompa-Pa. All homologous protein sequences were merged into one file and tagged using a BLAST-based annotation script (part of Bio-MUST-Drivers) and aligned using MAFFT L-INS-i. Conserved sites were selected using ali2phylip.pl with the 'min' and 'max' options set to 0.2. A first guide tree was computed from the resulting matrix of 1070 sequences x 410 AAs using IQ-TREE with 1000 UFBoot under the LG4X+R4 model. From this guide tree and automated annotation, all sequences were manually tagged using 'treeplot' from the MUST software package (Philippe 1993). According to their annotation, protein sequences of each member of the MraY-like family were aligned using MAFFT L-INS-i, then all members were realigned using Two-Scalp v0.211710 (A. Bertrand, V. Lupo and D. Baurain; https://metacpan.org/dist/Bio-MUST-Apps-TwoScalp) with the 'linsi' option enabled. Finally, ali2phylip.pl was used to select conserved sites with the 'min' and 'max' options set to 0.2 and the resulting matrix of 1070 sequences x 408 AAs was used to infer phylogenetic trees with IQ-TREE under three models (i.e., LG4X+R4, C20+G4, C40+G4) and 1000 UFBoot.

Mur domain-containing family

After enrichment of the OGs with archaeal and bacterial homologues, the multiple OGs corresponding to the Mur domain-containing family were merged into one single (unaligned) file. In parallel, reference protein sequences from additional members of the Mur domain-containing family were downloaded into three separated files using the command-line version of the 'efetch' tool v10.4 from the NCBI Entrez Programming Utilities (E-utilities): CapB (P96736), MurT (Q8DNZ9, A0A0H3JUU7, A0A0H2WZQ7) and CphA (P56947, O86109, P58572). Forty-Two in BLASTP mode was run, in two rounds, on the four files, using both bacterial and archaeal databases as 'bank', in a final effort to sample the diversity of Mur domain-containing proteins. Then, fusion proteins were cut between the two protein domains and half-sequences with no Mur ligase domain were discarded. The enriched files were merged and protein sequences were deduplicated using the cdhit-tax-filter.pl with the 'keep-all' option enabled and the identity threshold set to 1. Mur domain-containing family proteins were tagged using a BLAST-based annotation script (part of Bio-MUST-Drivers) with an e-value threshold of 1e-20. Protein sequences were aligned using MAFFT (default mode) and conserved sites were selected using ali2phylip.pl with the 'max' option set to 0.3. A first guide tree was computed with IQ-TREE under the LG4X+R4 model with 1000 UFBoot. Based on the automatic annotation, all protein sequences were manually tagged following the guide tree using 'treeplot' from the MUST software package.

In order to improve phylogenetic analysis, the alignment of the Mur domain-containing family was refined as follows: 1) sequences from the different members of the family were exported to distinct files and aligned using MAFFT L-INS-i, 2) using the 'ed' programme from the MUST software package, misaligned sequences were manually transferred to a '.non' file, and then, reduced files were realigned using MAFFT L-INS-i, 3) realigned files and '.non' files were merged and all sequences were aligned using Two-Scalp with the 'linsi' and 'keep-length' options enabled. Conserved sites were selected using ali2phylip.pl with the 'max' and 'min' option set to 0.3. Phylogenetic analysis was performed on the resulting matrix of 3407 sequences x 550 AAs using IQ-TREE with 1000 UFBoot under three models of sequence evolution: LG4X+R4, C20+G4 and C40+G4.

From the alignment of the four bacterial muramyl ligases (MurCDEF), the four archaeal muramyl ligases (Murαßγδ) and the FGPS protein sequences, we have produced two more alignments: one where the N-ter and the C-ter domains of the protein sequences were trimmed, and another where we kept only the most conserved AAs. Fusion proteins were removed from those three alignments and protein sequences converted to a binary encoding to analyze indels (i.e., 0 for a gap or a missing character state and 1 for any AA). Short sequences were removed using ali2phylip.pl with the 'min' option set to 0.6. The three resulting matrices of 2997 sequences x 2243 AAs, 3001 sequences x 1799 AAs and 3004 sequences x 281 AAs, respectively, were used to infer phylogenetic trees with with RAxML v8.1.17 (Stamatakis 2014) under the BINGAMMAX model.

The jackknife.pl perl script (part of Bio-MUST-Drivers) was used for species resampling analysis with the 'linsi' option enabled, 'min' and 'max' set to 0.3 and 'n-process' to 1000. The one thousand resulting alignments were used to infer phylogenetic trees using IQ-TREE with 1000 UFBoot under the LG4X+R4, C20+G4 and C40+G4 models. Clan support values were assessed using the parse_consense_out.pl perl script (Baurain et al. 2010) with the 'mode' option set to 'tree'. Consensus trees were computed from the 1000 replicate trees using ASTRAL v5.7.7 (Zhang et al. 2018) with default options.

Acknowledgements

VL is supported by a FRIA (Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture) fellowship of the F.R.S.-FNRS. FK is a research associate of 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), and by the Consortium des Équipements de Calcul Intensif (CÉCI) founded by the F.R.S.-FNRS. The authors thank Rosa Gago for her help with the graphical design of the figures.

Authors' Contributions

VL conceived the study and designed experiments, performed experiments, analyzed the data, drafted and drew the figures, wrote the manuscript and approved the final manuscript. DB conceived the study and designed experiments, analyzed the data, wrote and reviewed the manuscript and approved the final manuscript. FK conceived the study and designed experiments, performed experiments, analyzed the data, drew the figures, wrote and reviewed the manuscript and approved the final manuscript. CR, ER, LO and OJ performed experiments and approved the final manuscript.

References

Aboulmagd E, Oppermann-Sanio FB, Steinbüchel A. 2001. Purification of Synechocystis sp. strain PCC6308 cyanophycin synthetase and its characterization with respect to substrate and primer specificity. *Appl. Environ. Microbiol.* 67:2176–2182.

Albers S-V, Meyer BH. 2011. The archaeal cell envelope. *Nat. Rev. Microbiol.* 9:414–426.

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37:420–423.

Amer AO, Valvano MA. 2001. Conserved amino acid residues found in a predicted cytosolic domain of the lipopolysaccharide biosynthetic protein WecA are implicated in the recognition of UDP-N-acetylglucosamine. *Microbiol. Read. Engl.* 147:3015–3025.

Anderssen S, Naômé A, Jadot C, Brans A, Tocquin P, Rigali S. 2022. AURTHO: Autoregulation of transcription factors as facilitator of cis-acting element discovery. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* 1865:194847.

Aouad M, Flandrois J-P, Jauffrit F, Gouy M, Gribaldo S, Brochier-Armanet C. 2022. A divide-and-conquer phylogenomic approach based on character supermatrices resolves early steps in the evolution of the Archaea. *BMC Ecol. Evol.* 22:1.

Aoyama S, Ueno Y. 2018. Multiple sulfur isotope constraints on microbial sulfate reduction below an Archean seafloor hydrothermal system. *Geobiology* 16:107–120.

Ashiuchi M. 2013. Microbial production and chemical transformation of poly-γ-glutamate. *Microb. Biotechnol.* 6:664–674.

Bapteste E, Brochier C, Boucher Y. 2005. Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. *Archaea Vanc. BC*

1:353–363.

Baurain D, Brinkmann H, Petersen J, Rodríguez-Ezpeleta N, Stechmann A, Demoulin V, Roger AJ, Burger G, Lang BF, Philippe H. 2010. Phylogenomic Evidence for Separate Acquisition of Plastids in Cryptophytes, Haptophytes, and Stramenopiles. *Mol. Biol. Evol.* 27:1698–1709.

Bertrand JA, Auger G, Martin L, Fanchon E, Blanot D, Le Beller D, van Heijenoort J, Dideberg O. 1999. Determination of the MurD mechanism through crystallographic analysis of enzyme complexes11Edited by R. Huber. *J. Mol. Biol.* 289:579–590.

Bhattacharjee MK. 2016. Antibiotics That Inhibit Cell Wall Synthesis. In: Bhattacharjee MK, editor. Chemistry of Antibiotics and Related Drugs. Cham: Springer International Publishing. p. 49–94. Available from: https://doi.org/10.1007/978-3-319-40746-3_3

Braun F, Recalde A, Bähre H, Seifert R, Albers S-V. 2021. Putative Nucleotide-Based Second Messengers in the Archaeal Model Organisms Haloferax volcanii and Sulfolobus acidocaldarius. *Front. Microbiol.* 12:779012.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Cammarano P, Gribaldo S, Johann A. 2002. Updating carbamoylphosphate synthase (CPS) phylogenies: occurrence and phylogenetic identity of archaeal CPS genes. *J. Mol. Evol.* 55:153–160.

Campbell JA, Davies GJ, Bulone V, Henrissat B. 1997. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem. J.* 326 (Pt 3):929–939.

Catling DC, Zahnle KJ. 2020. The Archean atmosphere. *Sci. Adv.* 6:eaax1420.

Chang AB, Lin R, Keith Studley W, Tran CV, Saier MHJ. 2004. Phylogeny as a guide to structure and function of membrane transport proteins. *Mol. Membr. Biol.* 21:171–181.

Chen H, Gan Q, Fan C. 2020. Methyl-Coenzyme M Reductase and Its Post-translational Modifications. *Front. Microbiol.* 11:578356.

Cheng YS, Shen Y, Rudolph J, Stern M, Stubbe J, Flannigan KA, Smith JM. 1990. Glycinamide ribonucleotide synthetase from Escherichia coli: cloning, overproduction, sequencing, isolation, and characterization. *Biochemistry* 29:218–227.

Cristianini N, Hahn MW. 2006. Introduction to Computational Genomics: A Case Studies Approach. Cambridge University Press Available from: https://books.google.be/books?id=t3lkngEACAAJ

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

Dal Nogare AR, Dan N, Lehrman MA. 1998. Conserved sequences in enzymes of the UDP-GlcNAc/MurNAc family are essential in hamster UDP-GlcNAc:dolichol-P GlcNAc-1-P transferase. *Glycobiology* 8:625–632.

Demoulin CF, Lara YJ, Cornet L, François C, Baurain D, Wilmotte A, Javaux EJ. 2019. Cyanobacteria evolution: Insight from the fossil record. *Early Life Earth Oxidative Stress* 140:206–223.

Diesterhaft MD, Freese E. 1973. Role of pyruvate carboxylase, phosphoenolpyruvate carboxykinase, and malic enzyme during growth and sporulation of Bacillus subtilis. *J. Biol. Chem.* 248:6062–6070.

Dodd MS, Papineau D, Grenne T, Slack JF, Rittner M, Pirajno F, O'Neil J, Little CTS. 2017. Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature* 543:60–64.

Drake H, Reiners PW. 2021. Thermochronologic perspectives on the deep-time evolution of the deep biosphere. *Proc. Natl. Acad. Sci. U. S. A.* 118.

Egan AJF, Errington J, Vollmer W. 2020. Regulation of peptidoglycan synthesis and remodelling. *Nat. Rev. Microbiol.* 18:446–460.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157.

Fawaz MV, Topper ME, Firestine SM. 2011. The ATP-grasp enzymes. *Bioorganic Chem.* 39:185–191.

Garcia PS, Jauffrit F, Grangeasse C, Brochier-Armanet C. 2019. GeneSpy, a user-friendly and flexible genomic context visualizer. *Bioinformatics* 35:329–331.

Gorelova V, Bastien O, De Clerck O, Lespinats S, Rébeillé F, Van Der Straeten D. 2019. Evolution of folate biosynthesis and metabolism across algae and land plant lineages. *Sci. Rep.* 9:5731.

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

Graham DE, Huse HK. 2008. Methanogens with pseudomurein use diaminopimelate aminotransferase in lysine biosynthesis. *FEBS Lett.* 582:1369–1374.

Gribaldo S, Brochier-Armanet C, Gribaldo S, Brochier-Armanet C. 2006. The origin and evolution of Archaea: a state of the art. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 361:1007–1022.

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

Hartmann E, König H. 1994. A novel pathway of peptide biosynthesis found in methanogenic Archaea. *Arch. Microbiol.* 162:430–432.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* 35:518–522.

Hofmann HJ. 1976. Precambrian Microflora, Belcher Islands, Canada: Significance and Systematics. *J. Paleontol.* 50:1040–1073.

Hou J, Xiang H, Han J. 2015. Propionyl coenzyme A (propionyl-CoA) carboxylase in Haloferax mediterranei: Indispensability for propionyl-CoA assimilation and impacts on global metabolism. *Appl. Environ. Microbiol.* 81:794–804.

Hsueh Y-H, Huang K-Y, Kunene SC, Lee T-Y. 2017. Poly-γ-glutamic Acid Synthesis, Gene Regulation, Phylogenetic Relationships, and Role in Fermentation. *Int. J. Mol. Sci.* 18:2644.

Illergård K, Ardell DH, Elofsson A. 2009. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins* 77:499–508.

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. *Nat. Ecol. Evol.* 1:1370–1378.

Ithurbide S, Gribaldo S, Albers S-V, Pende N. 2022. Spotlight on FtsZ-based cell division in Archaea. *Trends Microbiol.* 30:665–678.

Jeske O, Schüler M, Schumann P, Schneider A, Boedeker C, Jogler M, Bollschweiler D, Rohde M, Mayer C, Engelhardt H, et al. 2015. Planctomycetes do possess a peptidoglycan cell wall. *Nat. Commun.* 6:7116.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236–1240.

Joyce MA, Fraser ME, Brownie ER, James MN, Bridger WA, Wolodko WT. 1999.

Probing the nucleotide-binding site of Escherichia coli succinyl-CoA synthetase. *Biochemistry* 38:7273–7283.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589.

Kandler O, Konig H. 1993. Chapter 8 Cell envelopes of archaea: Structure and chemistry. In: Kates M, Kushner DJ, Matheson AT, editors. New Comprehensive Biochemistry. Vol. 26. Elsevier. p. 223–259. Available from: https://www.sciencedirect.com/science/article/pii/S0167730608602574

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.

König H, Hartmann E, Kärcher U. 1993. Pathways and Principles of the Biosynthesis of Methanobacterial Cell Wall Polymers. *Syst. Appl. Microbiol.* 16:510–517.

König H, Kralik R, Kandler O. 1982. Structure and Modifications of Pseudomurein in Methano-bacleriales. *Zentralblatt Für Bakteriol. Mikrobiol. Hyg. Abt Orig. C Allg. Angew. Ökol. Mikrobiol.* 3:179–191.

Kordus SL, Baughn AD. 2019. Revitalizing antifolates through understanding mechanisms that govern susceptibility and resistance. *MedChemComm* 10:880–895.

Kouidmi I, Levesque RC, Paradis-Bleau C. 2014. The biology of Mur ligases as an antibacterial target. *Mol. Microbiol.* 94:242–253.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305:567–580.

Laddomada F, Miyachiro MM, Jessop M, Patin D, Job V, Mengin-Lecreulx D, Le Roy A, Ebel C, Breyton C, Gutsche I, et al. 2019. The MurG glycosyltransferase provides an oligomeric scaffold for the cytoplasmic steps of peptidoglycan biosynthesis in the human pathogen Bordetella pertussis. *Sci. Rep.* 9:4656.

Lawson FS, Charlebois RL, Dillon JA. 1996. Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life. *Mol. Biol. Evol.* 13:970–977.

Leahy SC, Kelly WJ, Altermann E, Ronimus RS, Yeoman CJ, Pacheco DM, Li D, Kong Z, McTavish S, Sang C, et al. 2010. The genome sequence of the rumen methanogen Methanobrevibacter ruminantium reveals new possibilities for controlling ruminant methane emissions. *PloS One* 5:e8926.

Léonard RR, Leleu M, Vlierberghe MV, Cornet L, Kerff F, Baurain D. 2021. ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies. *PeerJ* 9:e11348.

Léonard RR, Sauvage E, Lupo V, Perrin A, Sirjacobs D, Charlier P, Kerff F, Baurain D. 2022. Was the Last Bacterial Common Ancestor a Monoderm after All? *Genes* 13:376.

Levin I, Giladi M, Altman-Price N, Ortenberg R, Mevarech M. 2004. An alternative pathway for reduced folate biosynthesis in bacteria and halophilic archaea. *Mol. Microbiol.* 54:1307–1318.

Liechti G, Kuru E, Packiam M, Hsu Y-P, Tekkam S, Hall E, Rittichier JT, VanNieuwenhze M, Brun YV, Maurelli AT. 2016. Pathogenic Chlamydia Lack a Classical Sacculus but Synthesize a Narrow, Mid-cell Peptidoglycan Ring, Regulated by MreB, for Cell Division. *PLoS Pathog.* 12:e1005590.

Liechti GW, Kuru E, Hall E, Kalinda A, Brun YV, VanNieuwenhze M, Maurelli AT. 2014. A new metabolic cell-wall labelling method reveals peptidoglycan in Chlamydia trachomatis. *Nature* 506:507–510.

Lin L-H, Wang P-L, Rumble D, Lippmann-Pipke J, Boice E, Pratt LM, Sherwood Lollar B, Brodie EL, Hazen TC, Andersen GL, et al. 2006. Long-term sustainability of a high-energy, low-diversity crustal biome. *Science* 314:479–482.

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42:D490-495.

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, et al. 2020. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 48:D265–D268.

Lupo V, Mercuri PS, Frère J-M, Joris B, Galleni M, Baurain D, Kerff F. 2022. An Extended Reservoir of Class-D Beta-Lactamases in Non-Clinical Bacterial Strains. *Microbiol. Spectr.* 10:e0031522.

Lupo V, Van Vlierberghe M, Vanderschuren H, Kerff F, Baurain D, Cornet L. 2021. Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics. *Front. Microbiol.* 12:755101.

Maitra A, Nukala S, Dickman R, Martin LT, Munshi T, Gupta A, Shepherd AJ, Arnvig KB, Tabor AB, Keep NH, et al. 2021. Characterization of the MurT/GatD complex in Mycobacterium tuberculosis towards validating a novel anti-tubercular drug target. *JAC-Antimicrob. Resist.* 3:dlab028.

Makino S, Uchida I, Terakado N, Sasakawa C, Yoshikawa M. 1989. Molecular characterization and protein analysis of the cap region, which is essential for encapsulation in Bacillus anthracis. *J. Bacteriol.* 171:722–730.

Marolewski AE, Mattia KM, Warren MS, Benkovic SJ. 1997. Formyl phosphate: a proposed intermediate in the reaction catalyzed by Escherichia coli PurT GAR transformylase. *Biochemistry* 36:6709–6716.

Martínez-Torró C, Torres-Puig S, Marcos-Silva M, Huguet-Ramón M, Muñoz-Navarro C, Lluch-Senar M, Serrano L, Querol E, Piñol J, Pich OQ. 2021. Functional Characterization of the Cell Division Gene Cluster of the Wall-less Bacterium Mycoplasma genitalium. *Front. Microbiol.* 12:695572.

Mengin-Lecreulx D, Texier L, Rousseau M, van Heijenoort J. 1991. The murG gene of Escherichia coli codes for the UDP-N-acetylglucosamine: N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase involved in the membrane steps of peptidoglycan synthesis. *J. Bacteriol.* 173:4625–4636.

Meyer BH, Albers S-V. 2020. Archaeal Cell Walls. In: eLS. p. 1–14. Available from: https://doi.org/10.1002/9780470015902.a0000384.pub3

Mingorance J, Tamames J. 2004. The bacterial dcw gene cluster: an island in the genome? In: Vicente M, Tamames J, Valencia A, Mingorance J, editors. Molecules in Time and Space: Bacterial Shape, Division and Phylogeny. Boston, MA: Springer US. p. 249–271. Available from: https://doi.org/10.1007/0-306-48579-6_13

Mißbach H, Duda J-P, van den Kerkhof AM, Lüders V, Pack A, Reitner J, Thiel V. 2021. Ingredients for microbial life preserved in 3.5 billion-year-old fluid inclusions. *Nat. Commun.* 12:1101.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41:e121.

Mol Clifford D., Brooun Alexei, Dougan Douglas R., Hilgers Mark T., Tari Leslie W., Wijnands Robert A., Knuth Mark W., McRee Duncan E., Swanson Ronald V.

2003. Crystal Structures of Active Fully Assembled Substrate- and Product-Bound Complexes of UDP-N-Acetylmuramic Acid:l-Alanine Ligase (MurC) from Haemophilus influenzae. *J. Bacteriol.* 185:4152–4162.

Morlot C, Straume D, Peters K, Hegnar OA, Simon N, Villard A-M, Contreras-Martel C, Leisico F, Breukink E, Gravier-Pelletier C, et al. 2018. Structure of the essential peptidoglycan amidotransferase MurT/GatD complex from Streptococcus pneumoniae. *Nat. Commun.* 9:3180.

Mueller EJ, Meyer E, Rudolph J, Davisson VJ, Stubbe J. 1994. N5-carboxyaminoimidazole ribonucleotide: evidence for a new intermediate and two new enzymatic activities in the de novo purine biosynthetic pathway of Escherichia coli. *Biochemistry* 33:2269–2278.

Münch D, Roemer T, Lee SH, Engeser M, Sahl HG, Schneider T. 2012. Identification and in vitro analysis of the GatD/MurT enzyme-complex catalyzing lipid II amidation in Staphylococcus aureus. *PLoS Pathog.* 8:e1002509.

Musfeldt M, Schönheit P. 2002. Novel Type of ADP-Forming Acetyl Coenzyme A Synthetase in Hyperthermophilic *Archaea*: Heterologous Expression and Characterization of Isoenzymes from the Sulfate Reducer *Archaeoglobus fulgidus* and the Methanogen *Methanococcus jannaschii*. *J. Bacteriol.* 184:636–644.

Muyzer G, Stams AJM. 2008. The ecology and biotechnology of sulphate-reducing bacteria. *Nat. Rev. Microbiol.* 6:441–454.

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

Nöldeke ER, Muckenfuss LM, Niemann V, Müller A, Störk E, Zocher G, Schneider T, Stehle T. 2018. Structural basis of cell wall peptidoglycan amidation by the GatD/MurT complex of Staphylococcus aureus. *Sci. Rep.* 8:12953.

Ozuolmez D, Na H, Lever MA, Kjeldsen KU, Jørgensen BB, Plugge CM. 2015. Methanogenic archaea and sulfate reducing bacteria co-cultured on acetate: teamwork or coexistence? *Front. Microbiol.* 6:492.

Packiam M, Weinrick B, Jacobs WRJ, Maurelli AT. 2015. Structural characterization of muropeptides from Chlamydia trachomatis peptidoglycan by mass spectrometry resolves “chlamydial anomaly”. *Proc. Natl. Acad. Sci. U. S. A.* 112:11660–11665.

Papineau D, She Z, Dodd MS, Iacoviello F, Slack JF, Hauri E, Shearing P, Little CTS. 2022. Metabolically diverse primordial microbial communities in Earth's oldest seafloor-hydrothermal jasper. *Sci. Adv.* 8:eabm2296.

Pazos M, Peters K. 2019. Peptidoglycan. In: Kuhn A, editor. Bacterial Cell Walls and Membranes. Cham: Springer International Publishing. p. 127–168. Available from: https://doi.org/10.1007/978-3-030-18768-2_5

Philippe H. 1993. MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res.* 21:5264–5272.

Philippe H, Forterre P. 1999. The rooting of the universal tree of life is not reliable. *J. Mol. Evol.* 49:509–523.

Pilhofer M, Rappl K, Eckl C, Bauer AP, Ludwig W, Schleifer K-H, Petroni G. 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.

Ponce-Toledo RI, López-García P, Moreira D. 2019. Horizontal and endosymbiotic gene transfer in early plastid evolution. *New Phytol.* 224:618–624.

Popa E, Perera N, Kibédi-Szabó CZ, Guy-Evans H, Evans DR, Purcarea C. 2012. The smallest active carbamoyl phosphate synthetase was identified in the human gut archaeon Methanobrevibacter smithii. *J. Mol. Microbiol. Biotechnol.* 22:287–299.

Poux S, Arighi CN, Magrane M, Bateman A, Wei C-H, Lu Z, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B, et al. 2017. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinforma. Oxf. Engl.* 33:3454–3460.

Price NP, Momany FA. 2005. Modeling bacterial UDP-HexNAc: polyprenol-P HexNAc-1-P transferases. *Glycobiology* 15:29R-42R.

Real G, Henriques AO. 2006. Localization of the Bacillus subtilis murB gene within the dcw cluster is important for growth and sporulation. *J. Bacteriol.* 188:1721–1732.

Rodrigues-Oliveira T, Belmok A, Vasconcellos D, Schuster B, Kyaw CM. 2017. Archaeal S-Layers: Overview and Current State of the Art. *Front. Microbiol.* 8:2597.

Samuel BS, Hansen EE, Manchester JK, Coutinho PM, Henrissat B, Fulton R, Latreille P, Kim K, Wilson RK, Gordon JI. 2007. Genomic and metabolic adaptations of Methanobrevibacter smithii to the human gut. *Proc. Natl. Acad. Sci. U. S. A.* 104:10643–10648.

Sato N. 2021. Are Cyanobacteria an Ancestor of Chloroplasts or Just One of the Gene Donors for Plants and Algae? *Genes* 12.

Scheffers D-J, Pinho MG. 2005. Bacterial cell wall synthesis: new insights from localization studies. *Microbiol. Mol. Biol. Rev. MMBR* 69:585–607.

Sharon I, Haque AS, Grogg M, Lahiri I, Seebach D, Leschziner AE, Hilvert D, Schmeing TM. 2021. Structures and function of the amino acid polymerase cyanophycin synthetase. *Nat. Chem. Biol.* 17:1101–1110.

Shen Y, Buick R, Canfield DE. 2001. Isotopic evidence for microbial sulphate reduction in the early Archaean era. *Nature* 410:77–81.

Shen Y, Chou C-Y, Chang G-G, Tong L. 2006. Is dimerization required for the catalytic activity of bacterial biotin carboxylase? *Mol. Cell* 22:807–818.

Shi D, Caldovic L, Tuchman M. 2018. Sources and Fates of Carbamyl Phosphate: A Labile Energy-Rich Molecule with Multiple Facets. *Biology* 7:34.

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, et al. 2017. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr. Biol. CB* 27:958–967.

Slesarev AI, Mezhevaya KV, Makarova KS, Polushin NN, Shcherbinina OV, Shakhova VV, Belova GI, Aravind L, Natale DA, Rogozin IB, et al. 2002. The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. *Proc. Natl. Acad. Sci. U. S. A.* 99:4644–4649.

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, et al. 1997. Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. *J. Bacteriol.* 179:7135–7155.

Sorokin DY, Makarova KS, Abbas B, Ferrer M, Golyshin PN, Galinski EA, Ciordia S, Mena MC, Merkel AY, Wolf YI, et al. 2017. Discovery of extremely halophilic, methyl-reducing euryarchaea provides insights into the evolutionary origin of methanogenesis. *Nat. Microbiol.* 2:17081.

Spang A, Ettema TJG. 2017. Archaeal evolution: The methanogenic roots of Archaea. *Nat. Microbiol.* 2:17109.

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

Subedi BP, Martin WF, Carbone V, Duin EC, Cronin B, Sauter J, Schofield LR, Sutherland-Smith AJ, Ronimus RS. 2021. Archaeal pseudomurein and bacterial murein cell wall biosynthesis share a common evolutionary ancestry. *FEMS Microbes* 2:xtab012.

Subedi BP, Schofield LR, Carbone V, Wolf M, Martin WF, Ronimus RS, Sutherland-Smith AJ. 2022. Structural characterisation of methanogen pseudomurein cell wall peptide ligases homologous to bacterial MurE/F murein peptide ligases. *Microbiology,* [Internet] 168. Available from: https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.001235

Tamames J. 2001. Evolution of gene order conservation in prokaryotes. *Genome Biol.* 2:RESEARCH0020.

van Teeseling MCF, Mesman RJ, Kuru E, Espaillat A, Cava F, Brun YV, VanNieuwenhze MS, Kartal B, van Niftrik L. 2015. Anammox Planctomycetes have a peptidoglycan cell wall. *Nat. Commun.* 6:6878.

The UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49:D480–D489.

Ueno Y, Yamada K, Yoshida N, Maruyama S, Isozaki Y. 2006. Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. *Nature* 440:516–519.

Vollmer W, Blanot D, de Pedro MA. 2008. Peptidoglycan structure and architecture. *FEMS Microbiol. Rev.* 32:149–167.

Walker MC, van der Donk WA. 2016. The many roles of glutamate in metabolism. *J. Ind. Microbiol. Biotechnol.* 43:419–430.

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

Wolfe JM, Fournier GP. 2018. Horizontal gene transfer constrains the timing of methanogen evolution. *Nat. Ecol. Evol.* 2:897–903.

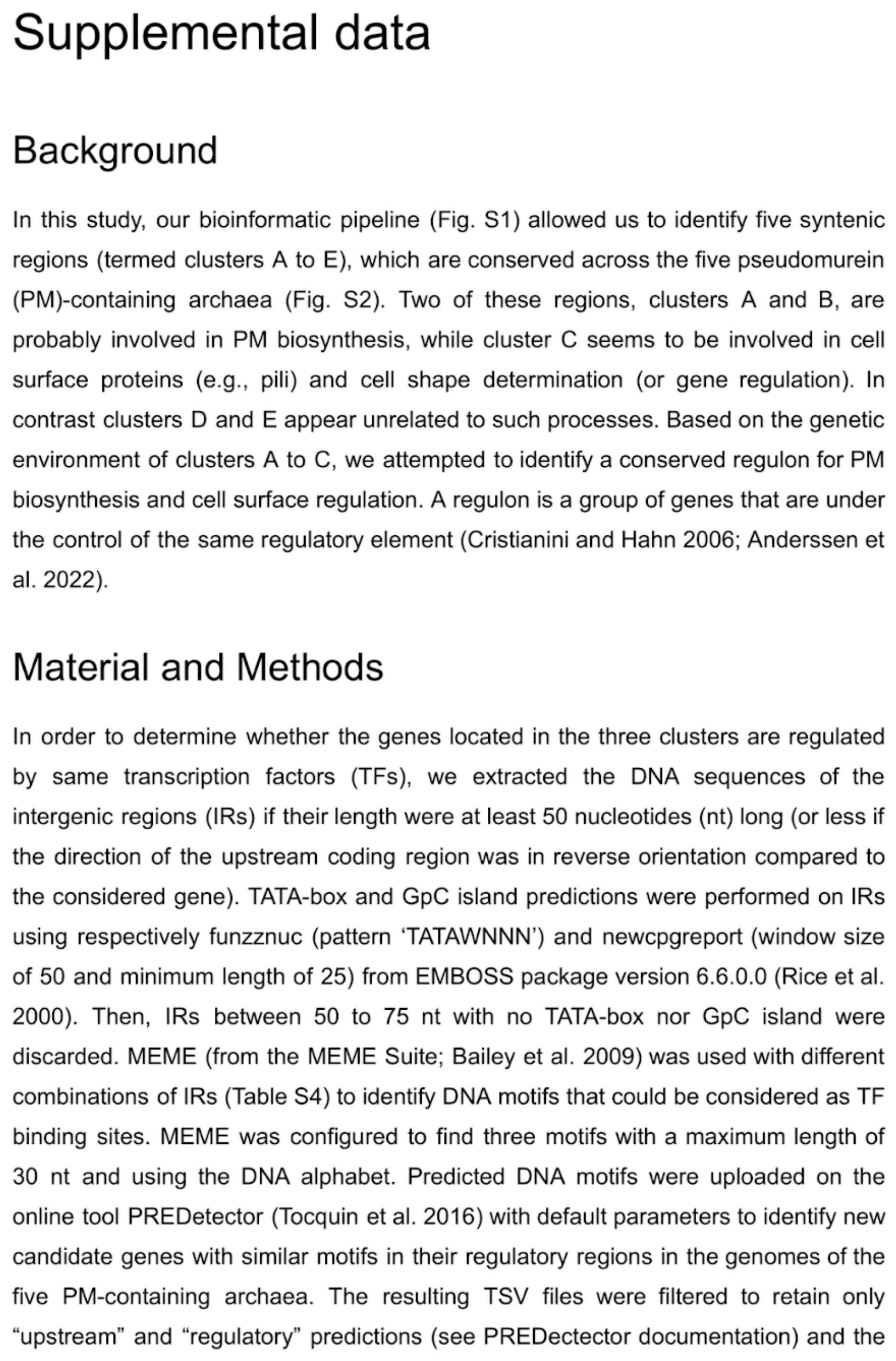
Zawadzke LE, Norcia M, Desbonnet CR, Wang H, Freeman-Cook K, Dougherty TJ. 2008. Identification of an inhibitor of the MurC enzyme, which catalyzes an essential step in the peptidoglycan precursor synthesis pathway. *Assay Drug*

Dev. Technol. 6:95–103.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19:153.

Zouch H, Karray F, Armougom F, Chifflet S, Hirschler-Réa A, Kharrat H, Kamoun L, Ben Hania W, Ollivier B, Sayadi S, et al. 2017. Microbial Diversity in Sulfate-Reducing Marine Sediment Enrichment Cultures Associated with Anaerobic Biotransformation of Coastal Stockpiled Phosphogypsum (Sfax, Tunisia). *Front. Microbiol.* 8:1583.

2.3.2. Supplementary Material



Supplemental data

Background

In this study, our bioinformatic pipeline (Fig. S1) allowed us to identify five syntenic regions (termed clusters A to E), which are conserved across the five pseudomurein (PM)-containing archaea (Fig. S2). Two of these regions, clusters A and B, are probably involved in PM biosynthesis, while cluster C seems to be involved in cell surface proteins (e.g., pili) and cell shape determination (or gene regulation). In contrast clusters D and E appear unrelated to such processes. Based on the genetic environment of clusters A to C, we attempted to identify a conserved regulon for PM biosynthesis and cell surface regulation. A regulon is a group of genes that are under the control of the same regulatory element (Cristianini and Hahn 2006; Anderssen et al. 2022).

Material and Methods

In order to determine whether the genes located in the three clusters are regulated by same transcription factors (TFs), we extracted the DNA sequences of the intergenic regions (IRs) if their length were at least 50 nucleotides (nt) long (or less if the direction of the upstream coding region was in reverse orientation compared to the considered gene). TATA-box and GpC island predictions were performed on IRs using respectively funzznuc (pattern ‘TATAWNNN’) and newcpgreport (window size of 50 and minimum length of 25) from EMBOSS package version 6.6.0.0 (Rice et al. 2000). Then, IRs between 50 to 75 nt with no TATA-box nor GpC island were discarded. MEME (from the MEME Suite; Bailey et al. 2009) was used with different combinations of IRs (Table S4) to identify DNA motifs that could be considered as TF binding sites. MEME was configured to find three motifs with a maximum length of 30 nt and using the DNA alphabet. Predicted DNA motifs were uploaded on the online tool PREDetector (Tocquin et al. 2016) with default parameters to identify new candidate genes with similar motifs in their regulatory regions in the genomes of the five PM-containing archaea. The resulting TSV files were filtered to retain only “upstream” and “regulatory” predictions (see PREDectector documentation) and the gene loci were used to fetch the corresponding protein accessions from GeneSpy (Garcia et al. 2019) GFM files.

Results

DNA motifs identified with MEME did not show significant E-values (from 1.5e-002 to 7.1e+003), indicating that MEME struggled to discover reliable motifs. Nevertheless, we retained the five best motifs across all combinations to identify genes presenting potential TF-binding sites. Despite MEME not using all input sequences to discover motifs (Table S4), we searched for the discovered motifs in all PM-containing archaea. However, quite unsurprisingly, PREDectector predictions solely worked in organisms from which DNA sequences had been used for motif discovery, except in one case: while no sequence from *Methanobrevibacter smithii* had been used for motif prediction, some gene loci were identified in this organism using the second motif. The OGs of the corresponding protein products were then filtered using classify-ali.pl (see Material and Methods in the main text) keeping only those with proteins found in the five PM-containing archaea or at least four of them (one Methanopyrales and three Methanobacteriales), thereby allowing one missing gene in Methanobacteriales. In total, 112 OGs with at least five PM-containing archaea proteins and 19 with at least four were identified, of which 21 OGs had already been identified in the main pipeline for identifying PM biosynthesis candidate proteins (see main text). Among the new OGs, two were identified with different motifs (Table S1, sheet 5), (1) OG0000311 (a glutamate--tRNA ligase) with motifs 1, 4 and 5 in *Methanopyrus sp.*, yet its upstream region (**M-a**; see Figure S2 and Table S4) had been used for motif discovery, (2) OG0000359 with motif 2 in *Methanopyrus sp.* and with motif 3 in *Methanobacterium congolense*. It is a single-copy gene present in the ten archaea and it corresponds to the transcription factor Pcc1, which regulates the cell cycle and polar growth (Kisseleva-Romanova et al. 2006). To investigate whether OG0000359 could actually include the transcription factor regulating PM biosynthesis, we performed the same pipeline of analysis (see Material and Methods above). Due to the difficulty to define a DNA upstream sequence in *Methanothermus fervidus*, we did not include it for motif discovery. The best motif identified by MEME in the upstream region of OG0000359 genes is an 8-nt motif with an E-value of 9.0e-002. It allowed us to identify 185 OGs with at least five PM-containing archaea proteins and 40 with at least four of them. As above, out of these 225 OGs, 25 had already been identified (see main text). The newly identified OGs were then intersected with those selected using classify-ali.pl (i.e., multi-copy genes in PM-containing archaea (paralogs) but existing in a single copy in other archaea). Unfortunately, none of the new OGs passed this filter. Unlike bacteria (Anderssen et al. 2022), the obtained results showed that such a regulon-oriented pipeline is ineffective when predicting regulatory elements in archaea. However, it is not clear whether it is only ineffective in this specific case or if the approach cannot be applied to any extant archaea.

References

Anderssen S, Naômé A, Jadot C, Brans A, Tocquin P, Rigali S. 2022. AURTHO: Autoregulation of transcription factors as facilitator of cis-acting element discovery. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* 1865:194847.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* 37:W202–W208.

Cristianini N, Hahn MW. 2006. Introduction to Computational Genomics: A Case Studies Approach. Cambridge University Press Available from: https://books.google.be/books?id=t3lkngEACAAJ

Garcia PS, Jauffrit F, Grangeasse C, Brochier-Armanet C. 2019. GeneSpy, a user-friendly and flexible genomic context visualizer. *Bioinformatics* 35:329–331.

Kisseleva-Romanova E, Lopreiato R, Baudin-Baillieu A, Rousselle J-C, Ilan L, Hofmann K, Namane A, Mann C, Libri D. 2006. Yeast homolog of a cancer-testis antigen defines a new transcription complex. *EMBO J.* 25:3576–3585.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet. TIG* 16:276–277.

Tocquin P, Naômé A, Jourdan S, Anderssen S, Hiard S, van Wezel GP, Hanikenne M, Baurain D, Rigali S. 2016. PREDetector 2.0: Online and Enhanced Version of the Prokaryotic Regulatory Elements Detector Tool. *bioRxiv* [Internet]. Available from: https://www.biorxiv.org/content/early/2016/11/01/084780

(a)

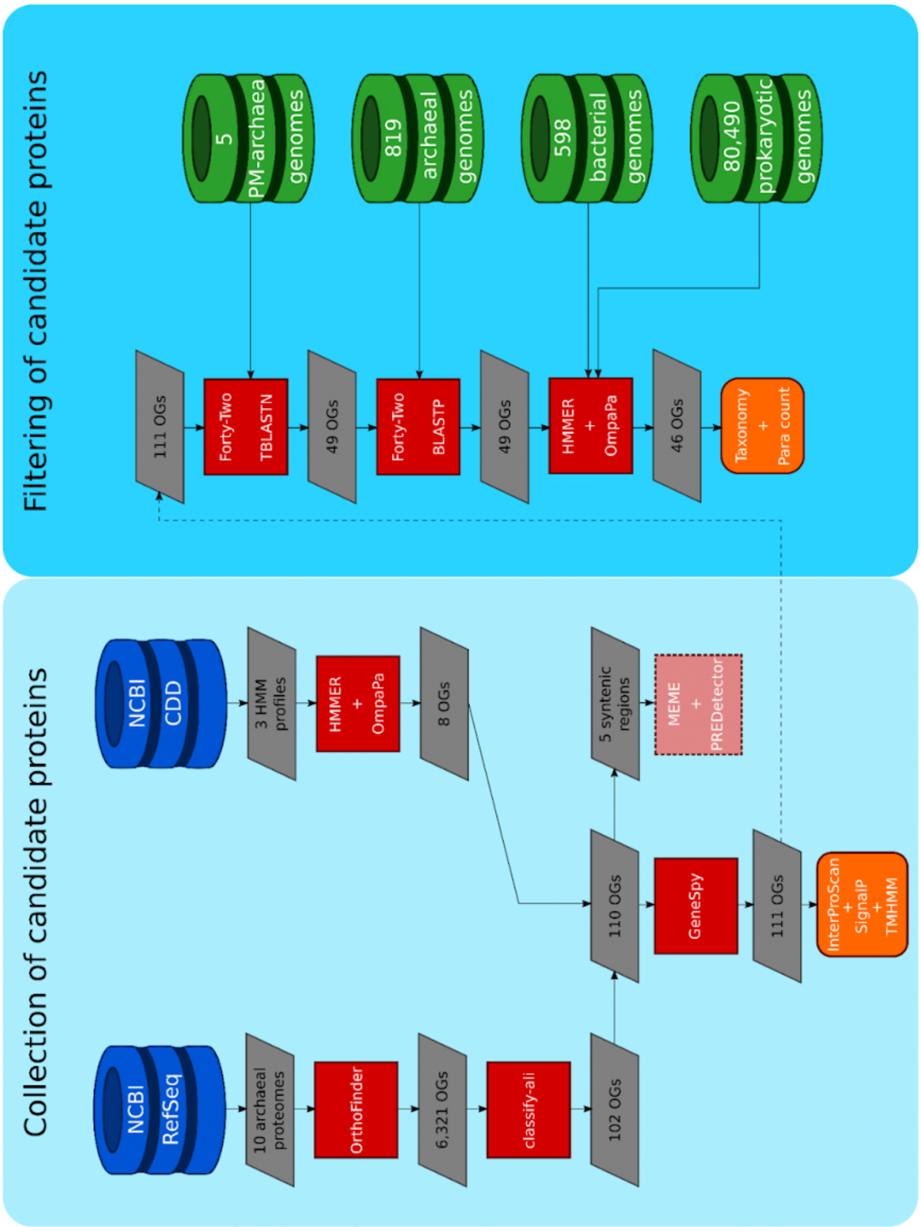
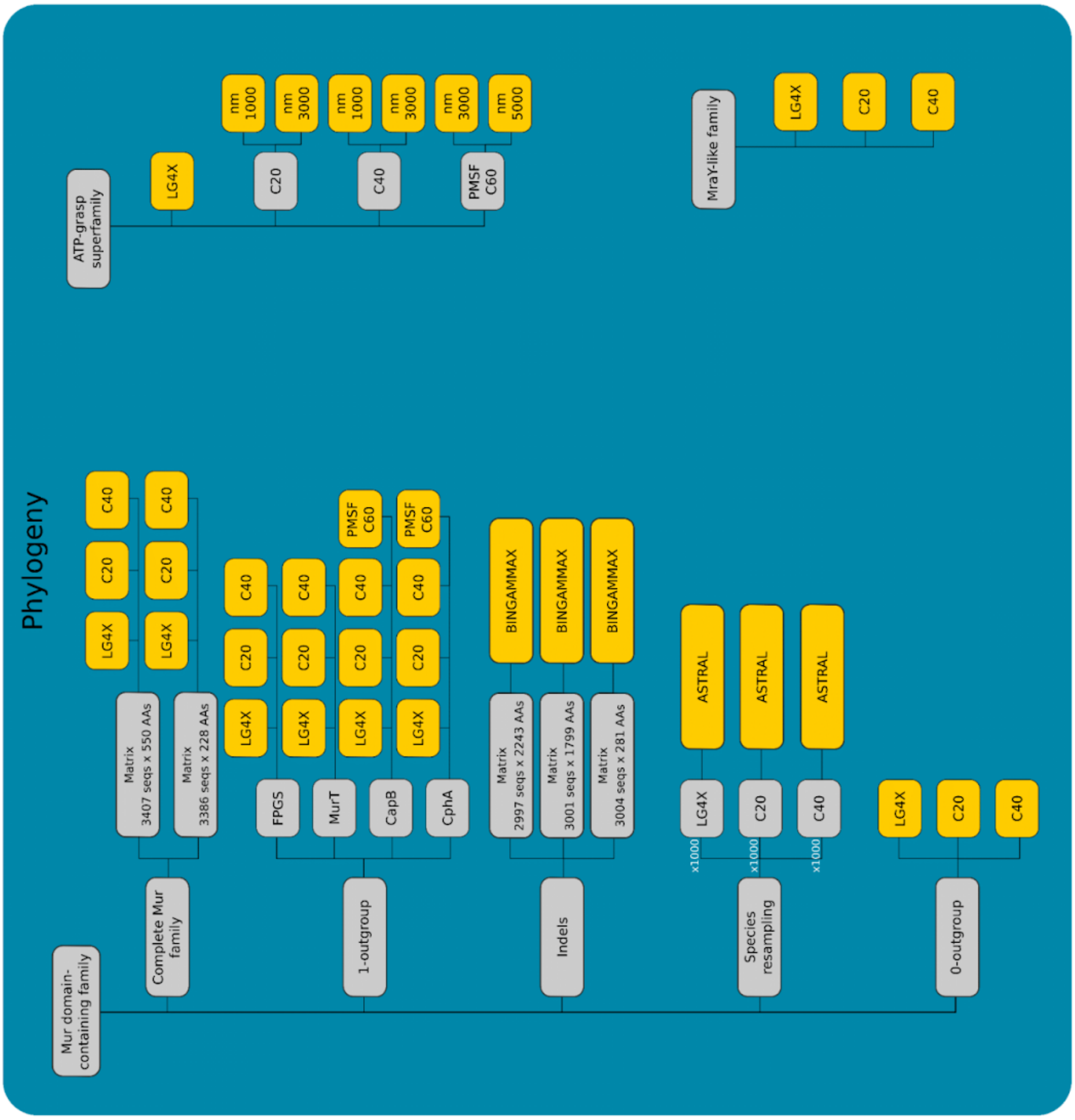
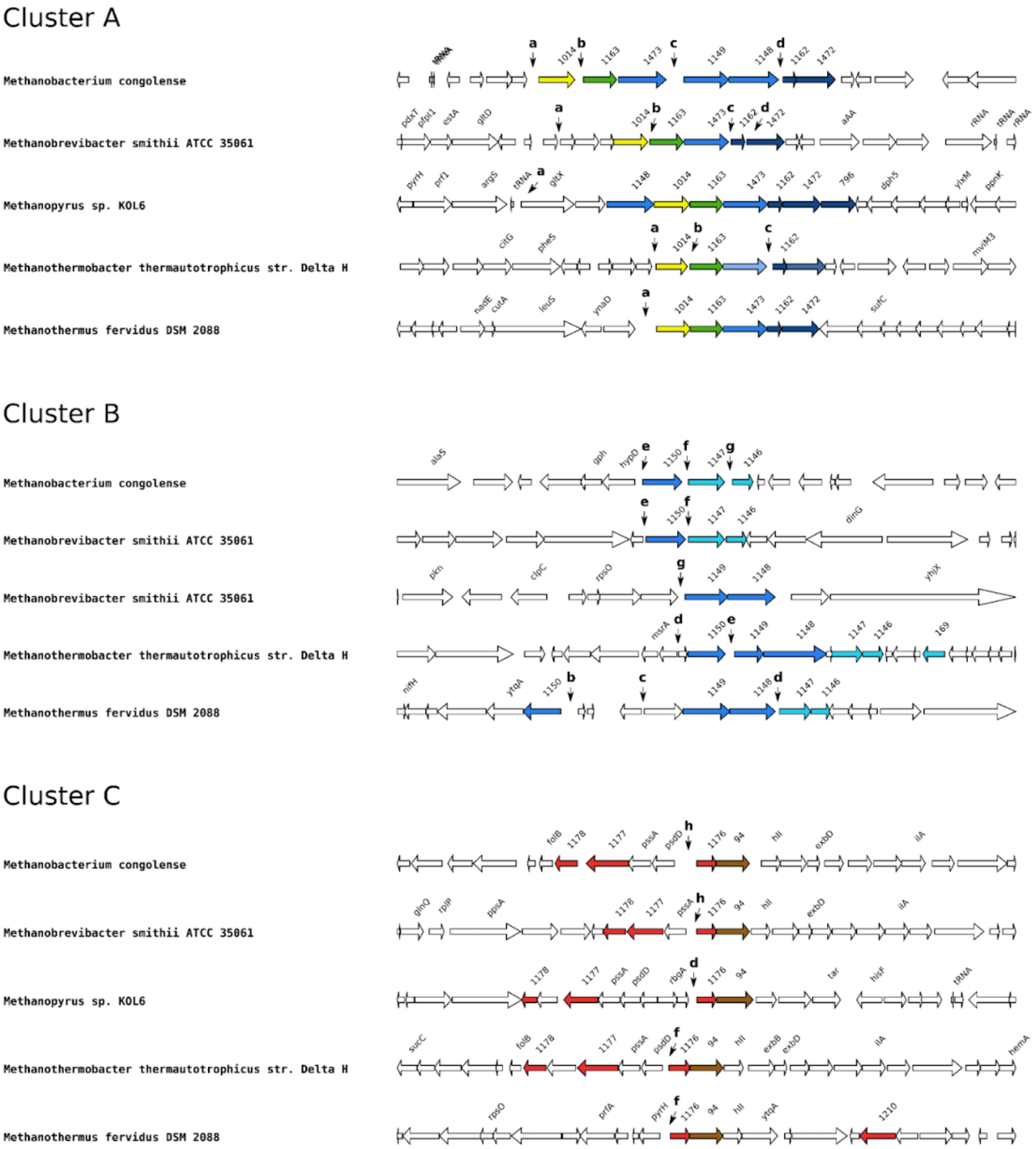


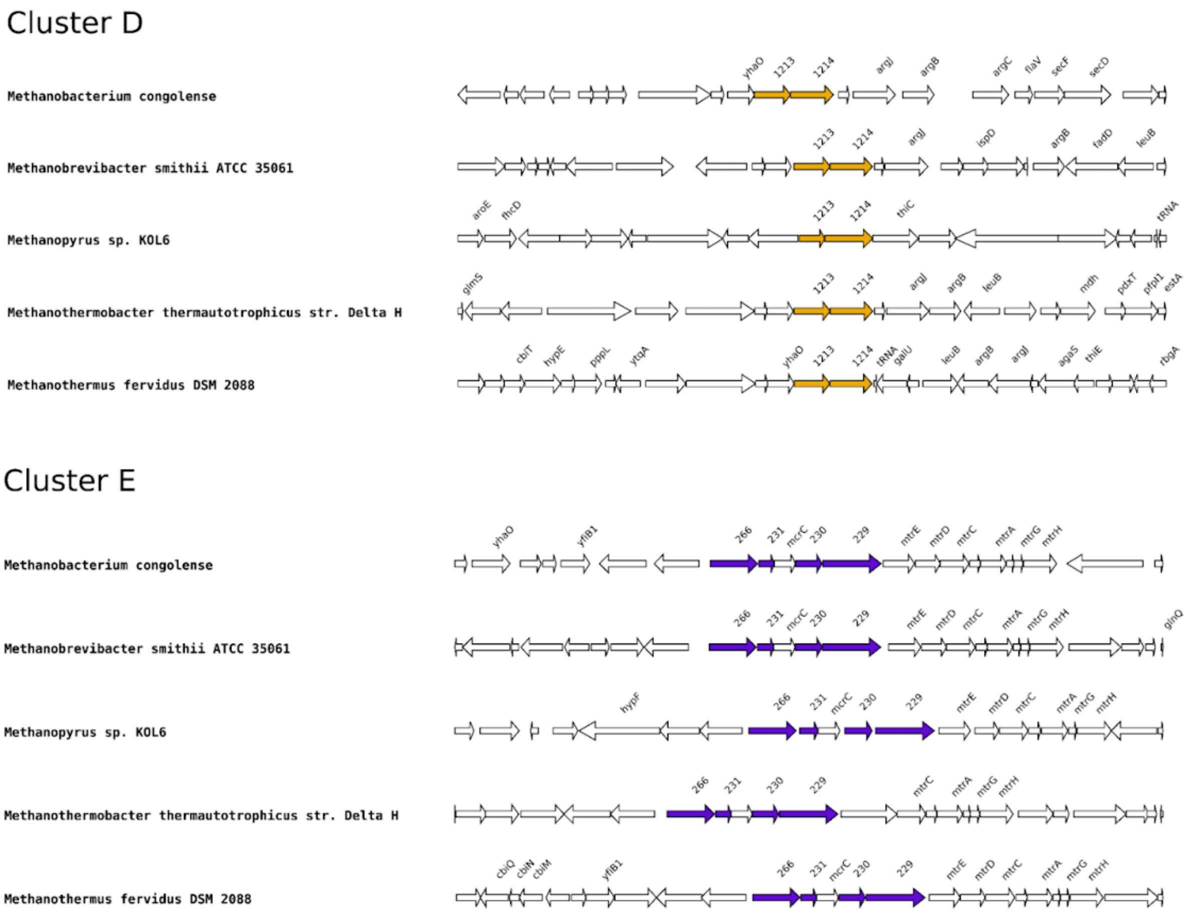
Figure S1. Overview of the methods used for the bioinformatic analyses carried out during this study (see Material and Methods of the main text for

details). (a, previous page) Main steps of the pipeline for the identification and filtering of the OGs potentially involved in PM biosynthesis. **(b, this page)** Details of the datasets and AA substitution models used for phylogenetic analyses.

(b)




Cluster A
Methanobacterium congolense
tRNA
a
1014
b
1163
1473
c
1149
1148
d
1162
1472
Methanobrevibacter smithii ATCC 35061
pdxT
pfpI1
estA
gltD
a
1014
b
1163
1473
c
1162
d
1472
aAA
rRNA
tRNA
rRNA
Methanopyrus sp. KOL6
pyrH
prf1
argS
tRNA
a
gltX
1148
1014
1163
1473
1162
1472
796
dph5
ylxM
ppnK
Methanothermobacter thermautotrophicus str. Delta H
citG
pheS
a
1014
b
1163
c
1162
mvIM3
Methanothermus fervidus DSM 2088
nadE
cutA
leuS
ynaD
a
1014
1163
1473
1162
1472
sufC
Cluster B
Methanobacterium congolense
alaS
gph
hypD
e
1150
f
1147
g
1146
Methanobrevibacter smithii ATCC 35061
e
1150
f
1147
1146
dinG
Methanobrevibacter smithii ATCC 35061
pkn
clpC
rpsO
g
1149
1148
yhjX
Methanothermobacter thermautotrophicus str. Delta H
msrA
d
1150
e
1149
1148
1147
1146
169
Methanothermus fervidus DSM 2088
nifH
ytqA
1150
b
c
1149
1148
d
1147
1146
Cluster C
Methanobacterium congolense
folB
1178
1177
pssA
psdD
h
1176
94
hII
exbD
ilA
Methanobrevibacter smithii ATCC 35061
glnQ
rplP
ppsA
1178
1177
pssA
h
1176
94
hII
exbD
ilA
Methanopyrus sp. KOL6
1178
1177
pssA
psdD
rbgA
d
1176
94
tar
hisF
tRNA
Methanothermobacter thermautotrophicus str. Delta H
sucC
folB
1178
1177
pssA
psdD
f
1176
94
hII
exbB
exbD
ilA
hemA
Methanothermus fervidus DSM 2088
rpsO
prfA
pyrH
f
1176
94
hII
ytqA
1210



Cluster D

Cluster E

Figure S2. Genetic organization of the five clusters. Identified clustered genes are colored, whereas other genes are left in white. Above each identified gene is indicated its corresponding orthologous group (OG) number (without 'OG' and the following 0s). Intergenic regions used during the regulon pipeline (see Supplemental data) are indicated with a lowercase letter (Table S4). The intergenic regions "b" and "c" of *Methanopyrus* sp. are not shown in this figure and correspond to the upstream region of the OG0001150 and OG0001147, genes respectively.

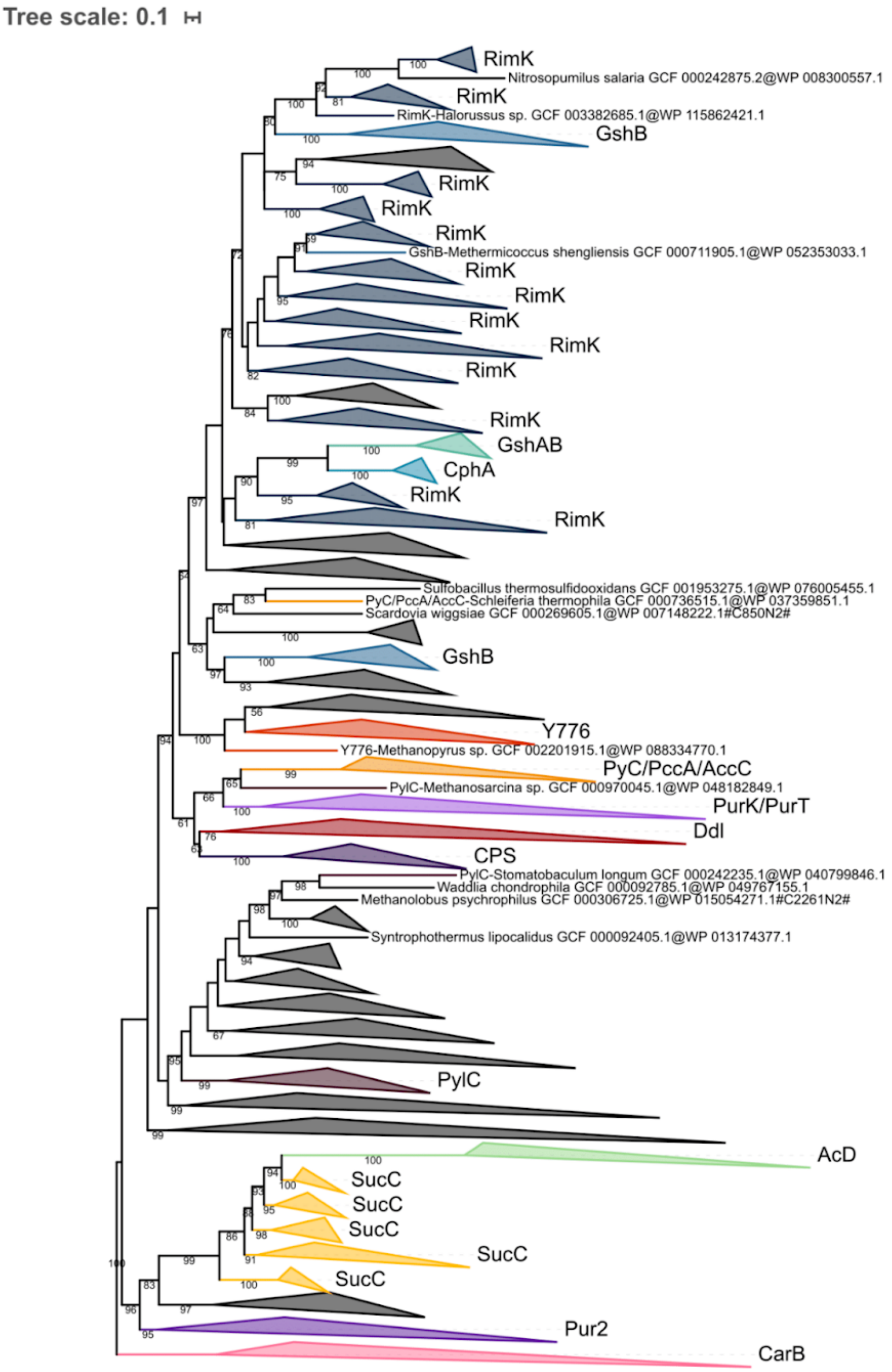

Tree scale: 0.1
RimK
Nitrosopumilus salaria GCF 000242875.2@WP 008300557.1
RimK
RimK-Halorussus sp. GCF 003382685.1@WP 115862421.1
GshB
RimK
RimK
RimK
GshB-Methermicoccus shengliensis GCF 000711905.1@WP 052353033.1
RimK
RimK
RimK
RimK
RimK
RimK
GshAB
CphA
RimK
RimK
Sulfobacillus thermosulfidooxidans GCF 001953275.1@WP 076005455.1
PyC/PccA/AccC-Schleiferia thermophila GCF 000736515.1@WP 037359851.1
Scardovia wiggsiae GCF 000269605.1@WP 007148222.1#C850N2#
GshB
Y776
Y776-Methanopyrus sp. GCF 002201915.1@WP 088334770.1
PyC/PccA/AccC
PylC-Methanosarcina sp. GCF 000970045.1@WP 048182849.1
PurK/PurT
Ddl
CPS
PylC-Stomatobaculum longum GCF 000242235.1@WP 040799846.1
Waddlia chondrophila GCF 000092785.1@WP 049767155.1
Methanolobus psychrophilus GCF 000306725.1@WP 015054271.1#C2261N2#
Syntrophothermus lipocalidus GCF 000092405.1@WP 013174377.1
PylC
AcD
SucC
SucC
SucC
SucC
SucC
Pur2
CarB

Figure S3. Phylogenetic tree of the ATP-grasp superfamily rooted on CarB. The tree was inferred from a matrix of 2,194 sequences x 180 unambiguously aligned AAs using IQ-TREE **under the C20+G4 model with 1000 iterations**. Tree visualization was performed using iTOL. Bootstrap support values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation based on reference sequences. Black collapsed branches correspond to unannotated sequences.

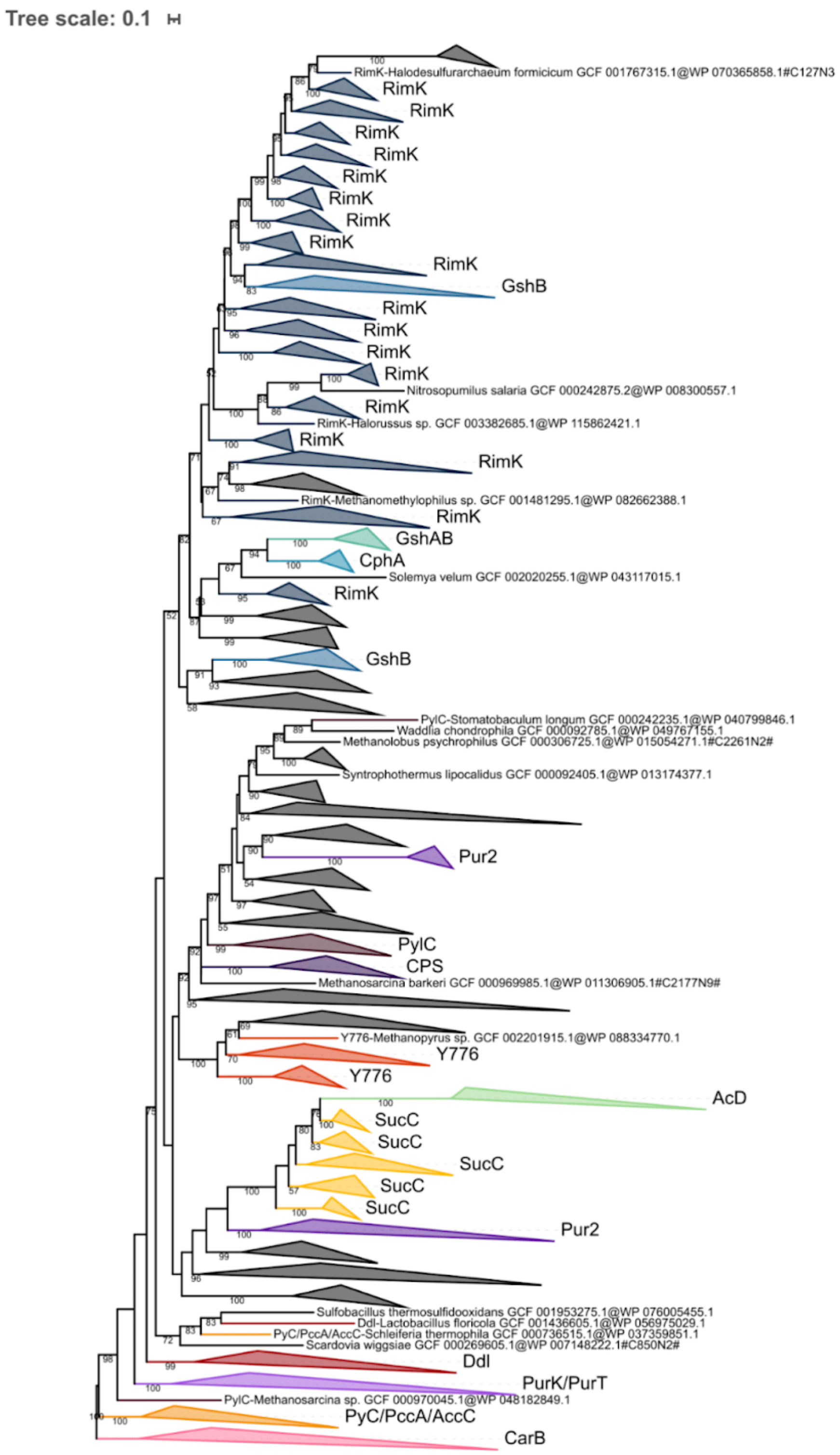

Tree scale: 0.1
RimK-Halodesulfurarchaeum formicicum GCF 001767315.1@WP 070365858.1#C127N3
RimK
RimK
RimK
RimK
RimK
RimK
RimK
RimK
RimK
GshB
RimK
RimK
RimK
RimK
Nitrosopumilus salaria GCF 000242875.2@WP 008300557.1
RimK
RimK-Halorussus sp. GCF 003382685.1@WP 115862421.1
RimK
RimK
RimK-Methanomethylophilus sp. GCF 001481295.1@WP 082662388.1
RimK
GshAB
CphA
Solemya velum GCF 002020255.1@WP 043117015.1
RimK
GshB
PylC-Stomatobaculum longum GCF 000242235.1@WP 040799846.1
Waddlia chondrophila GCF 000092785.1@WP 049767155.1
Methanolobus psychrophilus GCF 000306725.1@WP 015054271.1#C2261N2#
Syntrophothermus lipocalidus GCF 000092405.1@WP 013174377.1
Pur2
PylC
CPS
Methanosarcina barkeri GCF 000969985.1@WP 011306905.1#C2177N9#
Y776-Methanopyrus sp. GCF 002201915.1@WP 088334770.1
Y776
Y776
AcD
SucC
SucC
SucC
SucC
SucC
Pur2
Sulfobacillus thermosulfidooxidans GCF 001953275.1@WP 076005455.1
Ddl-Lactobacillus floricola GCF 001436605.1@WP 056975029.1
PyC/PccA/AccC-Schleiferia thermophila GCF 000736515.1@WP 037359851.1
Scardovia wiggsiae GCF 000269605.1@WP 007148222.1#C850N2#
Ddl
PurK/PurT
PylC-Methanosarcina sp. GCF 000970045.1@WP 048182849.1
PyC/PccA/AccC
CarB

Figure S4. Phylogenetic tree of the ATP-grasp superfamily rooted on CarB. The tree was inferred from a matrix of 2,194 sequences x 180 unambiguously aligned AAs using IQ-TREE **under the C20+G4 model with 3000 iterations**. Tree visualization was performed using iTOL. Bootstrap support values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation based on reference sequences. Black collapsed branches correspond to unannotated sequences.

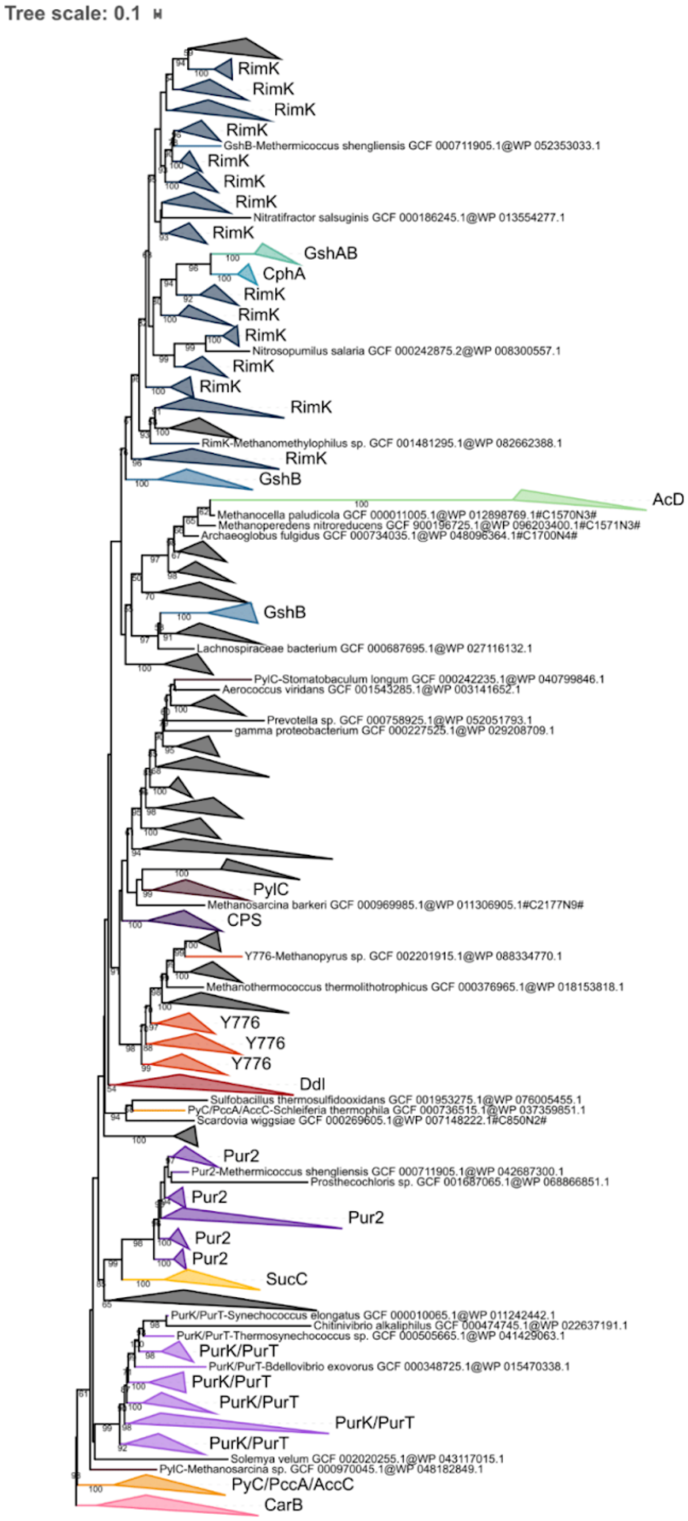

Tree scale: 0.1
RimK
RimK
RimK
RimK
GshB-Methermicoccus shengliensis GCF 000711905.1@WP 052353033.1
RimK
RimK
RimK
Nitratifractor salsuginis GCF 000186245.1@WP 013554277.1
RimK
GshAB
CphA
RimK
RimK
RimK
Nitrosopumilus salaria GCF 000242875.2@WP 008300557.1
RimK
RimK
RimK
RimK-Methanomethylophilus sp. GCF 001481295.1@WP 082662388.1
RimK
GshB
AcD
Methanocella paludicola GCF 000011005.1@WP 012898769.1#C1570N3#
Methanoperedens nitroreducens GCF 900196725.1@WP 096203400.1#C1571N3#
Archaeoglobus fulgidus GCF 000734035.1@WP 048096364.1#C1700N4#
GshB
Lachnospiraceae bacterium GCF 000687695.1@WP 027116132.1
PylC-Stomatobaculum longum GCF 000242235.1@WP 040799846.1
Aerococcus viridans GCF 001543285.1@WP 003141652.1
Prevotella sp. GCF 000758925.1@WP 052051793.1
gamma proteobacterium GCF 000227525.1@WP 029208709.1
PylC
Methanosarcina barkeri GCF 000969985.1@WP 011306905.1#C2177N9#
CPS
Y776-Methanopyrus sp. GCF 002201915.1@WP 088334770.1
Methanothermococcus thermolithotrophicus GCF 000376965.1@WP 018153818.1
Y776
Y776
Y776
Ddl
Sulfobacillus thermosulfidooxidans GCF 001953275.1@WP 076005455.1
PyC/PccA/AccC-Schleiferia thermophila GCF 000736515.1@WP 037359851.1
Scardovia wiggsiae GCF 000269605.1@WP 007148222.1#C850N2#
Pur2
Pur2-Methermicoccus shengliensis GCF 000711905.1@WP 042687300.1
Prosthecochloris sp. GCF 001687065.1@WP 068866851.1
Pur2
Pur2
Pur2
Pur2
SucC
PurK/PurT-Synechococcus elongatus GCF 000010065.1@WP 011242442.1
Chitinivibrio alkaliphilus GCF 000474745.1@WP 022637191.1
PurK/PurT-Thermosynechococcus sp. GCF 000505665.1@WP 041429063.1
PurK/PurT
PurK/PurT-Bdellovibrio exovorus GCF 000348725.1@WP 015470338.1
PurK/PurT
PurK/PurT
PurK/PurT
PurK/PurT
Solemya velum GCF 002020255.1@WP 043117015.1
PylC-Methanosarcina sp. GCF 000970045.1@WP 048182849.1
PyC/PccA/AccC
CarB

Figure S5. Phylogenetic tree of the ATP-grasp superfamily rooted on CarB. The tree was inferred from a matrix of 2,194 sequences x 180 unambiguously aligned AAs using IQ-TREE **under the C40+G4 model with 1000 iterations**. Tree visualization was performed using iTOL. Bootstrap support values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation based on reference sequences. Black collapsed branches correspond to unannotated sequences.

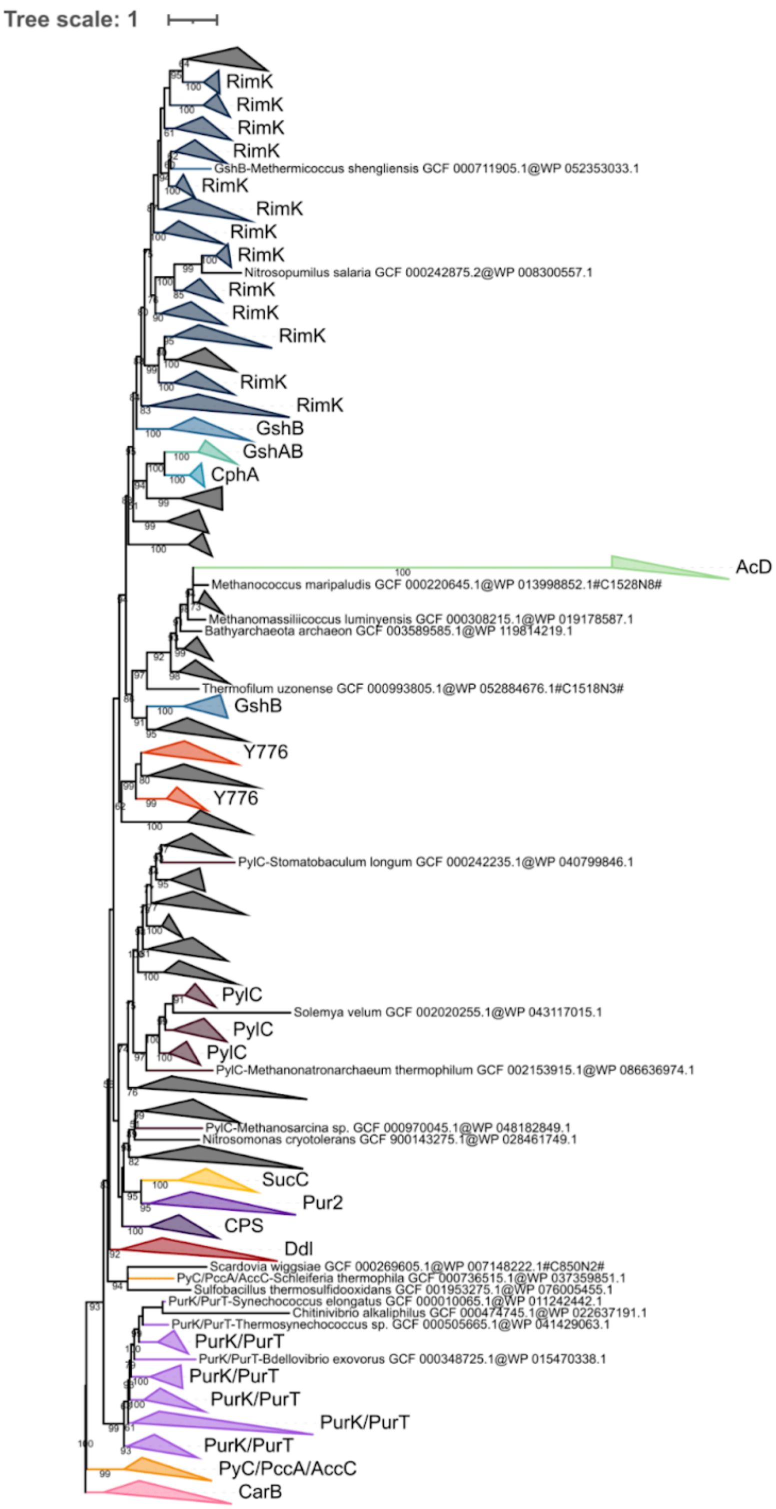

Tree scale: 1
RimK
RimK
RimK
RimK
GshB-Methermicoccus shengliensis GCF 000711905.1@WP 052353033.1
RimK
RimK
RimK
RimK
Nitrosopumilus salaria GCF 000242875.2@WP 008300557.1
RimK
RimK
RimK
RimK
RimK
GshB
GshAB
CphA
AcD
100
Methanococcus maripaludis GCF 000220645.1@WP 013998852.1#C1528N8#
Methanomassiliicoccus luminyensis GCF 000308215.1@WP 019178587.1
Bathyarchaeota archaeon GCF 003589585.1@WP 119814219.1
Thermofilum uzonense GCF 000993805.1@WP 052884676.1#C1518N3#
GshB
Y776
Y776
PylC-Stomatobaculum longum GCF 000242235.1@WP 040799846.1
PylC
Solemya velum GCF 002020255.1@WP 043117015.1
PylC
PylC
PylC-Methanonatronarchaeum thermophilum GCF 002153915.1@WP 086636974.1
PylC-Methanosarcina sp. GCF 000970045.1@WP 048182849.1
Nitrosomonas cryotolerans GCF 900143275.1@WP 028461749.1
SucC
Pur2
CPS
Ddl
Scardovia wiggsiae GCF 000269605.1@WP 007148222.1#C850N2#
PyC/PccA/AccC-Schleiferia thermophila GCF 000736515.1@WP 037359851.1
Sulfobacillus thermosulfidooxidans GCF 001953275.1@WP 076005455.1
PurK/PurT-Synechococcus elongatus GCF 000010065.1@WP 011242442.1
Chitinivibrio alkaliphilus GCF 000474745.1@WP 022637191.1
PurK/PurT-Thermosynechococcus sp. GCF 000505665.1@WP 041429063.1
PurK/PurT
PurK/PurT-Bdellovibrio exovorus GCF 000348725.1@WP 015470338.1
PurK/PurT
PurK/PurT
PurK/PurT
PurK/PurT
PyC/PccA/AccC
CarB

Figure S6. Phylogenetic tree of the ATP-grasp superfamily rooted on CarB. The tree was inferred from a matrix of 2,194 sequences x 180 unambiguously aligned AAs using IQ-TREE **under the C40+G4 model with 3000 iterations**. Tree visualization was performed using iTOL. Bootstrap support values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation based on reference sequences. Black collapsed branches correspond to unannotated sequences.

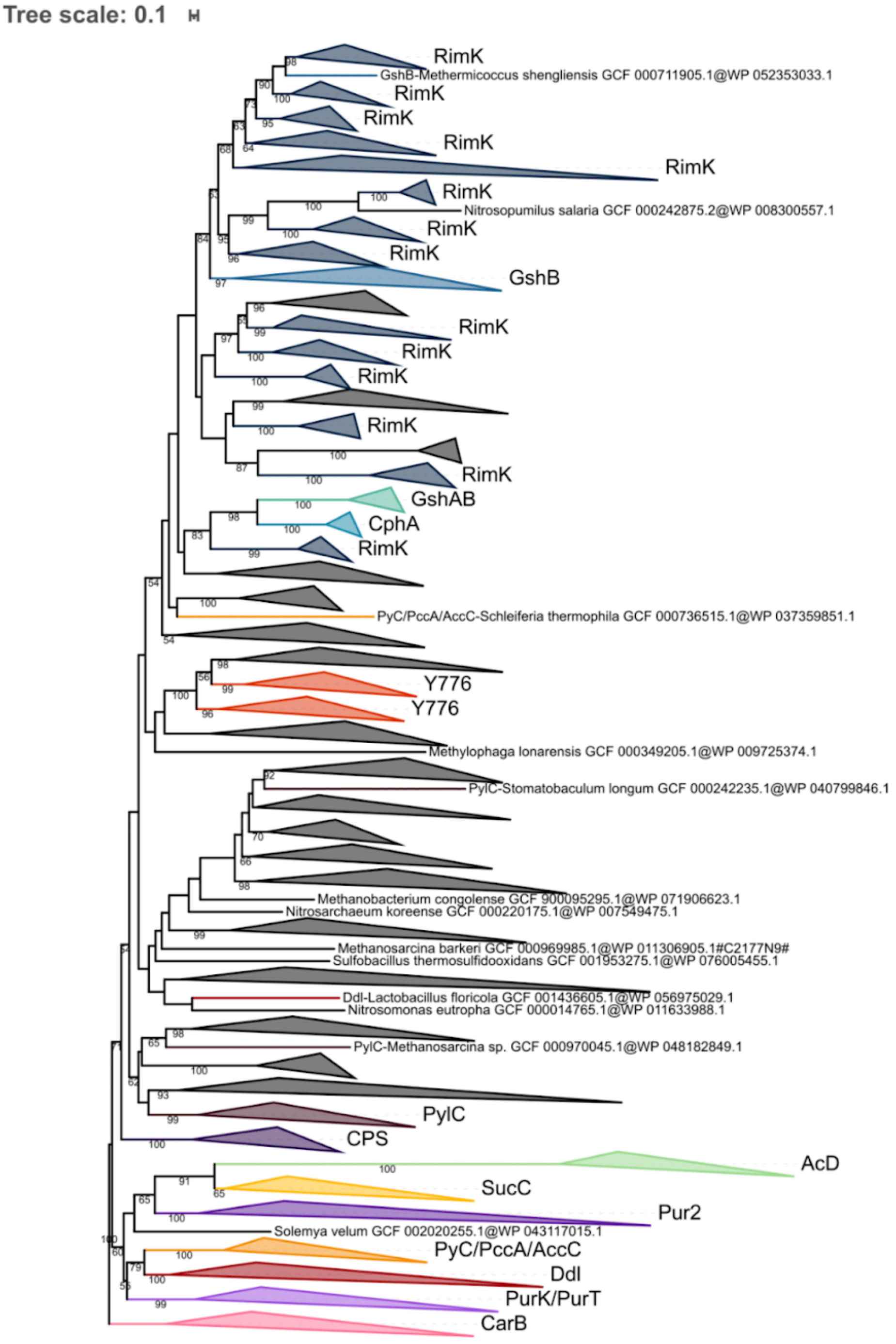

Tree scale: 0.1
RimK
GshB-Methermicoccus shengliensis GCF 000711905.1@WP 052353033.1
RimK
RimK
RimK
RimK
RimK
Nitrosopumilus salaria GCF 000242875.2@WP 008300557.1
RimK
RimK
GshB
RimK
RimK
RimK
RimK
RimK
GshAB
CphA
RimK
PyC/PccA/AccC-Schleiferia thermophila GCF 000736515.1@WP 037359851.1
Y776
Y776
Methylophaga lonarensis GCF 000349205.1@WP 009725374.1
PylC-Stomatobaculum longum GCF 000242235.1@WP 040799846.1
Methanobacterium congolense GCF 900095295.1@WP 071906623.1
Nitrosarchaeum koreense GCF 000220175.1@WP 007549475.1
Methanosarcina barkeri GCF 000969985.1@WP 011306905.1#C2177N9#
Sulfobacillus thermosulfidooxidans GCF 001953275.1@WP 076005455.1
Ddl-Lactobacillus floricola GCF 001436605.1@WP 056975029.1
Nitrosomonas eutropha GCF 000014765.1@WP 011633988.1
PylC-Methanosarcina sp. GCF 000970045.1@WP 048182849.1
PylC
CPS
AcD
SucC
Pur2
Solemya velum GCF 002020255.1@WP 043117015.1
PyC/PccA/AccC
Ddl
PurK/PurT
CarB

Figure S7. Phylogenetic tree of the ATP-grasp superfamily rooted on CarB. The tree was inferred from a matrix of 2,194 sequences x 180 unambiguously aligned AAs using IQ-TREE **under the PMSF LG+C60+G4 model with 3000 iterations**. Tree visualization was performed using iTOL. Bootstrap support values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation based on reference sequences. Black collapsed branches correspond to unannotated sequences.

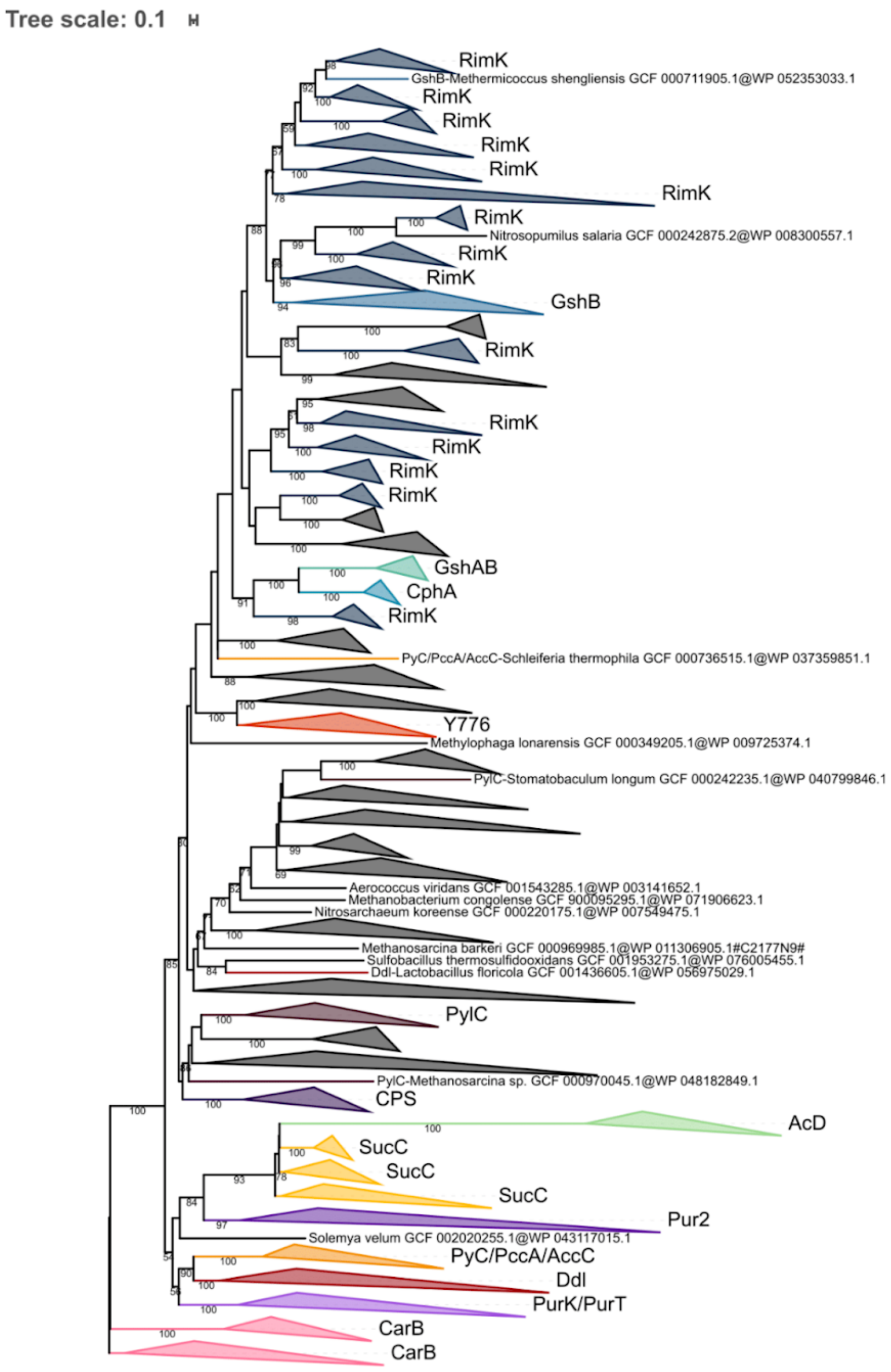

Tree scale: 0.1
RimK
GshB-Methermicoccus shengliensis GCF 000711905.1@WP 052353033.1
RimK
RimK
RimK
RimK
RimK
RimK
Nitrosopumilus salaria GCF 000242875.2@WP 008300557.1
RimK
RimK
GshB
RimK
RimK
RimK
RimK
RimK
GshAB
CphA
RimK
PyC/PccA/AccC-Schleiferia thermophila GCF 000736515.1@WP 037359851.1
Y776
Methylophaga lonarensis GCF 000349205.1@WP 009725374.1
PylC-Stomatobaculum longum GCF 000242235.1@WP 040799846.1
Aerococcus viridans GCF 001543285.1@WP 003141652.1
Methanobacterium congolense GCF 900095295.1@WP 071906623.1
Nitrosarchaeum koreense GCF 000220175.1@WP 007549475.1
Methanosarcina barkeri GCF 000969985.1@WP 011306905.1#C2177N9#
Sulfobacillus thermosulfidooxidans GCF 001953275.1@WP 076005455.1
Ddl-Lactobacillus floricola GCF 001436605.1@WP 056975029.1
PylC
PylC-Methanosarcina sp. GCF 000970045.1@WP 048182849.1
CPS
AcD
SucC
SucC
SucC
Pur2
Solemya velum GCF 002020255.1@WP 043117015.1
PyC/PccA/AccC
Ddl
PurK/PurT
CarB
CarB

Figure S8. Phylogenetic tree of the ATP-grasp superfamily rooted on CarB. The tree was inferred from a matrix of 2,194 sequences x 180 unambiguously aligned AAs using IQ-TREE **under the PMSF LG+C60+G4 model with 5000 iterations**. Tree visualization was performed using iTOL. Bootstrap support values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation based on reference sequences. Black collapsed branches correspond to unannotated sequences.

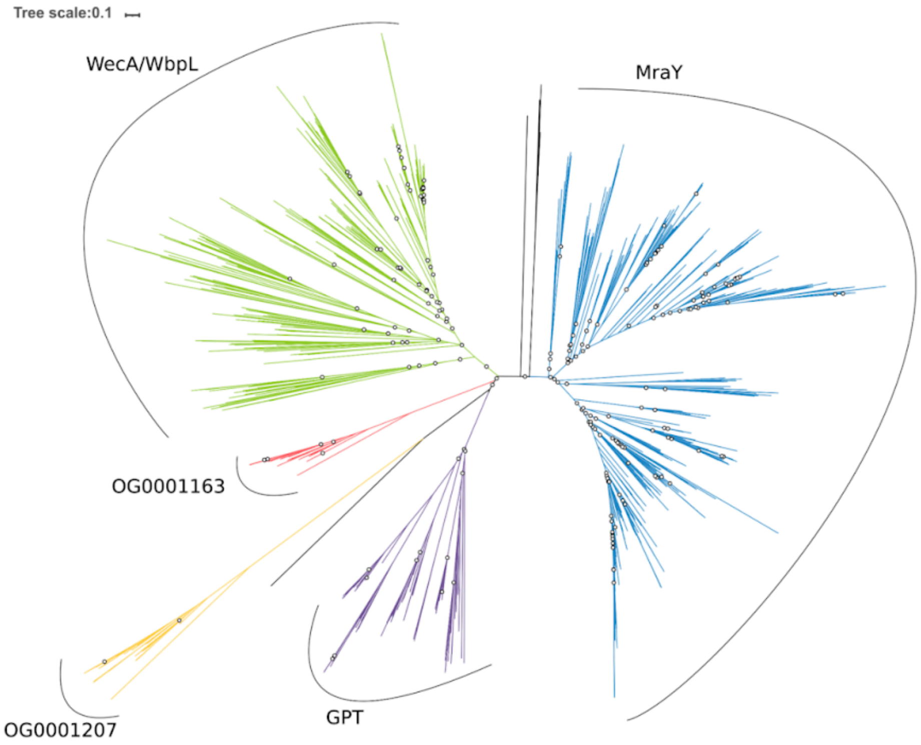


Figure S9. Unrooted phylogenetic tree of the MraY-like family. The tree was constructed from a matrix of 1,070 sequences x 408 unambiguously aligned AAs using IQ-TREE **under the LG4X+R4 model**. Open circles correspond to bootstrap support values lower than 90. Blue sequences correspond to a MraY annotation, green to WecA/WbpL, red to OG0001163 (MraY-like), yellow to OG0001207, purple to GPT, and black to unannotated bacterial sequences.

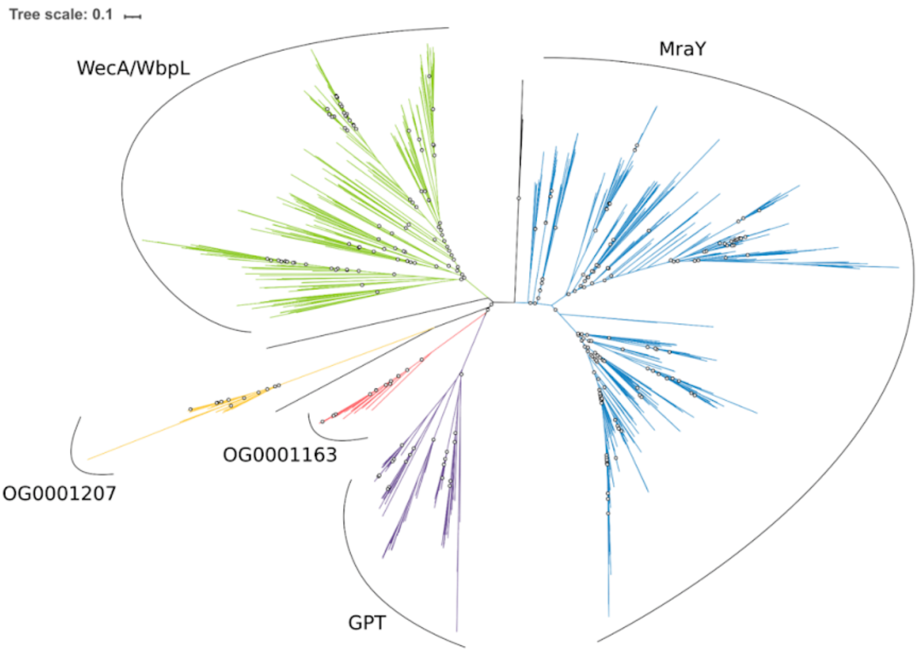


Figure S10. Unrooted phylogenetic tree of the MraY-like family. The tree was constructed from a matrix of 1,070 sequences x 408 unambiguously aligned AAs using IQ-TREE **under the C20+G4 model**. Open circles correspond to bootstrap support values lower than 90. Blue sequences correspond to a MraY annotation, green to WecA/WbpL, red to OG0001163 (MraY-like), yellow to OG0001207, purple to GPT, and black to unannotated bacterial sequences.

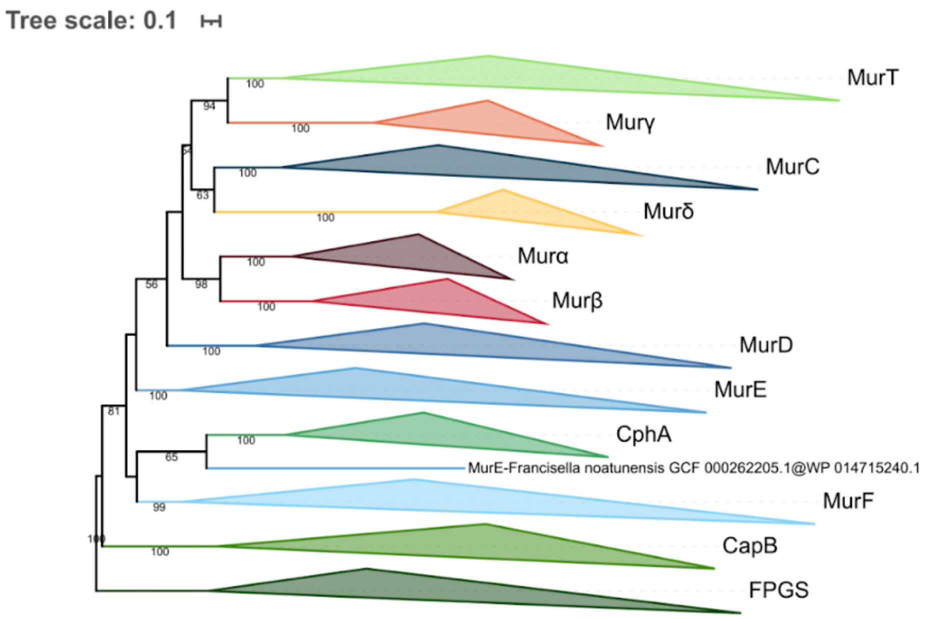


Figure S11. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. The tree was inferred from a matrix of **3407 sequences x 550 unambiguously aligned AAs** using IQ-TREE **under the LG4X+R4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

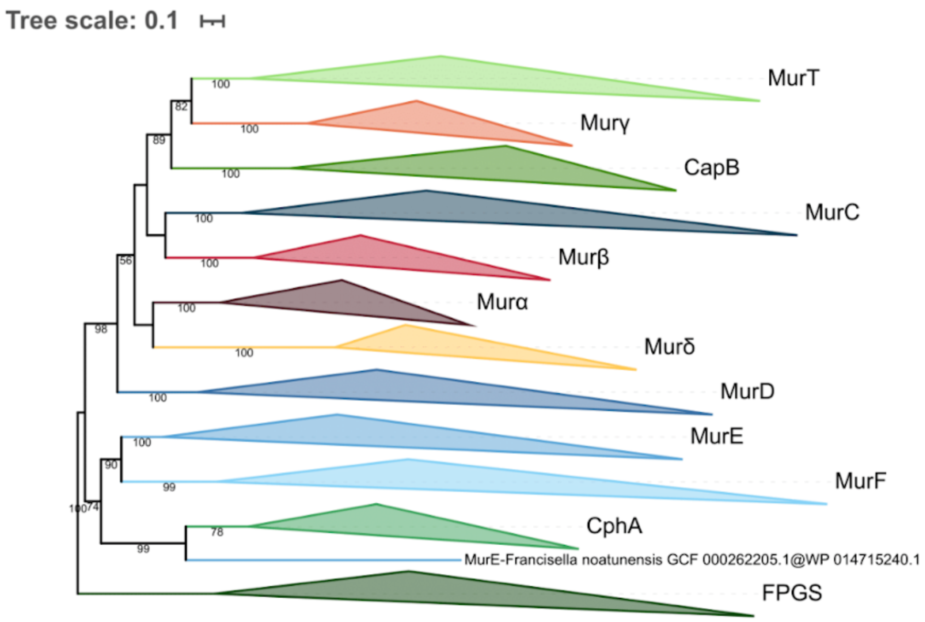


Figure S12. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. The tree was inferred from a matrix of **3407 sequences x 550 unambiguously aligned AAs** using IQ-TREE **under the C20+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

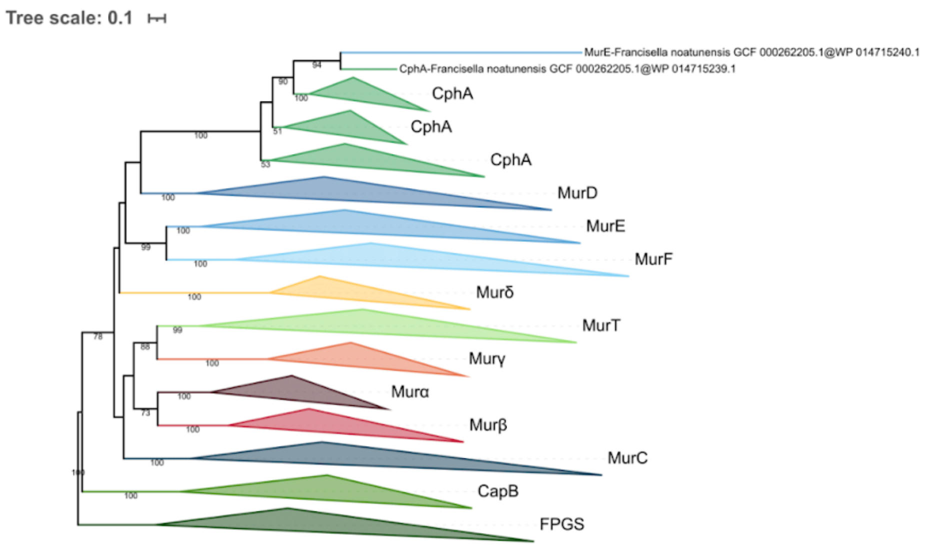


Figure S13. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. The tree was inferred from a matrix of **3407 sequences x 550 unambiguously aligned AAs** using IQ-TREE **under the C40+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

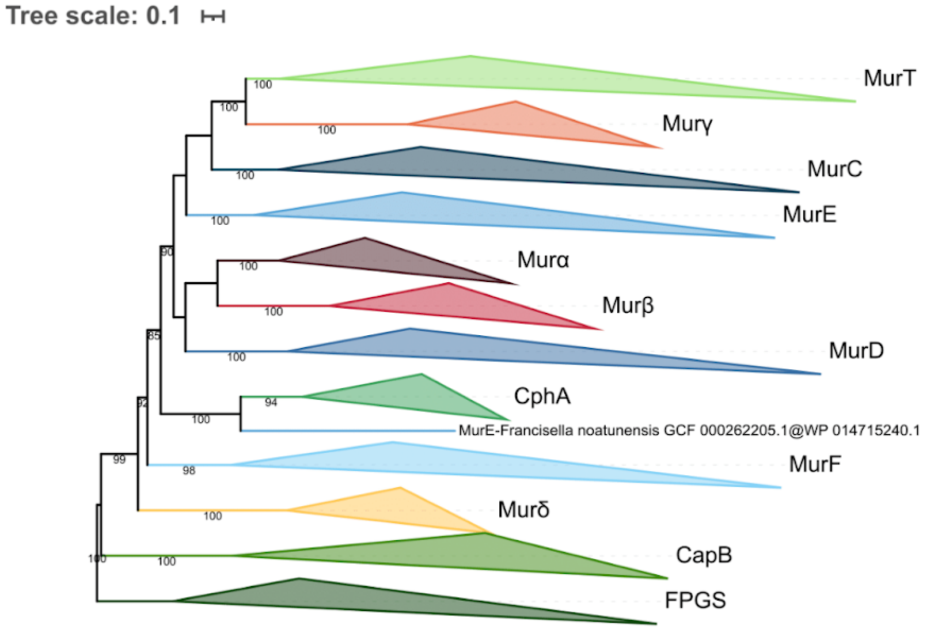


Figure S14. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. The tree was inferred from a matrix of **3386 sequences x 228 unambiguously aligned AAs** using IQ-TREE **under the LG4X+R4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

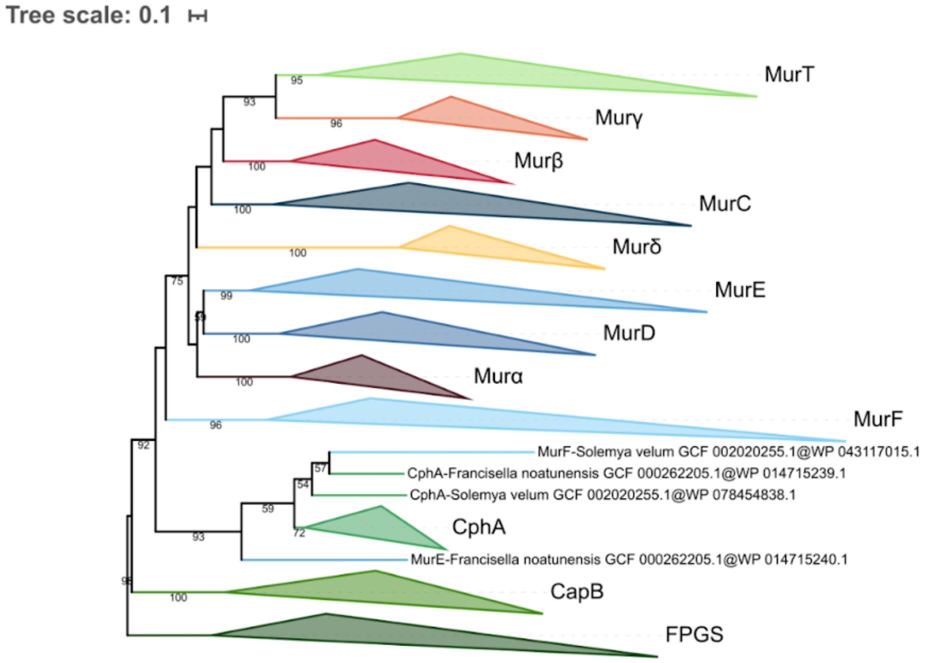

Tree scale: 0.1
95
MurT
93
96
Murγ
100
Murβ
100
MurC
100
Murδ
75
99
MurE
59
100
MurD
100
Murα
96
MurF
92
MurF-Solemya velum GCF 002020255.1@WP 043117015.1
57
CphA-Francisella noatunensis GCF 000262205.1@WP 014715239.1
54
CphA-Solemya velum GCF 002020255.1@WP 078454838.1
59
72
CphA
93
MurE-Francisella noatunensis GCF 000262205.1@WP 014715240.1
100
CapB
FPGS

Figure S15. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. The tree was inferred from a matrix of **3386 sequences x 228 unambiguously aligned AAs** using IQ-TREE **under the C20+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

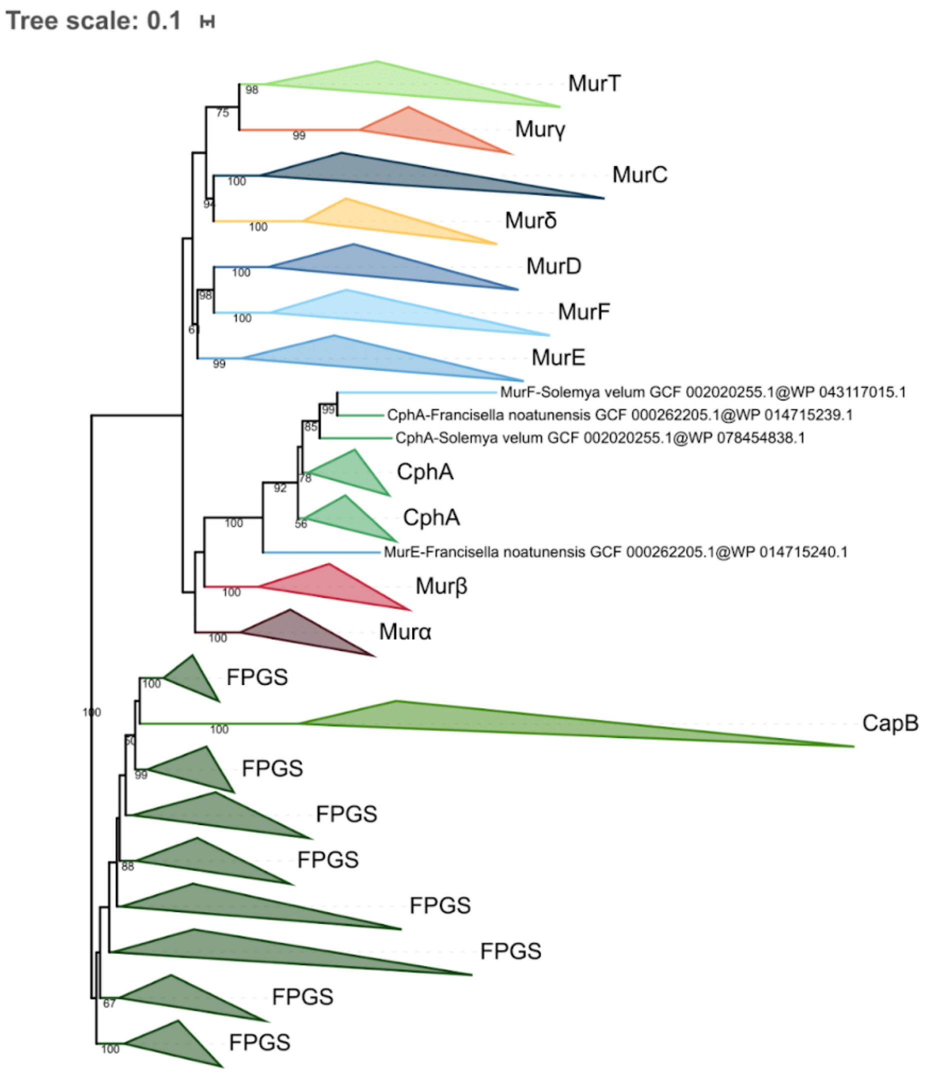

Tree scale: 0.1
MurT
98
75
Murγ
99
MurC
100
94
Murδ
100
MurD
100
98
MurF
100
61
MurE
99
MurF-Solemya velum GCF 002020255.1@WP 043117015.1
99
CphA-Francisella noatunensis GCF 000262205.1@WP 014715239.1
85
CphA-Solemya velum GCF 002020255.1@WP 078454838.1
CphA
78
92
CphA
56
100
MurE-Francisella noatunensis GCF 000262205.1@WP 014715240.1
Murβ
100
Murα
100
FPGS
100
100
CapB
100
50
FPGS
99
FPGS
FPGS
88
FPGS
FPGS
FPGS
67
FPGS
100

Figure S16. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. The tree was inferred from a matrix of **3386 sequences x 228 unambiguously aligned AAs** using IQ-TREE **under the C40+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

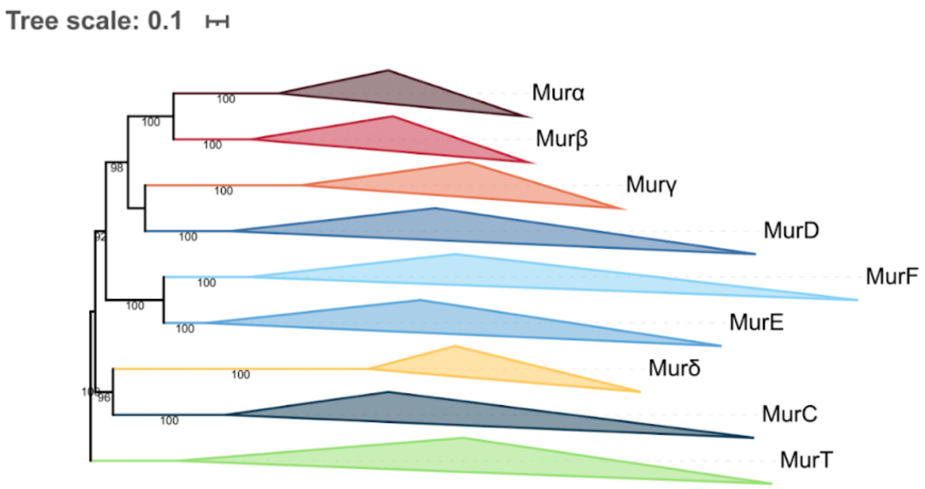


Figure S17. Phylogenetic tree of the Mur domain-containing family rooted on MurT. The tree was inferred from a matrix of **2677 sequences x 525 unambiguously aligned AAs** using IQ-TREE **under the LG4X+R4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

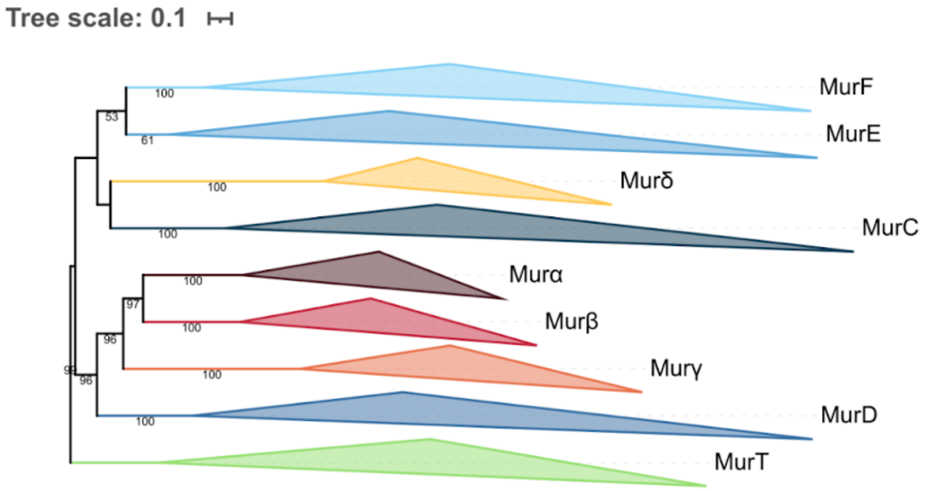


Figure S18. Phylogenetic tree of the Mur domain-containing family rooted on MurT. The tree was inferred from a matrix of **2677 sequences x 525 unambiguously aligned AAs** using IQ-TREE **under the C20+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

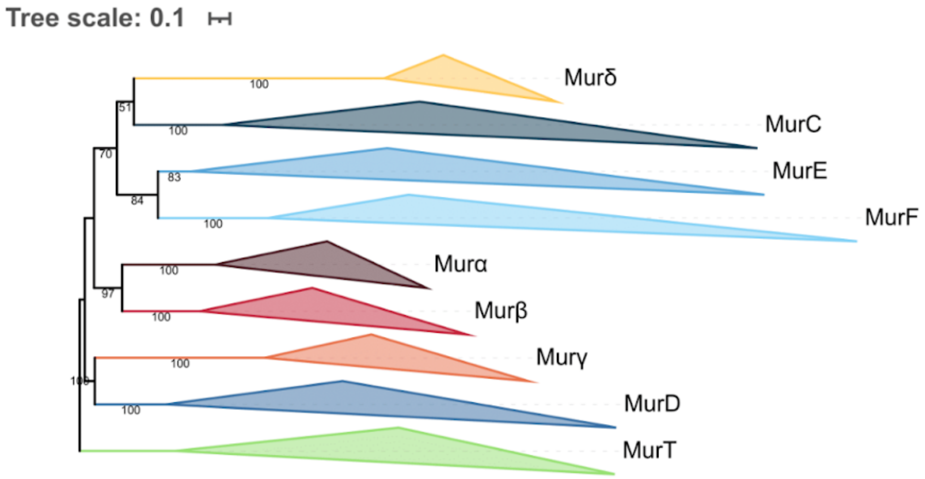


Figure S19. Phylogenetic tree of the Mur domain-containing family rooted on MurT. The tree was inferred from a matrix of **2677 sequences x 525 unambiguously aligned AAs** using IQ-TREE **under the C40+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

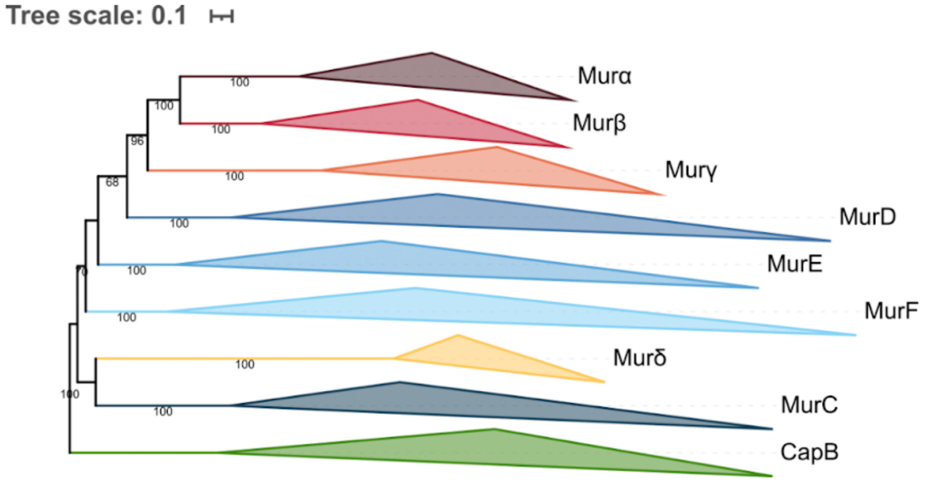


Figure S20. Phylogenetic tree of the Mur domain-containing family rooted on CapB. The tree was inferred from a matrix of **2519 sequences x 532 unambiguously aligned AAs** using IQ-TREE **under the LG4X+R4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

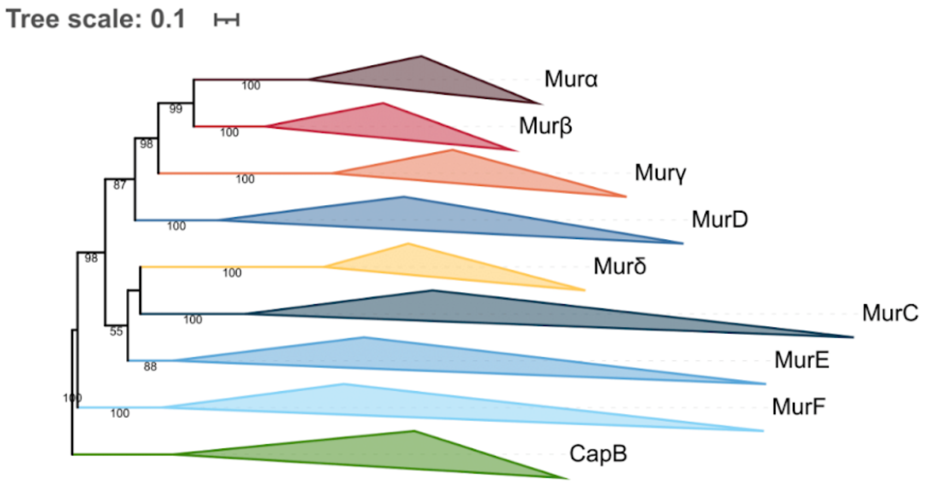


Figure S21. Phylogenetic tree of the Mur domain-containing family rooted on CapB. The tree was inferred from a matrix of **2519 sequences x 532 unambiguously aligned AAs** using IQ-TREE **under the C20+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

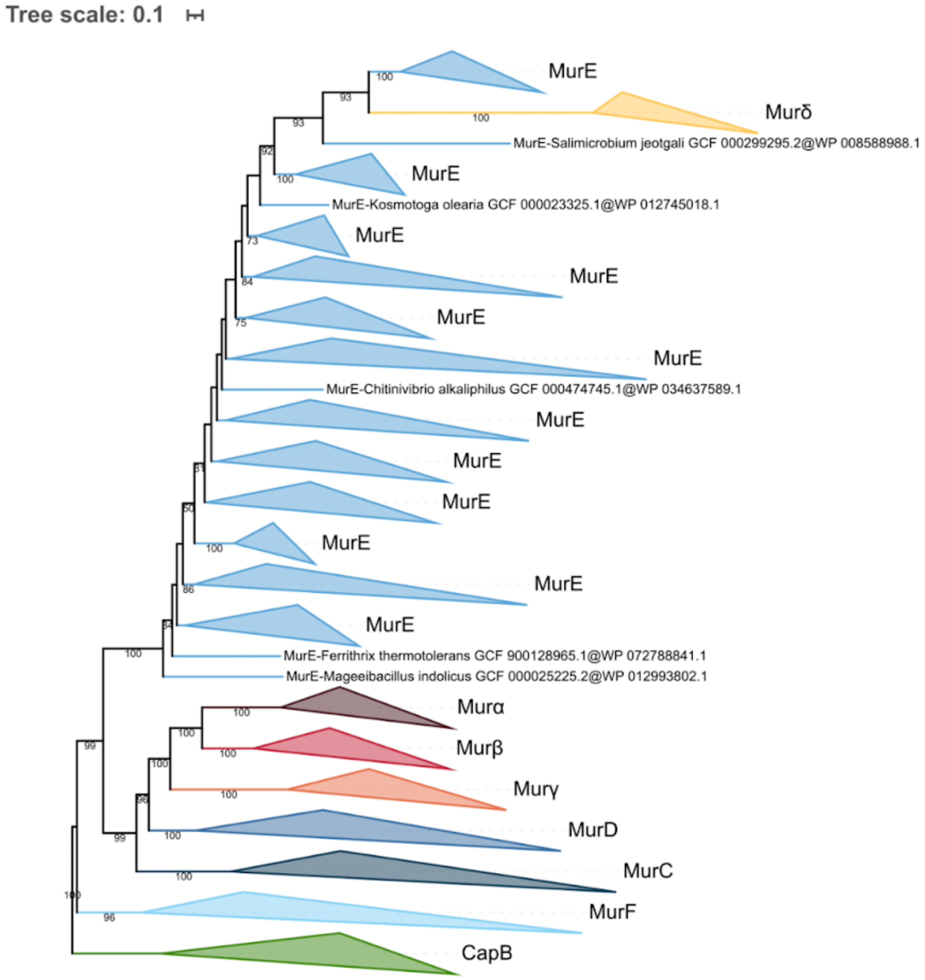

Tree scale: 0.1
MurE
100
93
Murδ
100
93
MurE-Salimicrobium jeotgali GCF 000299295.2@WP 008588988.1
92
MurE
100
MurE-Kosmotoga olearia GCF 000023325.1@WP 012745018.1
MurE
73
MurE
84
MurE
75
MurE
MurE-Chitinivibrio alkaliphilus GCF 000474745.1@WP 034637589.1
MurE
MurE
81
MurE
50
MurE
100
MurE
86
MurE
84
MurE-Ferrithrix thermotolerans GCF 900128965.1@WP 072788841.1
100
MurE-Mageeibacillus indolicus GCF 000025225.2@WP 012993802.1
Murα
100
100
Murβ
100
99
100
Murγ
100
96
MurD
100
99
MurC
100
100
MurF
96
CapB

Figure S22. Phylogenetic tree of the Mur domain-containing family rooted on CapB. The tree was inferred from a matrix of **2519 sequences x 532 unambiguously aligned AAs** using IQ-TREE **under the C40+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

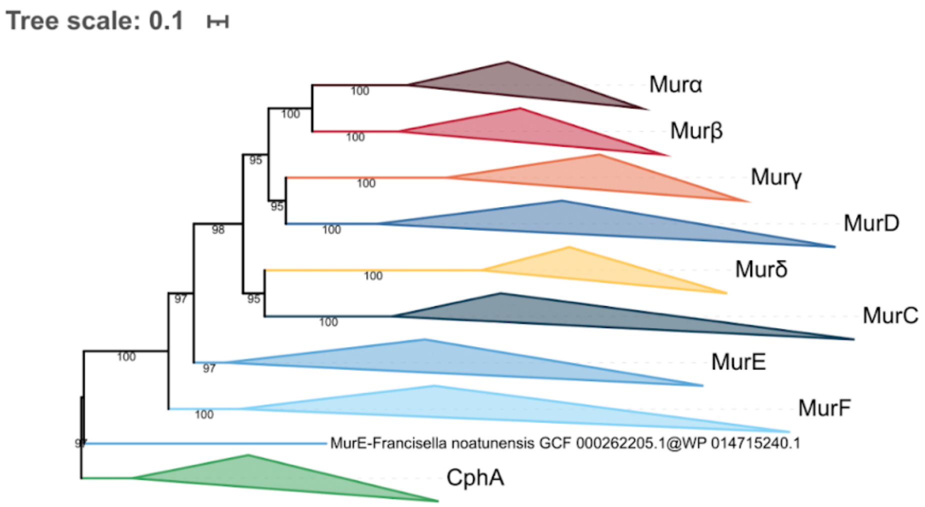


Figure S23. Phylogenetic tree of the Mur domain-containing family rooted on CphA. The tree was inferred from a matrix of **2461 sequences x 539 unambiguously aligned AAs** using IQ-TREE **under the LG4X+R4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

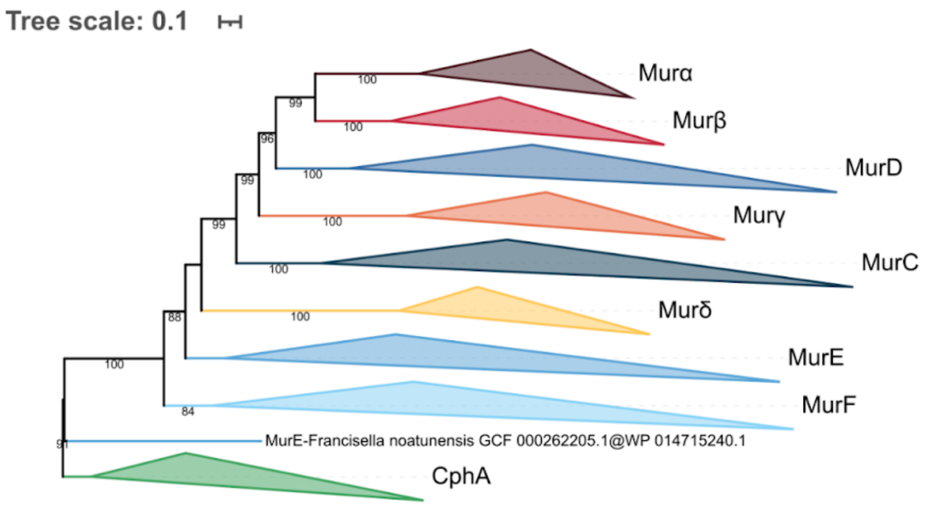


Figure S24. Phylogenetic tree of the Mur domain-containing family rooted on CphA. The tree was inferred from a matrix of **2461 sequences x 539 unambiguously aligned AAs** using IQ-TREE **under the C20+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

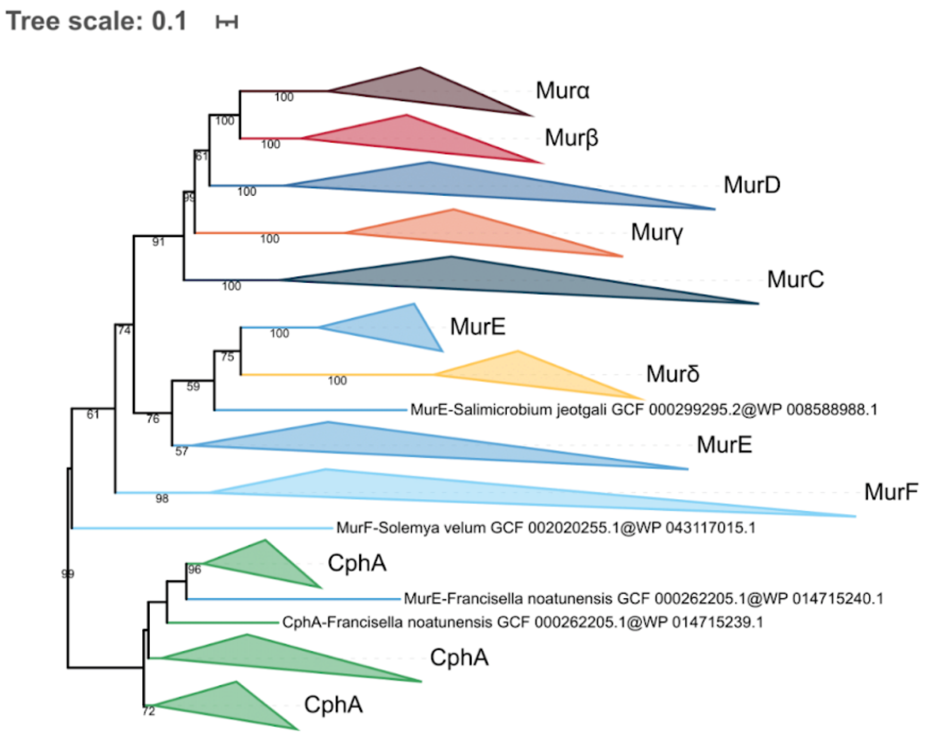

Tree scale: 0.1
Murα
100
100
Murβ
100
61
MurD
100
99
Murγ
100
91
MurC
100
74
MurE
100
75
Murδ
100
59
MurE-Salimicrobium jeotgali GCF 000299295.2@WP 008588988.1
61
76
MurE
57
MurF
98
MurF-Solemya velum GCF 002020255.1@WP 043117015.1
CphA
96
MurE-Francisella noatunensis GCF 000262205.1@WP 014715240.1
99
CphA-Francisella noatunensis GCF 000262205.1@WP 014715239.1
CphA
CphA
72

Figure S25. Phylogenetic tree of the Mur domain-containing family rooted on CphA. The tree was inferred from a matrix of **2461 sequences x 539 unambiguously aligned AAs** using IQ-TREE **under the C40+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

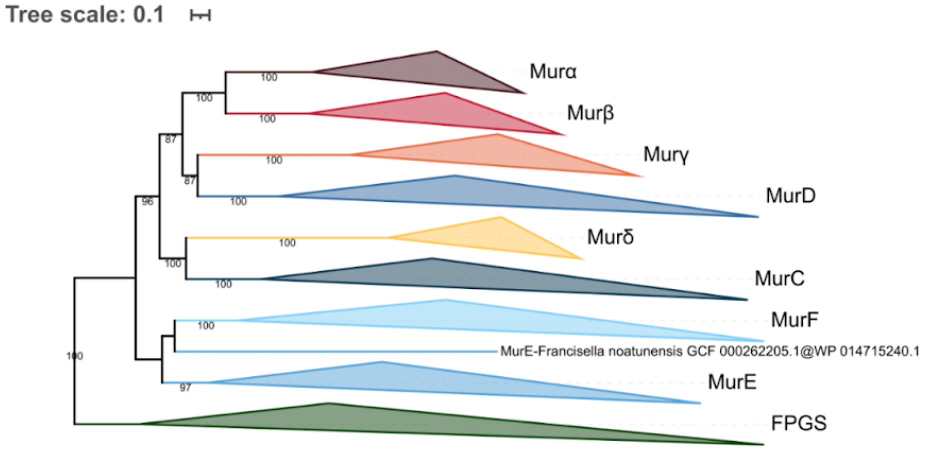


Figure S26. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. The tree was inferred from a matrix of **3,046 sequences x 543 unambiguously aligned AAs** using IQ-TREE **under the LG4X+R4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

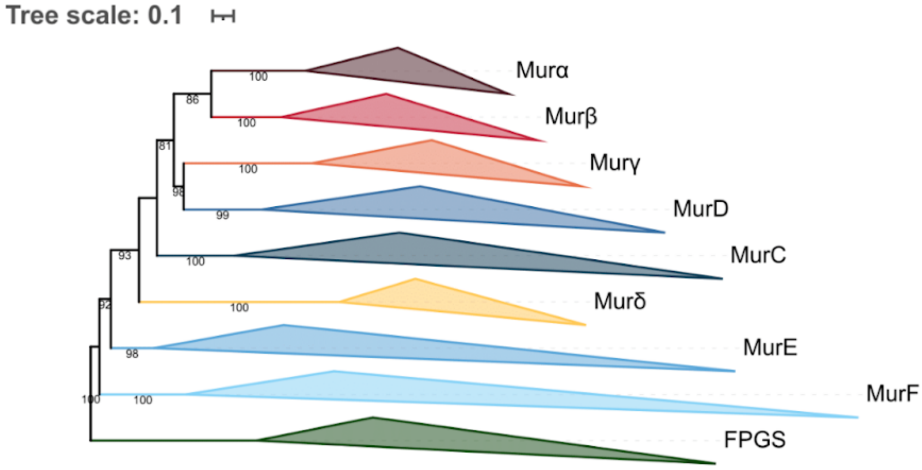


Figure S27. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. The tree was inferred from a matrix of **3,046 sequences x 543 unambiguously aligned AAs** using IQ-TREE **under the C20+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

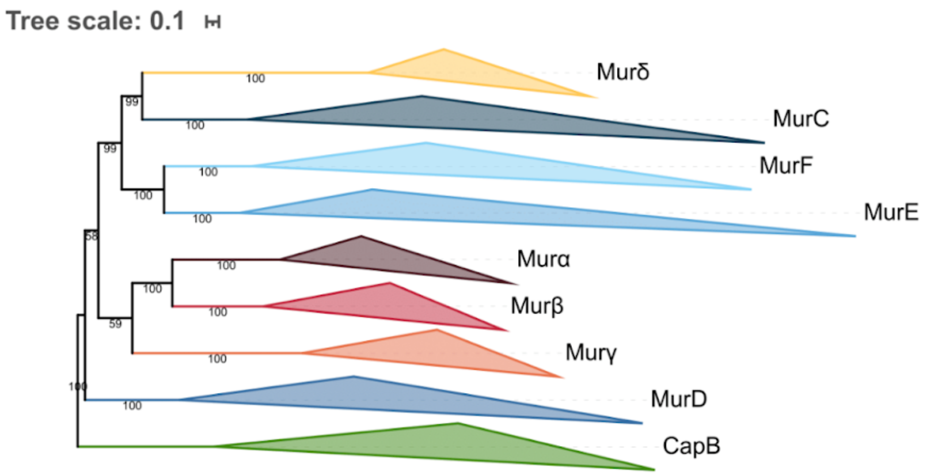


Figure S28. Phylogenetic tree of the Mur domain-containing family rooted on CapB. The tree was inferred from a matrix of **2519 sequences x 532 unambiguously aligned AAs** using IQ-TREE **under the PMSF LG+C60+G4 model** with 3000 iterations. The guide tree was the C40+G4 tree of Figure S22. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

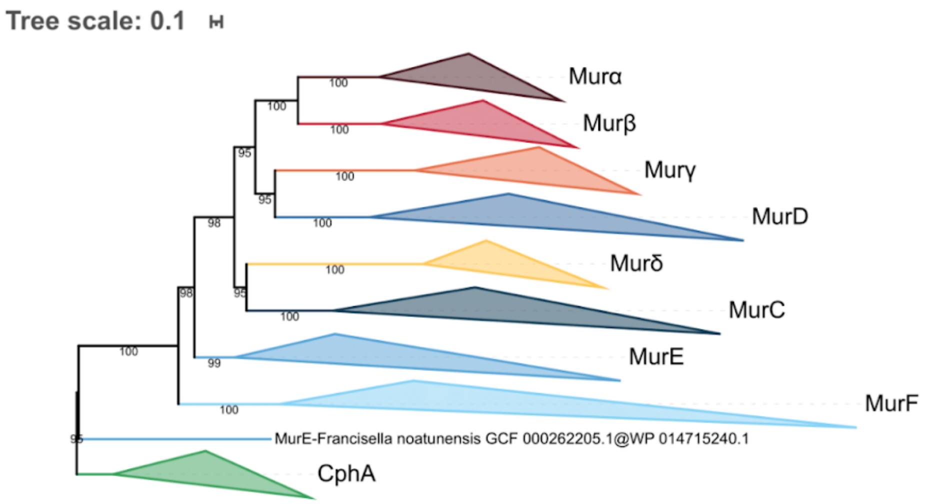


Figure S29. Phylogenetic tree of the Mur domain-containing family rooted on CphA. The tree was inferred from a matrix of **2461 sequences x 539 unambiguously aligned AAs** using IQ-TREE **under the PMSF LG+C60+G4 model** with 3000 iterations. The guide tree was the C40+G4 tree of Figure S25. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

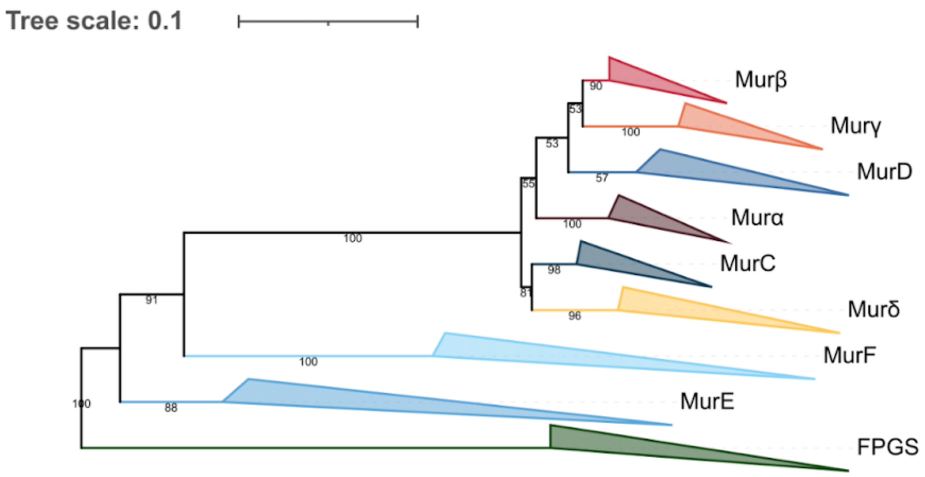


Figure S30. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. Indels tree inferred from a matrix of **3001 sequences x 1799 unambiguously aligned AAs** using RAxML **under the BINGAMMAX model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

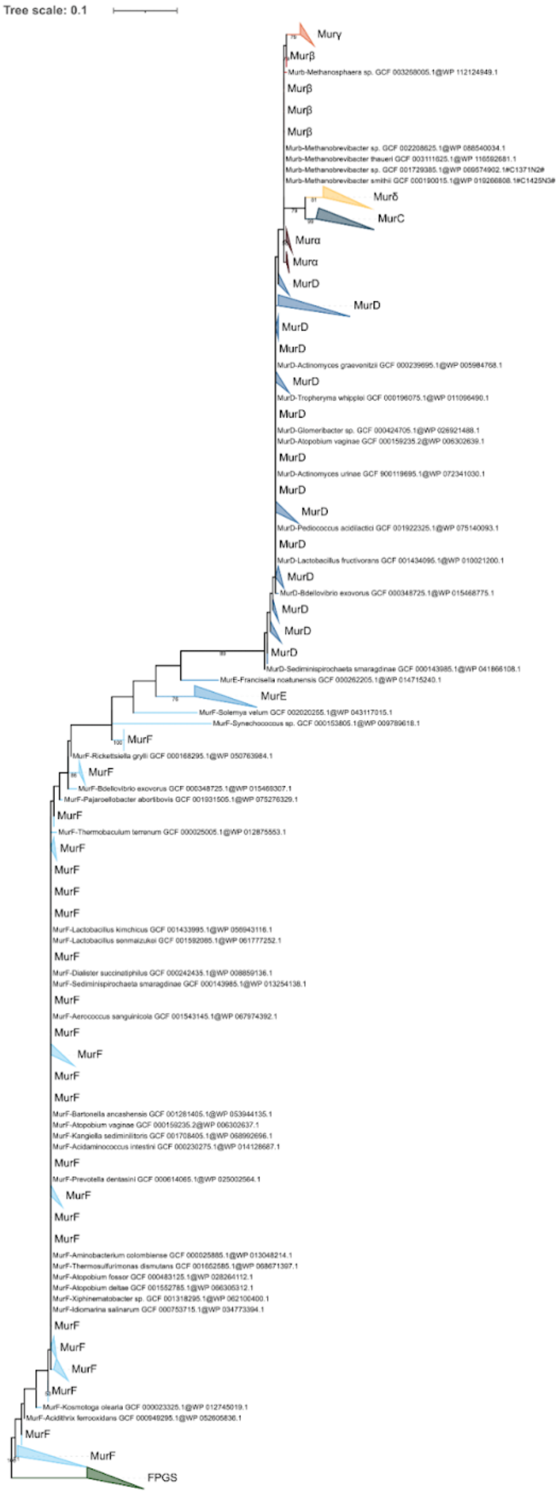

Tree scale: 0.1
76
Murγ
Murβ
Murb-Methanosphaera sp. GCF 003268005.1@WP 112124949.1
Murβ
Murβ
Murβ
Murb-Methanobrevibacter sp. GCF 002208625.1@WP 088540034.1
Murb-Methanobrevibacter thaueri GCF 003111625.1@WP 116592681.1
Murb-Methanobrevibacter sp. GCF 001729385.1@WP 069574902.1#C1371N2#
Murb-Methanobrevibacter smithii GCF 000190015.1@WP 019266808.1#C1425N3#
81
Murδ
79
99
MurC
Murα
Murα
MurD
MurD
MurD
MurD
MurD-Actinomyces graevenitzii GCF 000239695.1@WP 005984768.1
MurD
MurD-Tropheryma whipplei GCF 000196075.1@WP 011096490.1
MurD
MurD-Glomeribacter sp. GCF 000424705.1@WP 026921488.1
MurD-Atopobium vaginae GCF 000159235.2@WP 006302639.1
MurD
MurD-Actinomyces urinae GCF 900119695.1@WP 072341030.1
MurD
MurD
MurD-Pediococcus acidilactici GCF 001922325.1@WP 075140093.1
MurD
MurD-Lactobacillus fructivorans GCF 001434095.1@WP 010021200.1
MurD
MurD-Bdellovibrio exovorus GCF 000348725.1@WP 015468775.1
MurD
MurD
89
MurD
MurD-Sediminispirochaeta smaragdinae GCF 000143985.1@WP 041866108.1
MurE-Francisella noatunensis GCF 000262205.1@WP 014715240.1
76
MurE
MurF-Solemya velum GCF 002020255.1@WP 043117015.1
MurF-Synechococcus sp. GCF 000153805.1@WP 009789618.1
100
MurF
MurF-Rickettsiella grylli GCF 000168295.1@WP 050763984.1
86
MurF
MurF-Bdellovibrio exovorus GCF 000348725.1@WP 015468307.1
MurF-Pajaroellobacter abortibovis GCF 001931505.1@WP 075276329.1
MurF
MurF-Thermobaculum terrenum GCF 000025005.1@WP 012875553.1
MurF
MurF
MurF
MurF
MurF-Lactobacillus kimchicus GCF 001433995.1@WP 056943116.1
MurF-Lactobacillus senmaizukei GCF 001592085.1@WP 061777252.1
MurF
MurF-Dialister succinatiphilus GCF 000242435.1@WP 008859136.1
MurF-Sediminispirochaeta smaragdinae GCF 000143985.1@WP 013254138.1
MurF
MurF-Aerococcus sanguinicola GCF 001543145.1@WP 067974392.1
MurF
MurF
MurF
MurF
MurF-Bartonella ancashensis GCF 001281405.1@WP 053944135.1
MurF-Atopobium vaginae GCF 000159235.2@WP 006302637.1
MurF-Kangiella sediminilitoris GCF 001708405.1@WP 068992696.1
MurF-Acidaminococcus intestini GCF 000230275.1@WP 014128687.1
MurF
MurF-Prevotella dentasini GCF 000614065.1@WP 025002564.1
MurF
MurF
MurF
MurF-Aminobacterium colombiense GCF 000025885.1@WP 013048214.1
MurF-Thermosulfurimonas dismutans GCF 001652585.1@WP 068671397.1
MurF-Atopobium fossor GCF 000483125.1@WP 028264112.1
MurF-Atopobium deltae GCF 001552785.1@WP 066305312.1
MurF-Xiphinematobacter sp. GCF 001318295.1@WP 062100400.1
MurF-Idiomarina salinarum GCF 000753715.1@WP 034773394.1
MurF
MurF
MurF
MurF
MurF-Kosmotoga olearia GCF 000023325.1@WP 012745019.1
MurF-Acidithrix ferrooxidans GCF 000949295.1@WP 052605836.1
MurF
MurF
FPGS

Figure S31. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. Indels tree inferred from a matrix of **3004 sequences x 281 unambiguously aligned AAs** using RAxML **under the BINGAMMAX model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

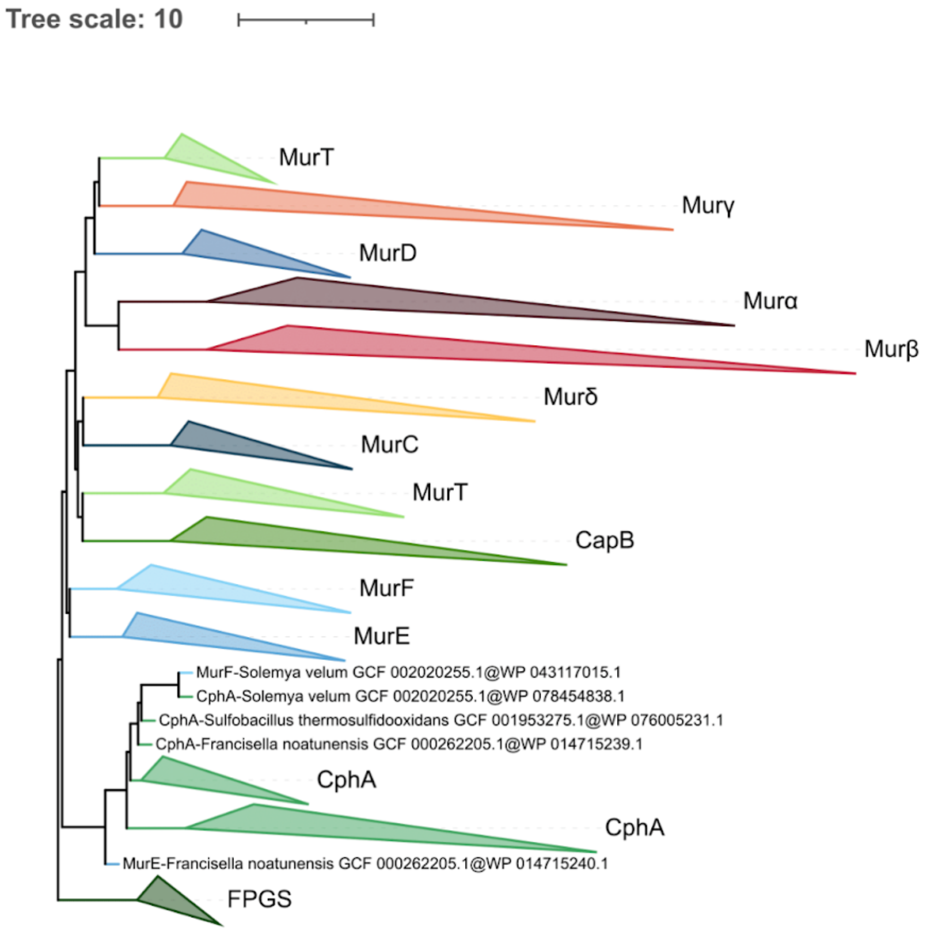

Tree scale: 10
MurT
Murγ
MurD
Murα
Murβ
Murδ
MurC
MurT
CapB
MurF
MurE
MurF-Solemya velum GCF 002020255.1@WP 043117015.1
CphA-Solemya velum GCF 002020255.1@WP 078454838.1
CphA-Sulfobacillus thermosulfidooxidans GCF 001953275.1@WP 076005231.1
CphA-Francisella noatunensis GCF 000262205.1@WP 014715239.1
CphA
CphA
MurE-Francisella noatunensis GCF 000262205.1@WP 014715240.1
FPGS

Figure S32. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. ASTRAL tree inferred computed **from the 1000 LG4X+R4 species resampling trees**. Tree visualization was performed using iTOL. Branches were collapsed on homogeneous sequence annotation.

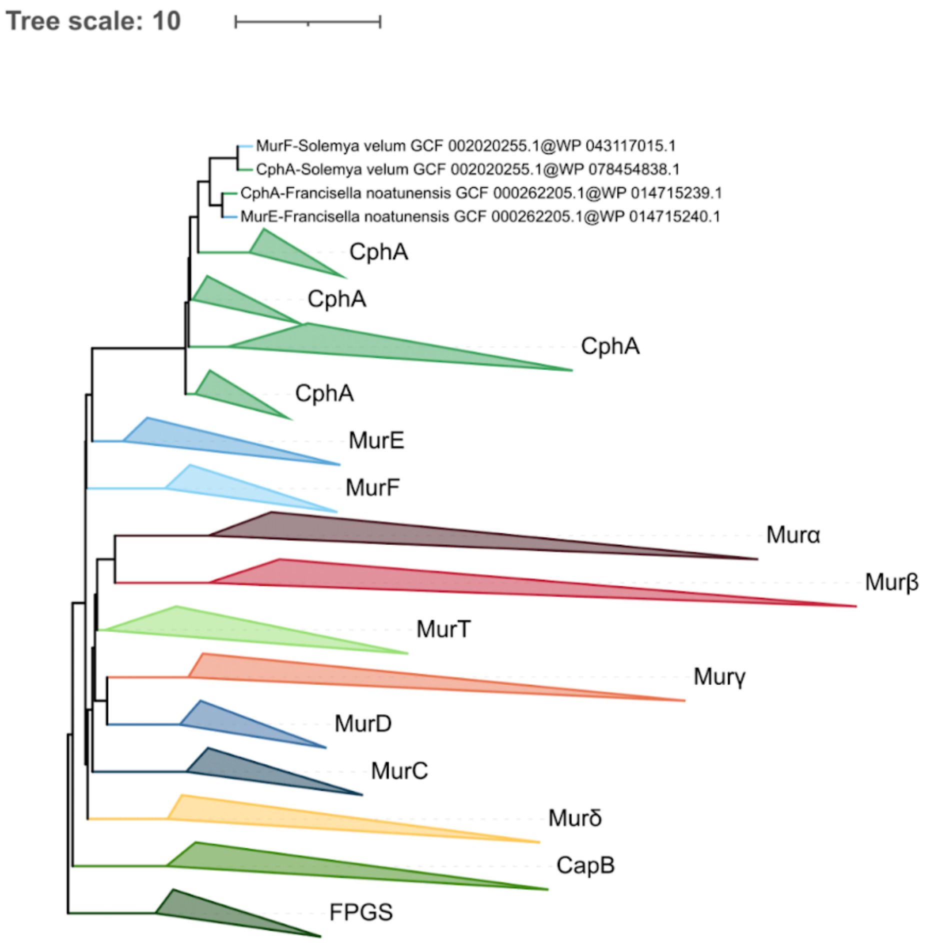

Tree scale: 10
MurF-Solemya velum GCF 002020255.1@WP 043117015.1
CphA-Solemya velum GCF 002020255.1@WP 078454838.1
CphA-Francisella noatunensis GCF 000262205.1@WP 014715239.1
MurE-Francisella noatunensis GCF 000262205.1@WP 014715240.1
CphA
CphA
CphA
CphA
MurE
MurF
Murα
Murβ
MurT
Murγ
MurD
MurC
Murδ
CapB
FPGS

Figure S33. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. ASTRAL tree computed **from the 1000 C20+G4 species resampling trees**. Tree visualization was performed using iTOL. Branches were collapsed on homogeneous sequence annotation.

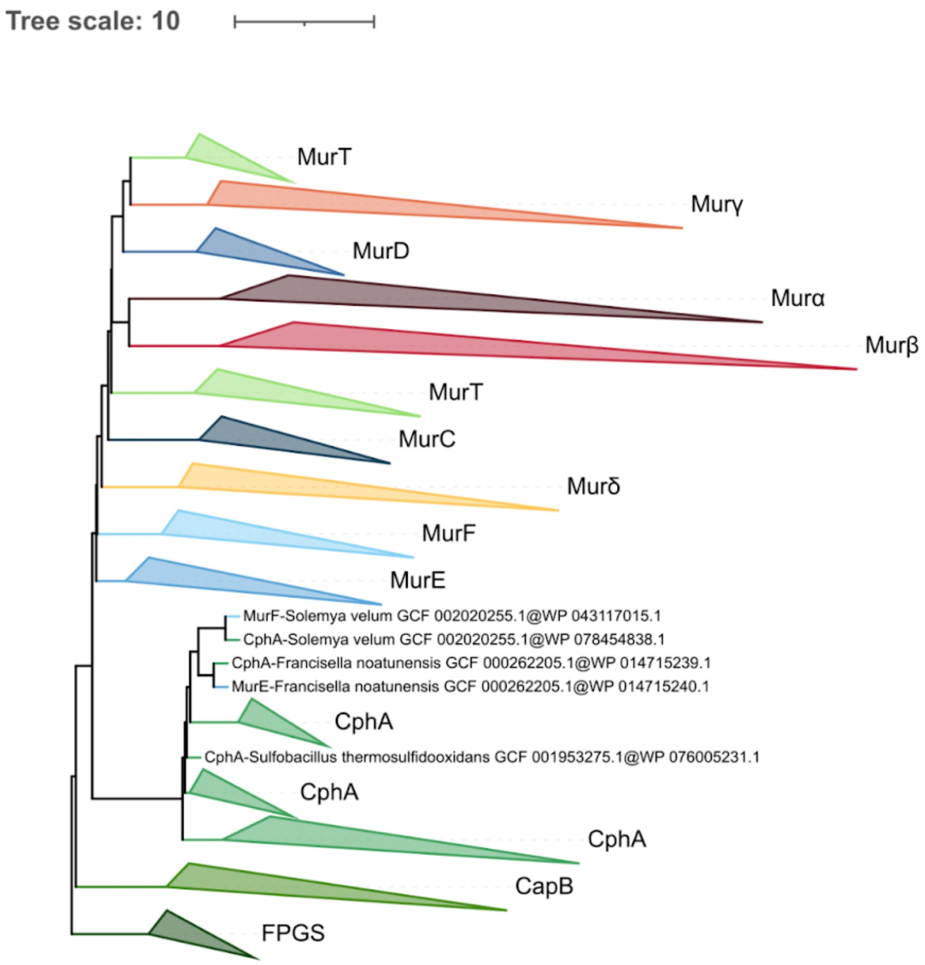

Tree scale: 10
MurT
Murγ
MurD
Murα
Murβ
MurT
MurC
Murδ
MurF
MurE
MurF-Solemya velum GCF 002020255.1@WP 043117015.1
CphA-Solemya velum GCF 002020255.1@WP 078454838.1
CphA-Francisella noatunensis GCF 000262205.1@WP 014715239.1
MurE-Francisella noatunensis GCF 000262205.1@WP 014715240.1
CphA
CphA-Sulfobacillus thermosulfidooxidans GCF 001953275.1@WP 076005231.1
CphA
CphA
CapB
FPGS

Figure S34. Phylogenetic tree of the Mur domain-containing family rooted on FPGS. ASTRAL tree computed **from the 1000 C40+G4 species resampling trees**. Tree visualization was performed using iTOL. Branches were collapsed on homogeneous sequence annotation.

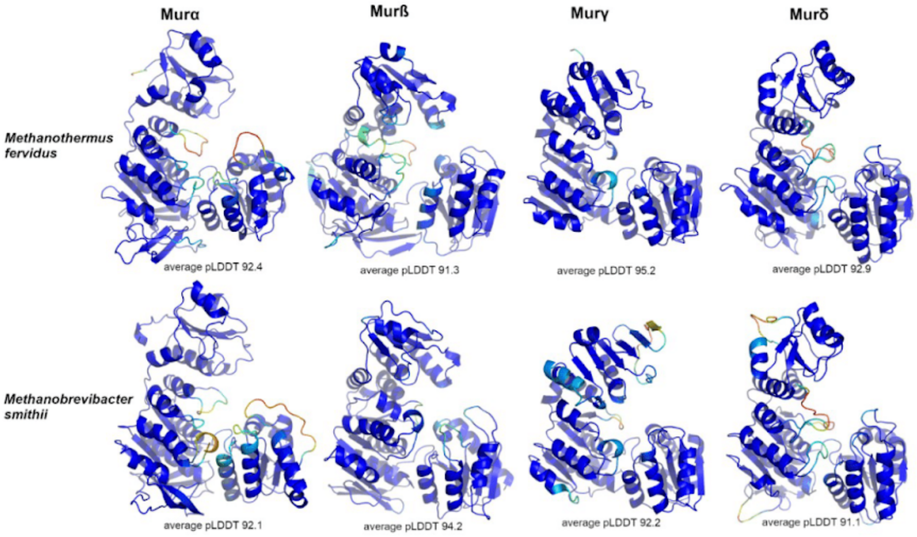


Figure S35. Cartoon representation of the AlphaFold models of the Murαßγδ enzymes from *M. fervidus* and *M. smithii*. Amino acids are colored according to their pLDDT value (from red <50 to blue >90). The average pLDDT value calculated for the entire protein is indicated below each model. Because our analysis is centered on the C-terminal domain, we selected the model with the highest average pLDDT for this specific domain among the five models generated by AlphaFold. For Murα and Murß from *M. fervidus* and Murδ from *M. smithii*., these models had a slightly lower total average pLDDT (less than 1%) compared to the overall best.

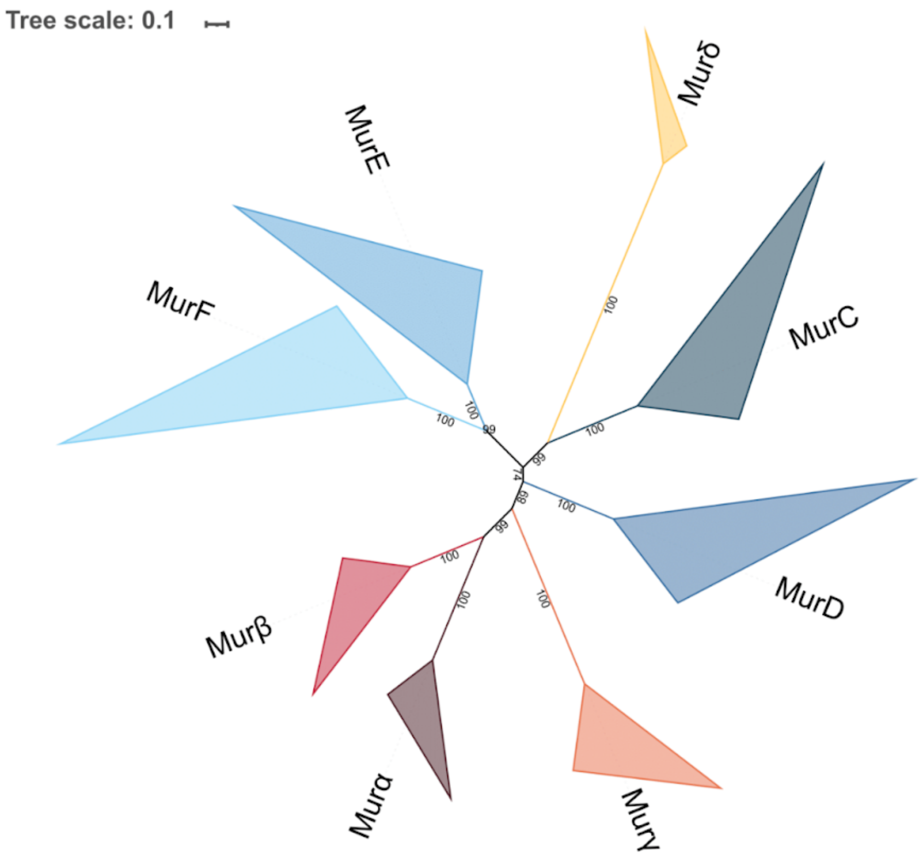


Figure S36. Unrooted phylogenetic tree of the Mur domain-containing family. The tree was inferred from a matrix of **2432 sequences x 528 unambiguously aligned AAs** using IQ-TREE **under the LG4X+R4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

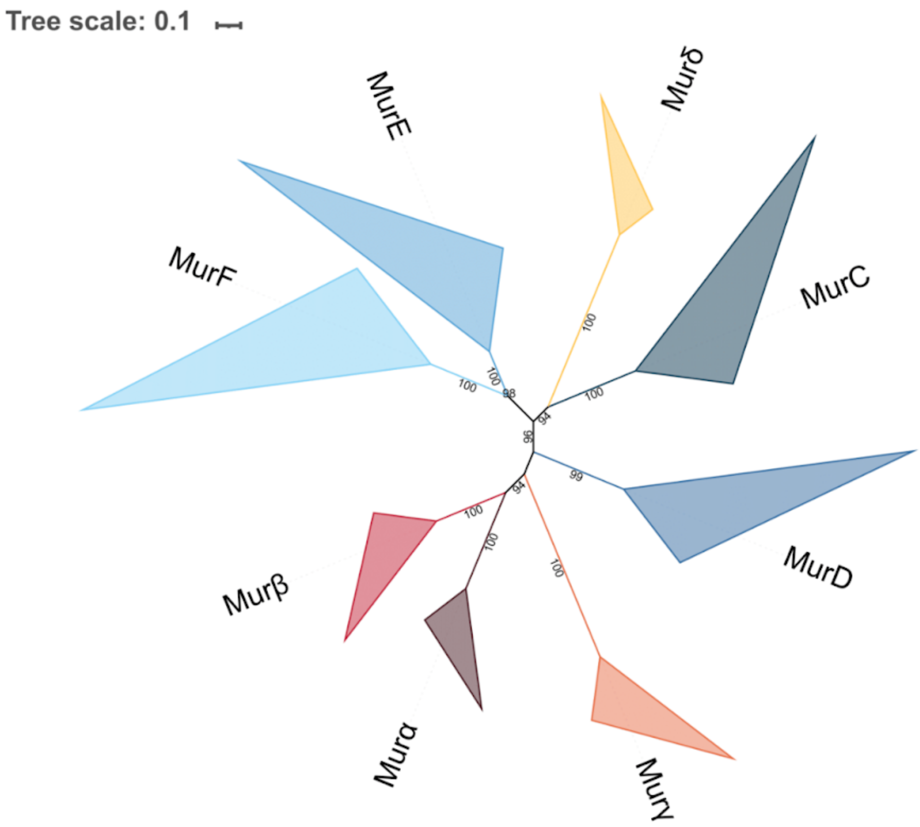


Figure S37. Unrooted phylogenetic tree of the Mur domain-containing family. The tree was inferred from a matrix of **2432 sequences x 528 unambiguously aligned AAs** using IQ-TREE **under the C20+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

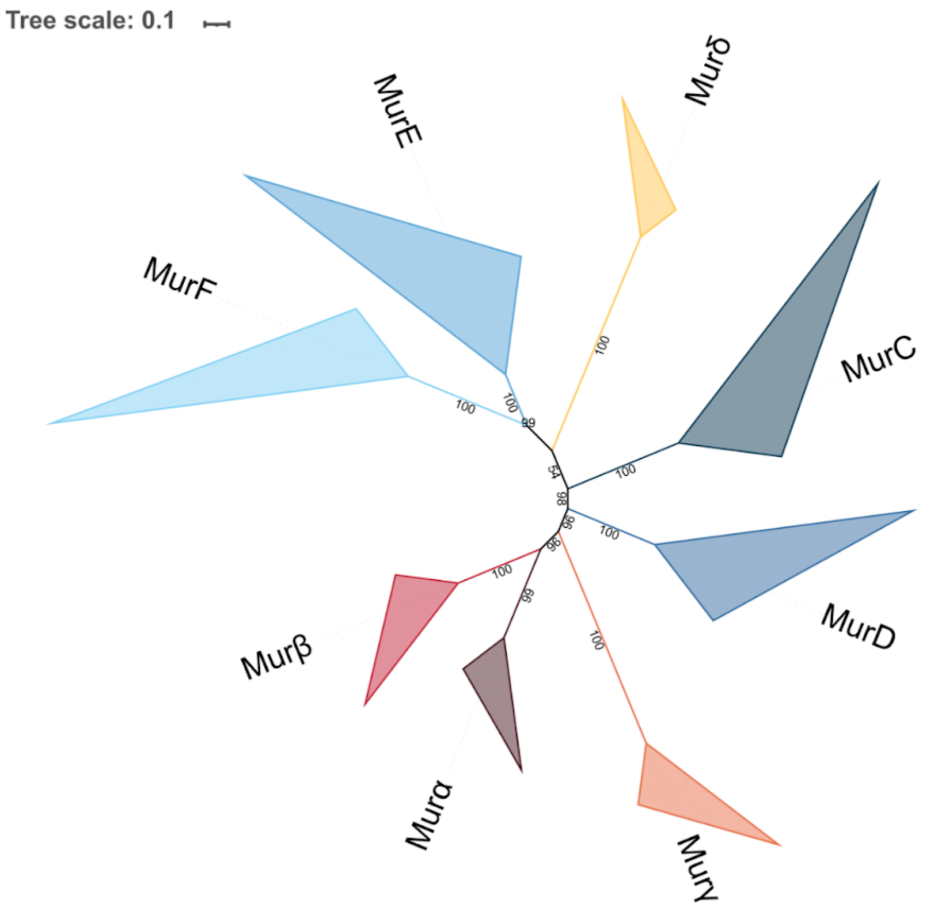


Figure S38. Unrooted phylogenetic tree of the Mur domain-containing family. The tree was inferred from a matrix of 2**432 sequences x 528 unambiguously aligned AAs** using IQ-TREE **under the C40+G4 model**. Tree visualization was performed using iTOL. Bootstrap values are shown if greater or equal to 50. Branches were collapsed on homogeneous sequence annotation.

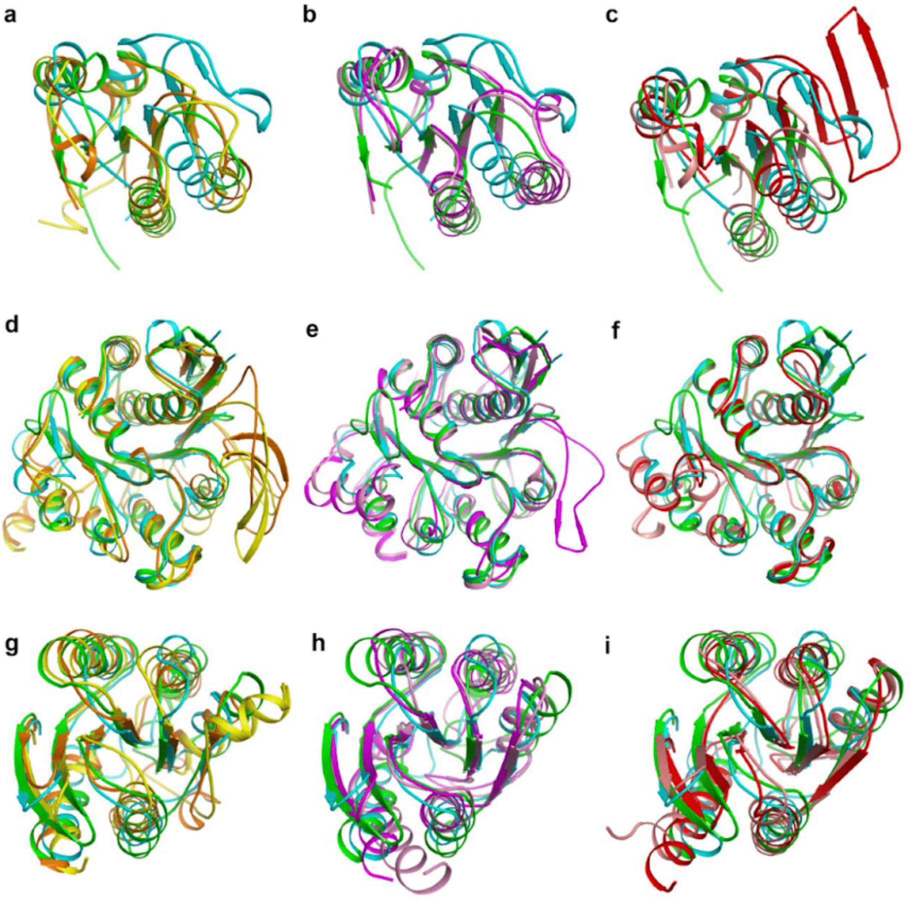


Figure S39. Superimposition of the three domains of MurC, MurD, Murα, Murß and Murγ. (**a**) Superimposition of the N-terminal domain of MurC from *Haemophilus influenzae* (PDB code 1P3D; green), MurD from *E. coli* (PDB code 4UAG; cyan), and Murα from *M. fervidus* (orange) and *M. smithii* (yellow) (**b**) same as (a) with Murß from *M. fervidus* (magenta) and *M. smithii* (pink) (**c**) same as (a) with Murγ from *M. fervidus* (salmon) and *M. smithii* (red). (**d, e, f**) Same as (a, b, c) for the middle domain. (**g, h, i**) Same as (a, b, c) for the C-terminal domain.

3. Discussion

3.1. On sequence mining of public databases

3.1.1. Contamination of public databases

Although RefSeq hosts high-quality genomes compared to GenBank, some issues of RefSeq still need to be addressed, notably regarding contamination. In Chapter 1 of the Results section (Lupo et al. 2021), we have demonstrated the presence of mis-affiliated genomes in NCBI RefSeq. This is problematic for comparative genomic or phylogenomic studies, because researchers often download genomes from public repositories based on their assigned taxonomy. Obviously, inclusion of unwanted organisms can lead to aberrant results and wrong conclusions (Cornet and Baurain 2022). Theoretically, Physeter is the only contamination detection tool able to detect mis-affiliated genomes by comparing the main detected organism and the associated taxonomy from NCBI (Schoch et al. 2020) or GTDB (Parks et al. 2018). However, it is difficult to distinguish between mis-affiliated genomes and these two cases: 1) the expected taxon is very scarce due to the heavy contamination of the genome and, 2) when assessing contamination of rare genomes with no close representative in reference databases (Cornet and Baurain 2022). Interestingly, these two cases are the same that cause the singletons to appear in genome deduplication (see section 3.1.4.). Furthermore, RefSeq is often used as a reference database for database-dependent tools designed for the detection of genomic contamination. Presence of issues such as mis-affiliated or contaminated genomes in those reference databases can lead to false-positive or false-negative results (Cornet and Baurain 2022). Physeter minimizes the effects of contaminated genomes in reference databases by using a leave-one-out approach (Lupo et al. 2021). Nevertheless, this strategy is successful only if the database is rich enough to maintain its taxonomic diversity during the leave-one-out step (Cornet and Baurain 2022). Altogether, these issues show the need to maintain high-quality and contamination-free reference databases.

3.1.2. Horizontal gene transfer and contamination

The distinction between contamination and horizontal gene transfer (HGT) events is a challenging issue that needs to be addressed in the future. HGTs have played an important role during the evolution of prokaryotes. Indeed, it has been shown that the majority of bacterial genes have been transferred at least once across organisms (Dagan and Martin 2007; Dagan et al. 2008). Moreover, HGT events have been reported in gut microbiota and thus affect metagenomic samples (Eme et al. 2017; Frazão et al. 2019). Interestingly, HGT involving bacteria and eukaryotes have also been reported (Keeling and Palmer 2008; Schmitt and Lumbsch 2009; McDonald et al. 2012; Soucy et al. 2015; Kominek et al. 2019; Yubuki et al. 2020), although some of them turned out to be the result of genomic contamination, such as in the tardigrade (Arakawa 2016; Delmont and Eren 2016; Koutsovoulos et al. 2016) or human genome (Salzberg 2017). This can be explained by the contamination background noise, which affects the identification of HGT events. Oppositely, "genuine" HGT can complicate the detection of contaminants. Consequently, a foreign sequence within an organism is so far considered either as a contaminant or the product of a HGT event, depending on the purpose of the software used, whereas it could be genuinely one or the other or even both at the same time (Cornet and Baurain 2022).

3.1.3. Facing sequence suppression in public databases

NCBI sequence databases receive a tremendous amount of data from direct submission by individual laboratories or batch submission by high-throughput sequencing centers (Bouadjenek et al. 2017; Sayers et al. 2022). Moreover, international collaborations, such as the INSDC (Arita et al. 2021), increase the number of the available sequences in those databases. To provide access to newly added sequences to the worldwide community, the NCBI regularly performs new releases of its databases. For instance, GenBank and RefSeq database releases occur every two months (https://www.ncbi.nlm.nih.gov/genbank/ Accessed 30 November 2022; https://ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/archive/ Accessed 30 November 2022). Each new release is accompanied by a list of nucleotide or protein sequences that have been removed from the corresponding database between two releases (https://ftp.ncbi.nlm.nih.gov/refseq/README Accessed 30 November 2022). A sequence is manually suppressed when curators, submitters or a third party report an error. It has been estimated that it takes on average 1 month between the submission and the deletion of a suspicious sequence

(Bouadjenek et al. 2017). The suppression of incorrect or dubious sequences is a critical point for the quality and accuracy of public databases. However, these suppressions of sequences are problematic for studies spread over several years. During the study of the class D beta-lactamases (see Chapter 2 of the Results section), we have faced sequence suppression (Lupo et al. 2022). Indeed, from the 3510 unique OXA-domain family sequences that we had identified, 763 are now tagged as “Removed record” from the NCBI RefSeq database. More surprisingly, 408 of the 763 removed sequences are representative sequences used to compute the phylogenetic tree. The suppressed sequences are not problematic for reproducibility studies, since removed records are still accessible through old release versions of the public databases (https://ftp.ncbi.nlm.nih.gov/refseq/removed/ Accessed 30 November 2022). However, suppressions are questionable regarding the interpretation of the results, particularly for the phylogenetic tree, in which about 30% of the sequences are probably incorrect (or at least not genuine in terms of organism source). Consequently, during the process of sequence selection for wet lab validation, we had to ensure the presence of the ten selected protein sequences in the last RefSeq database.

3.1.4. Divergent sequences and rare genomes

The 1413 representative OXA-domain family sequences have been selected by the clustering program CD-HIT (Fu et al. 2012) from the 3510 unique sequences. In addition to improving the performance of subsequent analyses (Fu et al. 2012), sequence deduplication can mechanically bring out sequences that are more divergent (i.e., singletons) than other sequences of a dataset (Evans and Denef 2020). The difference between divergent sequences and other sequences can be real and due to, e.g., fast evolution, or due to artifacts (e.g., sequencing errors or mispredictions of coding regions) (Di Franco et al. 2019), like our potentially incorrect OXA-domain family removed sequences. Similar observations have been done for genome assembly. Indeed, an aggressive deduplication of bacterial genomes showed that a genome can be considered as a singleton for two reasons: 1) the genome is truly different and forms a singleton cluster during deduplication (Léonard et al. 2021; Léonard 2021), or 2) the genome is so heavily contaminated (i.e., chimeric) that this makes it look very different from the other, genuinely related, genomes (Cornet and Baurain 2022). However, it is very difficult to assess the contamination level for a rare genome when there is no close representative in reference databases (Cornet and Baurain 2022). In Chapter 3 of the Results section, we have discussed the phylogeny of the gene *murT*. In this phylogenetic tree, we have observed that MurT, along with its partner GadT (Münch et al. 2012; Nöldeke et al. 2018), is found in some bacterial lineages and exclusively in Methanobacteriales. In contrast, an homologue of MurT, termed MurT-like, is only found in Methanobacteriales and also in the single Methanopyrales of our dataset. However, to assert that MurT is really absent from all Methanopyrales is adventurous at best, because only one Methanopyrales genome (i.e., *Methanopyrus sp.* KLO6) was available in RefSeq at the beginning of this study. Thus, this absence of the gene *murT* could instead result from artifacts affecting only this specific strain. To confirm or refute our observation, we have used the Forty-Two software package (Irisarri et al. 2017; Simion et al. 2017) to mine potential MurT homologues in two additional *Methanopyrus* genomes from RefSeq and one from GenBank. This control analysis confirmed the absence of MurT and the presence of MurT-like in all four *Methanopyrus* species. This also showed that GenBank is useful to validate observations made for rare genomes of RefSeq, not by using the entire genomes due to their lower quality compared to RefSeq but by mining individual sequences.

3.1.5. Public databases as a means to preserve biological diversity

Despite the aforementioned limitations, public databases offer researchers the ability to get their hand on large amounts of sequences and genomes. Such dataset collections also allow to preserve biological diversity (Mahilum-Tapy 2009), even more than microorganism collections, where contaminated or unpreserved cultures can be definitely lost (Sharma et al. 2019). Owing to these data, researchers can directly identify interesting sequences during genome mining analyses (Albarano et al. 2020), clone or synthesize the corresponding genes into vectors and express them in heterologous competent cells. In this respect, our study of class D beta-lactamases perfectly illustrates a complete analysis, going from bioinformatic analyses of public sequence data to wet lab validation of candidate protein sequences. Such an analytical workflow can certainly be applied to other protein families. In addition, those public sequence data can be used in synthetic biology, a field of biology that attempts to create new living organisms from synthetic components (Benner and Sismour 2005). A few application examples of synthetic biology are given in the review of Venter et al. 2022, among which the use of synthetic genomics to recover non-cultivable viruses. This strategy was applied in the recent pandemic of SARS-CoV-2 (i.e., a single-stranded RNA virus), where commercially synthesized DNA were reassembled in the yeast *Saccharomyces cerevisiae* and then transcripted into infectious RNA to rescue viable viruses (Thi Nhu Thao et al. 2020).

3.2. Cell-wall polymers, antibiotics and cell division

3.2.1. Other functions of the peptidoglycan

The peptidoglycan (PG) is a ubiquitous polymer found in the cell wall of almost all bacterial species (Pazos and Peters 2019). Most of the genes involved in PG biosynthesis lie in the division and cell-wall synthesis (*dcw*) cluster, of which the gene order and composition are relatively well conserved across the different bacterial lineages (Tamames 2001; Mingorance and Tamames 2004; Real and Henriques 2006). Recently, it has been shown that the last bacterial common ancestor (LBCA) was already a complex organism, with PG and a complete *dcw* cluster consisting of 17 genes: *mraZ*, *mraW*, *ftsL*, *ftsI*, *murE*, *murF*, *mraY*, *murD*, *ftsW*, *murG*, *murC*, *murB*, *murA*, *ddlB*, *fts*Q, *ftsA* and *ftsZ* (Léonard et al. 2022). In addition to acting as a protective layer, PG has an important role in bacterial division, which is initiated by the formation of the FtsZ (i.e., a cytoskeletal protein homologue to the eukaryotic tubulin) septal ring (McQuillen and Xiao 2020). Indeed, even species from the PVC group (e.g., Planctomycetes, Chlamydiae) once thought to lack PG, actually exhibit a thin layer of PG, which is notably synthesized during septal division (Liechti et al. 2014; Jeske et al. 2015; Packiam et al. 2015; van Teeseling et al. 2015; Liechti et al. 2016). Only Mollicutes, which are strict intracellular parasites including *Mycoplasma* spp, are the only known bacteria to lack a cell wall (Trachtenberg 1998). These bacteria have a reduced *dcw* cluster that is limited to a maximum of four genes: *mraZ*, *mraW*, *ftsA* (i.e., cytoskeletal protein homologue to the eukaryotic actin) and *ftsZ* (Martínez-Torró et al. 2021; Léonard et al. 2022). The mechanism of division of these cell-wall-less bacteria is poorly understood, but some results suggest that FtsZ plays an important role in their cell division (Martínez-Torró et al. 2021). Interestingly, there exist many bacteria able to switch into a cell-wall-less state named “L-form” (Allan et al. 2009). This state can be achieved through genetic mutation or chemical inhibition of the enzymes involved in PG biosynthesis. Without a protective layer of PG, L-form bacteria require an osmoprotective growth medium for survival (Osawa and Erickson 2019). Despite having no cell wall, L-form bacteria still have the ability to grow and proliferate (Errington 2017). It was shown that L-form *Escherichia coli* cells do not completely lack PG but have 7% of the normal amount, which allowed them to perform septal division (Joseleau-Petit et al. 2007). However, a L-form of *Bacillus subtilis* with no cell wall and knocked-down for the *ftsZ* gene proliferates in a strange manner, which does not follow the classical binary fission. Hence, the cell grows until the cell membrane forms a protrusion, followed by the eruption of multiple progeny (Leaver et al. 2009; Errington 2017). Interestingly, even though Chlamydia species use PG for division, they lack the two cytoskeletal proteins FtsZ and FtsA. Thus, it was proposed that another cytoskeletal protein, MreB, acts as a substitute for FtsZ (Ouellette et al. 2020). This latter protein has also been proposed as the substitute for FtsZ in the L-form *B. subtilis* mutant (Leaver et al. 2009). Although-cell wall-less bacteria can still divide, the small amount of PG found in PVC group species, notably during division suggests that PG is required for the classical septal division, while the function of FtsZ can be fulfilled by other cytoskeletal proteins. Therefore, it looks like PG has appeared so early in bacterial evolution that it is indissociable from the division complex (Pende et al. 2021). Thus, beyond its protective role the other function of PG is to scaffold the cell-division machinery.

3.2.2. Environmental bacteria as a reservoir of antimicrobial resistance genes

The bacterial cell wall is so important for cell survival that it is the target of many antimicrobial compounds, such as beta-lactam antibiotics (Bhattacharjee 2016). For protection against those antibiotics, bacteria have developed different strategies of resistance, notably the production of specific hydrolases termed beta-lactamases (Bush 2018). In our study of class D beta-lactamases, we have emphasized the idea that environmental bacteria that have never been exposed to the pressure of human and veterinary antibiotic therapy can constitute a reservoir of new beta-lactamases. In natural habitats, microorganisms live in communities as a multi-cellular network. When ressources begin to deplete, organisms start to secrete secondary metabolites, of which antibiotics, that give a competitive advantage to producers. In response, non-producer organisms developed antimicrobial resistance (AMR) mechanisms, and it has been documented that antibiotics could also act as signaling molecules between organisms of a microbial community (Sengupta et al. 2013). Although there are still a lot of new antimicrobial compounds to be isolated from naturally producing organisms (Adam et al. 2018), we can speculate that AMR mechanisms already exist for these compounds that are still to be discovered, due to the evolutionary arms race between producer and non-producer organisms. We have to be mindful that even if new antimicrobial compounds are released for human health therapy, an overuse of these antibiotics could increase the spread of the associated AMR.

3.2.3. Antimicrobial resistance in Archaea

In 2020, Diene et al. (Diene et al. 2020) have identified in archaeal genomes genes coding for putative class B and class C beta-lactamases. They cloned and expressed two sequences of *Methanosarcina* into *E. coli*, and showed that both sequences exhibit a weak beta-lactamase activity. Since no PG is found in archaea, and thus no penicillin binding proteins (PBPs) either, we can question the use for such beta-lactamases in archaeal species. Two hypotheses might explain the presence of beta-lactamases in some archaea. First, we can speculate that those archaeal enzymes are actually not beta-lactamases but far related archaeal homologues with a hydrolase activity profile similar to bacterial beta-lactamases but playing a different role in archaea (Colson et al. 2020). Second, these enzymes would have been acquired through HGT from bacteria. Consequently, the AMR genes were only maintained in archaea, probably because the corresponding antibiotics can target an undetermined transpeptidase in the archeal cell. Yet, detecting AMR gene in archaeal species is not problematic, at least so far. Indeed, no pathogenic archaea have been identified to date, even though archaea do possess characteristic of pathogens (e.g., interaction with eukaryotic cells, production of toxins) (Gill and Brinkman 2011). However, *Methanobrevibacter smithii*, a commensal methanogenic archaeon notably found in human gut (Borrel et al. 2020), is often co-cultured along with pathogenic bacteria during different infections (Collin et al. 2011; Grine, Lotte, et al. 2019; Grine, Drouet, et al. 2019; Djemai et al. 2021; Rasmussen and Collin 2021). Thus, we should reconsidered these methanogenic archaea as opportunistic pathogens (Hassani et al. 2020), and therefore search for anti-archaeal compounds, since several antibiotics (e.g., beta-lactams, glycopeptides) that target bacteria are not effective against archaea (Khelaifia and Drancourt 2012). Nevertheless, *M. smithii* and other pseudomurein-containing archaea are sensitive to two endoisopeptidases isolated from phages, PeiW and PeiP. Both enzymes cleave the isopeptide bond Ala-ε-Lys in the stem peptide of the pseudomurein (PM), which leads to the lysis of the cell (Schofield et al. 2015). Such enzymes might thus be considered as archaeal antibiotics in case of future need.

3.2.4. Advantages to acquire pseudomurein cell-wall

Unlike bacteria, there is not a cell wall or a cell wall polymer that is ubiquitous in archaea. Yet, the S-layer is the most commonly observed cell wall in the different archaeal lineages. Thus, it was suggested that S-layer was the most ancient archaeal cell wall structure to evolve (Klingl et al. 2019; Meyer and Albers 2020). However, a recent study suggests multiple independent evolutionary origins for the S-layer (Bharat et al. 2021). Beside this S-layer cell wall, some lineages of Euryarchaeota are characterized by the presence of a cell wall polymer: Methanopyrales and Methanobacteriales have PM, Methanosarcina have methanochondroitin, whereas Halobacteria have halomucin, glutaminylglycan or heteropolysaccharides. This taxonomic distribution suggests that cell wall polymers have appeared independently during archaeal evolution. Based on these observations, we speculate that the last archaeal common ancestor (LACA) had no cell wall. Knowing that some modern euryarchaeotal species do survive in extreme environments without any cell wall, what would be the advantages of acquiring a cell-wall polymer other than acting as a simple protective layer? Indeed, some cell-wall-less Thermoplasmata species (e.g., *Thermoplasma*, *Ferroplasma*) can thrive in thermophilic (around 60°C) and acidophilic (pH ≤ 2) environments

(Golyshina and Timmis 2005; Reysenbach 2015). Instead of a bilayered membrane, these euryarchaeota have a monolayered membrane to which are attached glycoproteins and lipoglycans, which confer them a resistance to hydrolysis (Klingl et al. 2019). This cell-wall-less state is analogous to Mollicutes in Bacteria. However, as the LBCA probably already had PG and Mollicutes have a reduced *dcw* cluster, the absence of a cell wall in Mollicutes is not an ancestral state and must be the result of a secondary simplification due to its parasitic lifestyle (Martínez-Torró et al. 2021). Above, we have argued that PG in Bacteria seems to be mandatory for the classical septal division. Therefore, does the acquisition of a cell-wall polymer could enhance the FtsZ-dependent septal division in Archaea? Despite a few exceptions, the cell division (cytokinesis) in Bacteria is mainly mediated by the cytoskeletal protein FtsZ. In eukaryotes, the cell division involves actin and proteins of the **E**ndosomal **S**orting **C**omplex **R**equired for **T**ransport (ESCRT), the latter notably mediating membrane abscission (i.e., separation of daughter cells) during cytokinesis. In Archaea, two division mechanisms are found, the FtsZ- and Cdv-based systems. Interestingly, some proteins of the Cdv-based division are homologous to the eukaryotic ESCRT-III sorting complex and are found in the TACK superphylum and the Asgard (Lindås et al. 2008; Caspi and Dekker 2018; Pende et al. 2021; Ithurbide et al. 2022). In addition, while most Archaea possess two homologues of FtsZ, namely FtsZ-1 and FtsZ-2, Methanopyrales and Methanobacteriales are the only cell-wall polymer containing organisms to contain only one homologue of FtsZ. Moreover, they are the only archaeal species to possess a MreB homologue, a cytoskeletal protein involved in cell elongation in Bacteria (Pende et al. 2021; Ithurbide et al. 2022). In the majority of bacteria, FtsZ is anchored to the cytoplasmic membrane through FtsA (Mura et al. 2017), whereas SepF is the anchor of FtsZ in Archaea. The SepF protein is also found in some Terrabacteria lineages and has an overlapping role with FtsA. In contrast, FtsA has been identified in Methanopyrales and is completely absent from the other archaeal lineages. Furthermore, evolutionary analyses have shown that both SepF and FtsZ date back to the last universal common ancestor (LUCA) (Pende et al. 2021; Ithurbide et al. 2022). Unfortunately, these observations do not suggest any correlation between the acquisition of a cell-wall polymer and the FtsZ-dependent septal division. Yet, it highlights similar cell-wall-related genetic determinants between PM-containing Archaea and Bacteria. Indeed, both divide through a FtsZ-based system relying on only one copy of the *ftsZ* gene, indicating that PM-containing Archaea have lost one copy of *ftsZ* during their evolution. Moreover, they have a *mreB* gene, which was probably acquired by HGT from Bacteria (Ithurbide et al. 2022). Beside protection and cell division, a third incentive for acquiring a rigid cell-wall polymer could consist in conferring a better resistance to phage entry and exit from the cell (Buchmann and Holmes 2015).

3.2.5. HGT has triggered the evolution of pseudomurein

In Chapter 3 of the Results section, we have discussed HGTs of some genes involved in PG biosynthesis that occurred between Bacteria to a common ancestor of Methanopyrales and Methanobacteriales. Moreover, recent studies have shown that Bacteria and PM-containing Archaea have a similar set of cytoskeletal proteins, of which at least two (i.e., MreB, FtsA) have been acquired through HGT (Pende et al. 2021; Ithurbide et al. 2022). These results indicate that several HGTs from Bacteria could have driven the evolution of PM in the ancestor of Methanopyrales and Methanobacteriales. Since only a small fraction of the genes involved in PG biosynthesis were transferred, their ancestor had to complete the missing steps for PM biosynthesis by its own set of genes. In relation to this, we do not know whether this organism had a simpler cell-wall polymer prior to the transfers, which then evolved into modern PM following the acquisition of bacterial genes, or whether it used the acquired genes along with repurposed pre-existing genes to synthesize the PM from scratch.

3.2.6. From modern traits to ancestral organisms

When phylogeny is used to retrace ancestral traits based on those of modern organisms (top-down approach), the reconstructed ancestors often exhibit complex features. For example, the LBCA was inferred to already have PG with a complete *dcw* cluster (Léonard et al. 2022) while the last eukaryotic common ancestor (LECA) has been modeled as having a mitochondrion, a nucleus, a complex endoplasmic membrane system etc (Lane 2015). Moreover, all “intermediate” organisms (i.e., organisms exhibiting traits that would reflect the transition from LUCA to LBCA and LECA or from prokaryote to eukaryote) have been lost during evolution (i.e., they were stem groups now extinct). Because of their basal position in the eukaryotic tree, some scientists have thought that Archezoa (i.e., eukaryotes that lack mitochondria) were a missing link of the transition between prokaryotic and eukaryotic state (Cavalier-Smith 1987). However, it was subsequently shown that the basal position of Archezoa was due to the long branch attraction artifact. Thus, this “primitive” state of Archezoa actually results from secondary simplification induced by the environment (Brinkmann and Philippe 2007), like Mollicutes in Bacteria (Martínez-Torró et al. 2021). In addition, the high rate of HGT that occurs between Bacteria and Archaea creates a background noise that complicates the phylogenetic reconstruction of ancestral traits (Dagan and Martin 2007; Dagan et al. 2008). Extinct stem groups, leading to complex LCAs, make the top-down approach uninformative about the order of acquisition of traits of interest and their evolutionary advantage at that time. Another approach, termed bottom-up, is advocated in the book of Nick Lane entitled “The Vital Question: Why is life the way it is?” (Lane 2015), where the author uses knowledge of present-day chemistry to imagine the prebiotic chemistry that has led to LUCA. Although some of his hypotheses could in principle be tested in the laboratory, this approach also has limitations. For example, it assumes that prebiotic chemistry is identical to present-day chemistry. Moreover, even if the predicted results are observed, nothing ensures that the original events really happened in this manner.

3.2.7. Complexification or secondary simplification?

For the layman audience, “evolution” often equates to “progress”, in the form of progressive complexification. This inaccurate yet popular view finds its iconic representation in the Great chain of beings (or *Scala Naturae*) (Ragan 2009), especially when it comes to the “evolution of Man”, going from crawling monkeys to upright Western males wearing office suits. According to S. J. Gould (1941–2002), such a misconception stems from the fact that “people are storytelling creatures” who are fond of so-called “trends”. Moreover, it is true that, at geological scales, the maximal level of organismic complexity is on the rise from the very beginning. However, the maximum can be a very poor measure to describe a statistical distribution (as can be the mean), especially a long-tail distribution. At a planetary scale, Life has always been (and always will be) dominated by microbes (Gould 1996). Even when some criterion is chosen to order extant organisms on a complexity scale, it is very difficult to determine the direction of the evolutionary process: towards an ever more complex state or back to a (much) simpler state, following a path of secondary simplification? A good illustration of this is the case of the RNA polymerase in eukaryotic mitochondria, which has been secondarily simplified by the recruitment of a phage enzyme early in eukaryotic evolution, whereas the other RNA polymerases increase in complexity from Bacteria and Archaea to the nuclei of Eukaryotes (Forterre and Philippe 1999). When repeatedly observing a complex feature scattered across multiple lineages of the Tree of Life, a common argument for secondary simplification is parsimony: it would be "unparsimonious" to assume multiple independent gains (i.e., convergent evolution) of complex features, and thus one has to postulate a single origin, followed by multiple independent losses. This line of thought underpinned a lot of the work of T. Cavalier-Smith (1942–2021), in particular his hypotheses about the evolution of eukaryotic organelles, such as the chloroplast (Cavalier-Smith 1999; Cavalier-Smith 2003). However, it is known today that most of the multiple examples of "complex plastids", once thought to trace back to only one or two events of "secondary endosymbiosis", are actually the result of a much more complex evolution involving lateral transfers of organelles across distant eukaryotic lineages (Petersen et al. 2014; Sibbald and Archibald 2020). Altogether, this brief argumentation, which could have been much longer considering the numerous examples of convergent complexification, e.g., in protists (Lukes et al. 2009), and secondary simplification, suggests that, for any feature displaying a patchy distribution across a group of given organisms, it is impossible to decide *a priori* whether that feature is the product of convergent complexification of secondary simplification. Only proper data collection and model-based analysis can help.

In our article "Was the Last Bacterial Common Ancestor a Monoderm after All?" (Léonard et al. 2022), we tried to determine the ancestral state of the bacterial cell envelope (i.e., monoderm or diderm) using statistical models. Indeed, there are two major hypotheses that are discussed regarding the cell envelope architecture in the LBCA: "diderm-first" (Cavalier-Smith 2006) and "monoderm-first" (Lake 2009; Gupta 2011). The first hypothesis assumes a diderm LBCA with multiple losses of the outer membrane (OM) (i.e., secondary simplification), whereas the second hypothesis assumes the opposite (i.e., convergent complexification). Our results indicate that the LBCA was a monoderm organism, equipped with a complete *dcw* cluster, thus indicating that modern diderm bacteria have arisen multiple times through complexification from a monoderm ancestor. Although a recent study supports our views regarding the *dcw* cluster (Megrian et al. 2022), other studies, using a parsimony-inspired approach, support the "diderm-first" hypothesis entailing secondary simplification (Antunes et al. 2016; Witwinowski et al. 2022). In section 3.2.5, we discussed that the acquisition and subsequent tinkering of bacterial genes by a common ancestor of Methanopyrales and Methanobacteriales had led to the complexification of its cell envelope, through the acquisition of PM. Moreover, we argued in section 3.2.4 that both S-layers and archaeal cell-wall polymers have appeared independently in different archaeal lineages. Finally, diderm archaea are observed in distantly related lineages, thereby suggesting convergent complexification is at work in the archaeal domain too (Rachel et al. 2002; Baker et al. 2006; Comolli et al. 2009; Dridi et al. 2012).

Even though this point was not the initial impetus for our thesis work, research on Archaea has been recently revivified by the discovery of eukaryotic-like archaeal lineages in metagenomic data obtained from marine sediments of Scandinavia (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017). Depending on the rooting of the Tree of Life considered (Gouy et al. 2015), these intermediate lineages can either be taken as transitional on the path from a prokaryotic LUCA (Krupovic et al. 2020) to Eukaryotes, following a "fusion" between an archaeon and at least one bacterium (Embley and Martin 2006), or as an intermediate stage in the simplification of a complex (i.e., eukaryotic-like) LUCA into the simpler "akaryotes" that are regular Archaea and Bacteria (Gouy et al. 2015). According to Nick Lane (Lane 2015), the acquisition of the mitochondrion by an ancient prokaryotic cell was the key to increasing complexity. Indeed, an average eukaryotic cell is 15,000 fold larger than a prokaryotic cell. Thus, a larger cell requires much more energy for protein synthesis. In eukaryotic cells, energy is supplied by mitochondria. In addition, he argued that a large bacterium would require numerous copies of its chromosome to efficiently translate the genes involved in the production of ATP required for protein synthesis. Recently, very large bacteria (up to 2 cm) were described (Angert et al. 1993; Volland et al. 2022). These organisms exhibit extreme polyploidy and thus support Lane's predictions. In their way, Thiomargaritales further blur the lines between the three domains of Life, as did the Asgard Archaea and other lineages with eukaryotic-like features, such as Planctomycetes in the bacterial domain (PVC group) (Devos and Reynaud 2010). In conclusion, despite recent and exciting progress, the jury is still out with respect to the rooting of the Tree of Life and the nature of LUCA (Gouy et al. 2015), including the architecture of its cell envelope. Some authors even suggest that the three domains might actually form a single domain, in which Archaea and Eukaryotes are two parallel experiments from a PVC-like common ancestor (Devos 2021), showing that cellular complexity is not easily mapped on a linear scale.

3.3. References

Adam D, Maciejewska M, Naômé A, Martinet L, Coppieters W, Karim L, Baurain D, Rigali S. 2018. Isolation, Characterization, and Antibacterial Activity of Hard-to-Culture Actinobacteria from Cave Moonmilk Deposits. *Antibiotics* 7:28.

Albarano L, Esposito R, Ruocco N, Costantini M. 2020. Genome Mining as New Challenge in Natural Products Discovery. *Mar. Drugs* 18:199.

Allan EJ, Hoischen C, Gumpert J. 2009. Chapter 1 Bacterial L-Forms. In: Advances in Applied Microbiology. Vol. 68. Academic Press. p. 1–39. Available from: https://www.sciencedirect.com/science/article/pii/S0065216409012015

Angert ER, Clements KD, Pace NR. 1993. The largest bacterium. *Nature* 362:239–241.

Antunes LCS, Poppleton D, Klingl A, Criscuolo A, Dupuy B, Brochier-Armanet C, Beloin C, Gribaldo S. 2016. Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the Firmicutes. *eLife* 5:1–21.

Arakawa K. 2016. No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc. Natl. Acad. Sci. U. S. A.* 113:E3057.

Arita M, Karsch-Mizrachi I, Cochrane G. 2021. The international nucleotide sequence database collaboration. *Nucleic Acids Res.* 49:D121–D124.

Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, Allen EE, Banfield JF. 2006. Lineages of acidophilic archaea revealed by community genomic analysis. *Science* 314:1933–1935.

Benner SA, Sismour AM. 2005. Synthetic biology. *Nat. Rev. Genet.* 6:533–543.

Bharat TAM, von Kügelgen A, Alva V. 2021. Molecular Logic of Prokaryotic Surface Layer Structures. *Trends Microbiol.* 29:405–415.

Bhattacharjee MK. 2016. Antibiotics That Inhibit Cell Wall Synthesis. In: Bhattacharjee MK, editor. Chemistry of Antibiotics and Related Drugs. Cham: Springer International Publishing. p. 49–94. Available from: https://doi.org/10.1007/978-3-319-40746-3_3

Borrel G, Brugère J-F, Gribaldo S, Schmitz RA, Moissl-Eichinger C. 2020. The host-associated archaeome. *Nat. Rev. Microbiol.* 18:622–636.

Bouadjenek MR, Verspoor K, Zobel J. 2017. Automated detection of records in biological sequence databases that are inconsistent with the literature. *J.*

Biomed. Inform. 71:229–240.

Brinkmann H, Philippe H. 2007. The Diversity Of Eukaryotes And The Root Of The Eukaryotic Tree. In: Eukaryotic Membranes and Cytoskeleton: Origins and Evolution. New York, NY: Springer New York. p. 20–37. Available from: https://doi.org/10.1007/978-0-387-74021-8_2

Buchmann JP, Holmes EC. 2015. Cell Walls and the Convergent Evolution of the Viral Envelope. *Microbiol. Mol. Biol. Rev. MMBR* 79:403–418.

Bush K. 2018. Past and Present Perspectives on β-Lactamases. *Antimicrob. Agents Chemother.* 62:e01076-18.

Caspi Y, Dekker C. 2018. Dividing the Archaeal Way: The Ancient Cdv Cell-Division Machinery. *Front. Microbiol.* 9:174.

Cavalier-Smith T. 1987. Eukaryotes with no mitochondria. *Nature* 326:332–333.

Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J. Eukaryot. Microbiol.* 46:347–366.

Cavalier-Smith T. 2003. Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358:109–133; discussion 133-134.

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

Collin S, Guilvout I, Nickerson NN, Pugsley AP. 2011. Sorting of an integral outer membrane protein via the lipoprotein-specific Lol pathway and a dedicated lipoprotein pilotin. 80:655–665.

Colson P, Pinault L, Azza S, Armstrong N, Chabriere E, La Scola B, Pontarotti P, Raoult D. 2020. A protein of the metallo-hydrolase/oxidoreductase superfamily with both beta-lactamase and ribonuclease activity is linked with translation in giant viruses. *Sci. Rep.* 10:21685.

Comolli LR, Baker BJ, Downing KH, Siegerist CE, Banfield JF. 2009. Three-dimensional analysis of the structure and ecology of a novel, ultra-small archaeon. *ISME J.* 3:159–167.

Cornet L, Baurain D. 2022. Contamination detection in genomic data: more is not enough. *Genome Biol.* 23:60.

Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci.* 105:10039–10044.

Dagan T, Martin W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl. Acad. Sci.* 104:870–875.

Delmont TO, Eren AM. 2016. Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ* 4:e1839.

Devos DP. 2021. Reconciling Asgardarchaeota Phylogenetic Proximity to Eukaryotes and Planctomycetes Cellular Features in the Evolution of Life. *Mol. Biol. Evol.* 38:3531–3542.

Devos DP, Reynaud EG. 2010. Evolution. Intermediate steps. *Science* 330:1187–1188.

Di Franco A, Poujol R, Baurain D, Philippe H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol. Biol.* 19:21.

Diene SM, Pinault L, Armstrong N, Azza S, Keshri V, Khelaifia S, Chabrière E, Caetano-Anolles G, Rolain J-M, Pontarotti P, et al. 2020. Dual RNase and β-lactamase Activity of a Single Enzyme Encoded in Archaea. *Life Basel Switz.* 10:280.

Djemai K, Gouriet F, Michel J, Radulesco T, Drancourt M, Grine G. 2021. Methanobrevibacter smithii tonsillar phlegmon: a case report. New microbes and new infections

Dridi B, Fardeau M-L, Ollivier B, Raoult D, Drancourt M. 2012. Methanomassiliicoccus luminyensis gen. nov., sp. nov., a methanogenic archaeon isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* 62:1902–1907.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. *Nature* 440:623–630.

Eme L, Spang A, Lombard J, Stairs CW, Ettema TJG. 2017. Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* 15:711–723.

Errington J. 2017. Cell wall-deficient, L-form bacteria in the 21st century: a personal perspective. *Biochem. Soc. Trans.* 45:287–295.

Evans JT, Denef VJ. 2020. To Dereplicate or Not To Dereplicate? *mSphere* 5.

Forterre P, Philippe H. 1999. Where is the root of the universal tree of life? *BioEssays News Rev. Mol. Cell. Dev. Biol.* 21:871–879.

Frazão N, Sousa A, Lässig M, Gordo I. 2019. Horizontal gene transfer overrides mutation in Escherichia coli colonizing the mammalian gut. *Proc. Natl. Acad. Sci. U. S. A.* 116:17906–17915.

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.

Gill EE, Brinkman FSL. 2011. The proportional lack of archaeal pathogens: Do viruses/phages hold the key? *BioEssays* 33:248–254.

Golyshina OV, Timmis KN. 2005. Ferroplasma and relatives, recently discovered cell wall-lacking archaea making a living in extremely acid, heavy metal-rich environments. *Environ. Microbiol.* 7:1277–1288.

Gould SJ. 1996. Full House. The spread of excellence from Plato to Darwin. New York: Harmony Books

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

Grine G, Drouet H, Fenollar F, Bretelle F, Raoult D, Drancourt M. 2019. Detection of Methanobrevibacter smithii in vaginal samples collected from women diagnosed with bacterial vaginosis. *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.* 38:1643–1649.

Grine G, Lotte R, Chirio D, Chevalier A, Raoult D, Drancourt M, Ruimy R. 2019. Co-culture of Methanobrevibacter smithii with enterobacteria during urinary infection. *EBioMedicine* 43:333–337.

Gupta RS. 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.

Hassani Y, Brégeon F, Aboudharam G, Drancourt M, Grine G. 2020. Detection of Methanobrevobacter smithii and Methanobrevibacter oralis in Lower Respiratory Tract Microbiota. *Microorganisms* 8:1866.

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. *Nat. Ecol. Evol.* 1:1370–1378.

Ithurbide S, Gribaldo S, Albers S-V, Pende N. 2022. Spotlight on FtsZ-based cell division in Archaea. *Trends Microbiol.* 30:665–678.

Jeske O, Schüler M, Schumann P, Schneider A, Boedeker C, Jogler M, Bollschweiler D, Rohde M, Mayer C, Engelhardt H, et al. 2015. Planctomycetes do possess a peptidoglycan cell wall. *Nat. Commun.* 6:7116.

Joseleau-Petit D, Liébart J-C, Ayala JA, D'Ari R. 2007. Unstable Escherichia coli L forms revisited: growth requires peptidoglycan synthesis. *J. Bacteriol.* 189:6512–6520.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* 9:605–618.

Khelaifia S, Drancourt M. 2012. Susceptibility of archaea to antimicrobial agents: applications to clinical microbiology. *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.* 18:841–848.

Klingl A, Pickl C, Flechsler J. 2019. Archaeal Cell Walls. In: Kuhn A, editor. Bacterial Cell Walls and Membranes. Cham: Springer International Publishing. p. 471–493. Available from: https://doi.org/10.1007/978-3-030-18768-2_14

Kominek J, Doering DT, Opulente DA, Shen X-X, Zhou X, DeVirgilio J, Hulfachor AB, Groenewald M, Mcgee MA, Karlen SD, et al. 2019. Eukaryotic Acquisition of a Bacterial Operon. *Cell* 176:1356-1366.e10.

Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker AA, Blaxter M. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. *Proc. Natl. Acad. Sci. U. S. A.* 113:5053–5058.

Krupovic M, Dolja VV, Koonin EV. 2020. The LUCA and its complex virome. *Nat. Rev. Microbiol.* 18:661–670.

Lake JA. 2009. Evidence for an early prokaryotic endosymbiosis. *Nature* 460:967–971.

Lane N. 2015. The Vital Question: Why is life the way it is? Profile Available from: https://books.google.be/books?id=IfJYBQAAQBAJ

Leaver M, Domínguez-Cuevas P, Coxhead JM, Daniel RA, Errington J. 2009. Life without a wall or division machine in Bacillus subtilis. *Nature* 457:849–853.

Léonard R. 2021. Bacterial cell-wall architecture: from automated genome selection to evolution of genes and traits.

Léonard RR, Leleu M, Vlierberghe MV, Cornet L, Kerff F, Baurain D. 2021. ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies. *PeerJ* 9:e11348.

Léonard RR, Sauvage E, Lupo V, Perrin A, Sirjacobs D, Charlier P, Kerff F, Baurain D. 2022. Was the Last Bacterial Common Ancestor a Monoderm after All? *Genes* 13:376.

Liechti G, Kuru E, Packiam M, Hsu Y-P, Tekkam S, Hall E, Rittichier JT, VanNieuwenhze M, Brun YV, Maurelli AT. 2016. Pathogenic Chlamydia Lack a Classical Sacculus but Synthesize a Narrow, Mid-cell Peptidoglycan Ring, Regulated by MreB, for Cell Division. *PLoS Pathog.* 12:e1005590.

Liechti GW, Kuru E, Hall E, Kalinda A, Brun YV, VanNieuwenhze M, Maurelli AT. 2014. A new metabolic cell-wall labelling method reveals peptidoglycan in Chlamydia trachomatis. *Nature* 506:507–510.

Lindås A-C, Karlsson EA, Lindgren MT, Ettema TJG, Bernander R. 2008. A unique cell division machinery in the Archaea. *Proc. Natl. Acad. Sci.* 105:18942–18946.

Lukes J, Leander BS, Keeling PJ. 2009. Cascades of convergent evolution: the corresponding evolutionary histories of euglenozoans and dinoflagellates. *Proc. Natl. Acad. Sci. U. S. A.* 106 Suppl 1:9963–9970.

Lupo V, Mercuri PS, Frère J-M, Joris B, Galleni M, Baurain D, Kerff F. 2022. An Extended Reservoir of Class-D Beta-Lactamases in Non-Clinical Bacterial Strains. *Microbiol. Spectr.* 10:e0031522.

Lupo V, Van Vlierberghe M, Vanderschuren H, Kerff F, Baurain D, Cornet L. 2021. Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics. *Front. Microbiol.* 12:755101.

Mahilum-Tapy L. 2009. The importance of microbial culture collection and Gene Banks in biotechnology. *Biotechnol. Encycl. Life Support Syst. EOLSS Eolss Publ. Oxf.*

Martínez-Torró C, Torres-Puig S, Marcos-Silva M, Huguet-Ramón M, Muñoz-Navarro C, Lluch-Senar M, Serrano L, Querol E, Piñol J, Pich OQ. 2021. Functional Characterization of the Cell Division Gene Cluster of the Wall-less Bacterium Mycoplasma genitalium. *Front. Microbiol.* 12:695572.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6:610–618.

McQuillen R, Xiao J. 2020. Insights into the Structure, Function, and Dynamics of the Bacterial Cytokinetic FtsZ-Ring. *Annu. Rev. Biophys.* 49:309–341.

Megrian D, Taib N, Jaffe AL, Banfield JF, Gribaldo S. 2022. Ancient origin and constrained evolution of the division and cell wall gene cluster in Bacteria. *Nat. Microbiol.* 7:2114–2127.

Meyer BH, Albers S-V. 2020. Archaeal Cell Walls. In: eLS. p. 1–14. Available from: https://doi.org/10.1002/9780470015902.a0000384.pub3

Mingorance J, Tamames J. 2004. The bacterial dcw gene cluster: an island in the genome? In: Vicente M, Tamames J, Valencia A, Mingorance J, editors. Molecules in Time and Space: Bacterial Shape, Division and Phylogeny. Boston, MA: Springer US. p. 249–271. Available from: https://doi.org/10.1007/0-306-48579-6_13

Münch D, Roemer T, Lee SH, Engeser M, Sahl HG, Schneider T. 2012. Identification and in vitro analysis of the GatD/MurT enzyme-complex catalyzing lipid II amidation in Staphylococcus aureus. *PLoS Pathog.* 8:e1002509.

Mura A, Fadda D, Perez AJ, Danforth ML, Musu D, Rico AI, Krupka M, Denapaite D, Tsui H-CT, Winkler ME, et al. 2017. Roles of the Essential Protein FtsA in Cell Growth and Division in Streptococcus pneumoniae. *J. Bacteriol.* 199:e00608-16.

Nöldeke ER, Muckenfuss LM, Niemann V, Müller A, Störk E, Zocher G, Schneider T, Stehle T. 2018. Structural basis of cell wall peptidoglycan amidation by the GatD/MurT complex of Staphylococcus aureus. *Sci. Rep.* 8:12953.

Osawa M, Erickson HP. 2019. L form bacteria growth in low-osmolality medium. *Microbiology,* 165:842–851.

Ouellette SP, Lee J, Cox JV. 2020. Division without Binary Fission: Cell Division in the FtsZ-Less Chlamydia. *J. Bacteriol.* 202.

Packiam M, Weinrick B, Jacobs WRJ, Maurelli AT. 2015. Structural characterization of muropeptides from Chlamydia trachomatis peptidoglycan by mass spectrometry resolves “chlamydial anomaly”. *Proc. Natl. Acad. Sci. U. S. A.* 112:11660–11665.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36:996–1004.

Pazos M, Peters K. 2019. Peptidoglycan. In: Kuhn A, editor. Bacterial Cell Walls and

Membranes. Cham: Springer International Publishing. p. 127–168. Available from: https://doi.org/10.1007/978-3-030-18768-2_5

Pende N, Sogues A, Megrian D, Sartori-Rupp A, England P, Palabikyan H, Rittmann SK-MR, Graña M, Wehenkel AM, Alzari PM, et al. 2021. SepF is the FtsZ anchor in archaea, with features of an ancestral cell division system. *Nat. Commun.* 12:3214.

Petersen J, Ludewig A-K, Michael V, Bunk B, Jarek M, Baurain D, Brinkmann H. 2014. Chromera velia, endosymbioses and the rhodoplex hypothesis--plastid evolution in cryptophytes, alveolates, stramenopiles, and haptophytes (CASH lineages). *Genome Biol. Evol.* 6:666–684.

Rachel R, Wyschkony I, Riehl S, Huber H. 2002. The ultrastructure of Ignicoccus: evidence for a novel outer membrane and for intracellular vesicle budding in an archaeon. *Archaea Vanc. BC* 1:9–18.

Ragan MA. 2009. Trees and networks before and after Darwin. *Biol. Direct* 4:43.

Rasmussen M, Collin M. 2021. Archaea in Blood Cultures: Coincidence or Coinfection? *Clin. Infect. Dis.* 73:e2580–e2581.

Real G, Henriques AO. 2006. Localization of the Bacillus subtilis murB gene within the dcw cluster is important for growth and sporulation. *J. Bacteriol.* 188:1721–1732.

Reysenbach A-L. 2015. Thermoplasmatales ord. nov. In: Bergey’s Manual of Systematics of Archaea and Bacteria. p. 1–1. Available from: https://doi.org/10.1002/9781118960608.obm00055

Salzberg SL. 2017. Horizontal gene transfer is not a hallmark of the human genome. *Genome Biol.* 18:85.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, et al. 2022. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 50:D20–D26.

Schmitt I, Lumbsch HT. 2009. Ancient horizontal gene transfer from bacteria enhances biosynthetic capabilities of fungi. *PloS One* 4:e4437.

Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O’Neill K, Robbertse B, et al. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* 2020:baaa062.

Schofield LR, Beattie AK, Tootill CM, Dey D, Ronimus RS. 2015. Biochemical

Characterisation of Phage Pseudomurein Endoisopeptidases PeiW and PeiP Using Synthetic Peptides. *Archaea Vanc. BC* 2015:828693.

Sengupta S, Chattopadhyay MK, Grossart H-P. 2013. The multifaceted roles of antibiotics and antibiotic resistance in nature. *Front. Microbiol.* 4:47.

Sharma SK, Saini S, Verma A, Sharma PK, Lal R, Roy M, Singh UB, Saxena AK, Sharma AK. 2019. National Agriculturally Important Microbial Culture Collection in the Global Context of Microbial Culture Collection Centres. *Proc. Natl. Acad. Sci. India Sect. B Biol. Sci.* 89:405–418.

Sibbald SJ, Archibald JM. 2020. Genomic Insights into Plastid Evolution. *Genome Biol. Evol.* 12:978–990.

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, et al. 2017. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr. Biol. CB* 27:958–967.

Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* 16:472–482.

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

Tamames J. 2001. Evolution of gene order conservation in prokaryotes. *Genome Biol.* 2:RESEARCH0020.

van Teeseling MCF, Mesman RJ, Kuru E, Espaillat A, Cava F, Brun YV, VanNieuwenhze MS, Kartal B, van Niftrik L. 2015. Anammox Planctomycetes have a peptidoglycan cell wall. *Nat. Commun.* 6:6878.

Thi Nhu Thao T, Labroussaa F, Ebert N, V’kovski P, Stalder H, Portmann J, Kelly J, Steiner S, Holwerda M, Kratzel A, et al. 2020. Rapid reconstruction of SARS-CoV-2 using a synthetic genomics platform. *Nature* 582:561–565.

Trachtenberg S. 1998. Mollicutes-wall-less bacteria with internal cytoskeletons. *J. Struct. Biol.* 124:244–256.

Venter JC, Glass JI, Hutchison CA, Vashee S. 2022. Synthetic chromosomes, genomes, viruses, and cells. *Cell* 185:2708–2724.

Volland J-M, Gonzalez-Rizzo S, Gros O, Tyml T, Ivanova N, Schulz F, Goudeau D, Elisabeth NH, Nath N, Udwary D, et al. 2022. A centimeter-long bacterium with DNA contained in metabolically active, membrane-bound organelles.

Science 376:1453–1458.

Witwinowski J, Sartori-Rupp A, Taib N, Pende N, Tham TN, Poppleton D, Ghigo J-M, Beloin C, Gribaldo S. 2022. An ancient divide in outer membrane tethering systems in bacteria suggests a mechanism for the diderm-to-monoderm transition. *Nat. Microbiol.* 7:411–422.

Yubuki N, Galindo LJ, Reboul G, López-García P, Brown MW, Pollet N, Moreira D. 2020. Ancient Adaptive Lateral Gene Transfers in the Symbiotic Opalina–Blastocystis Stramenopile Lineage. *Mol. Biol. Evol.* 37:651–659.

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

4. Annexes

4.1. Was the Last Bacterial Common Ancestor a Monoderm after All?

genes

MDPI

Article

Was the Last Bacterial Common Ancestor a Monoderm after All?

Raphaël R. Léonard [1,2], Eric Sauvage [1], Valérian Lupo [1,2], Amandine Perrin [3,4], Damien Sirjacobs [2], Paulette Charlier [1], Frédéric Kerff [1,*] and Denis Baurain [2,*]

[1] InBioS–Centre d'Ingénierie des Protéines, Université de Liège, 4000 Liege, Belgium; rleonard@doct.uliege.be (R.R.L.); ericsauvage8@gmail.com (E.S.); valerian.lupo@doct.uliege.be (V.L.); paulette.charlier@uliege.be (P.C.)

[2] InBioS–PhytoSYSTEMS, Unit of Eukaryotic Phylogenomics, Université de Liège, 4000 Liege, Belgium; d.sirjacobs@uliege.be

[3] University Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France; amandine.perrin@pasteur.fr

[4] Hub de Bioinformatique et Biostatistique-Département Biologie Computationnelle, Institut Pasteur, 75015 Paris, France

* Correspondence: fkerff@uliege.be (F.K.); denis.baurain@uliege.be (D.B.)

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 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 of cultivated and characterized bacteria as an evolutionary framework and a literature review of their 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 outer membrane in diderm bacteria. Quite unexpectedly, our analyses suggest that all cultivated and characterized bacteria might have evolved from a common ancestor with a monoderm cell-wall architecture. If true, this would indicate that the appearance of the outer membrane was not a unique event and that selective forces have led to the repeated adoption of such an architecture. Due to the lack of phenotypic information, our methodology cannot be applied to all extant bacteria. Consequently, our conclusion might change once enough information is made available to allow the use of an even more diverse organism selection.

Keywords: bacterial evolution; cell-wall; outer membrane (OM); Bayesian inference (BI); phylogenomics; comparative genomics; ancestral traits

check for updates

Citation: Léonard, R.R.; Sauvage, E.; Lupo, V.; Perrin, A.; Sirjacobs, D.; Charlier, P.; Kerff, F.; Baurain, D. Was the Last Bacterial Common Ancestor a Monoderm after All? *Genes* **2022**, *13*, 376. https://doi.org/10.3390/genes13020376

Academic Editors: Daisuke Kageyama and Renè Massimiliano Marsano

Received: 31 December 2021
Accepted: 15 February 2022
Published: 18 February 2022

1. Introduction

Cell-wall architecture has always been an important morphological character for bacterial classification [1]. 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 (and are negative to Gram coloration) [2,3]. However, cell-wall features are insufficient to yield a classification that would correlate with phylogenetic trees based on molecular data [4]. 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) [5]. Nonetheless, the evolution of the bacterial cell wall should be addressed considering 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 often 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 going from "diderm" to "monoderm" architecture has been proposed by Cavalier-Smith [6,7] but evolution from monoderm to diderm bacteria is usually favoured by other evolutionary biologists [8–11]. 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 [12], but how exactly this symbiosis could have further evolved to form a diderm bacterium still is to be detailed. An attractive hypothesis accounting for the emergence of the outer membrane 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 outer membrane of the forespore and ultimately of the germinating cell [13], leading to the assumption that the outer membrane of diderm bacteria could have evolved from monoderms via sporulation [11,13–15]. 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 the double membrane in the Firmicutes ancestor, and thus concluded that the outer membrane was lost multiple times in this phylum [16,17]. 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 [18]. 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 [18]. This cluster is found in many bacteria and its composition and gene order are generally well conserved [19,20]. It has also been shown to be one of the most stable gene clusters (the cluster itself and the gene synteny within the cluster are conserved in a broad taxonomic range of genomes) [18], on par with the ribosomal clusters [21,22]. 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 [23]. 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 well-known examples [24]. 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 outer membrane in diderm bacteria might provide useful information about their evolutionary status, ancestral or derived, with respect to the bacterial domain as a whole [5,17,25].

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 outer membrane of diderm bacteria appeared more than once, a hypothesis that is indeed supported by differences in the genetic machinery involved in its biosynthesis across the various diderm lineages, as shown by our extensive similarity searches. Our results also show that the LBCA already had a complete *dcw* cluster and that its organization does not correlate with cell-wall architecture.

2. Materials and Methods

2.1. Dataset Assembly

2.1.1. Data Download

The initial dataset of prokaryotic genomes and proteomes was downloaded from Ensembl Bacteria release 20 [26] using wget. This dataset had 8848 Bacteria and 238 Archaea represented.

2.1.2. Genome Dereplication and Selection

We first reduced the number of genomes based on genomic signatures [27] to regroup similar genomes into genome clusters with a prerelease version of our new software ToRQuEMaDA [28]. 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 [29]. The complete lineage of every genome was recovered from the NCBI Taxonomy database [30] using the program fetch-tax.pl from the Bio::MUST::Core distribution (D. Baurain, https://metacpan.org/dist/Bio-MUST-Core, accessed on 16 February 2022). Each signature file was further analysed in R [31] 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). 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 or 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) [28]. 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 [28]. After including a few other 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.

2.1.3. 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 [32] with the following parameters (evalue = 1×10^{-5}; accel = 1; threads = 64). Then, we used OrthoMCL v2.0.3 [33] to cluster protein sequences into orthologous groups based on USEARCH reports, using an e-value cut-off of 1×10^{-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 orthologous groups, whereas 326,269 sequences were considered as "singletons" by OrthoMCL (i.e., without homologues).

2.1.4. 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.

2.2. Evolution of the Bacterial Domain

2.2.1. 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 maximize the number of classes. Then, using classify-mcl-out.pl [34], we selected all orthologous groups 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 had 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 orthologous groups to ensure a maximum of 14 missing species in each individual orthologous group (Table S1). The corresponding orthologous groups were aligned with MAFFT v7.127b [35] using default parameters. The protein sequence alignments were then filtered with Gblocks v0.91b [36] using a set of "medium stringency" parameters (as predefined in Bio::MUST::Core) and concatenated with SCaFoS v1.30k [37]. Finally, the resulting concatenation was further filtered for sites >50% missing character states, yielding a supermatrix of 85 species × 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).

2.2.2. Phylogenomic Analyses

For Bayesian inference (BI), we used PhyloBayes MPI v1.5 [38] to produce six replicate Markov Chain Monte–Carlo (MCMC) chains of 50,000 cycles, with one tree sampled every 10 cycles, using the CAT+GTR+Γ model of sequence evolution [39–41]. 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 posterior probabilities) 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 decide the best-fit model (CAT+GTR+Γ) were carried out using PhyloBayes v3.3f [42], as suggested in PhyloBayes manual (p. 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 5000 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 [43].

2.2.3. Congruence Tests

Congruence tests were performed on the 85-species supermatrix genes with Phylo-MCOA v1.4 [44], then Maximum Likelihood (ML) reconstruction with RAxML v8.1.17 [45] 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) found by Phylo-MCOA.

2.3. Evolution of the Cell-Wall

2.3.1. Cell-Wall Architecture of Extant Organisms

For each one of the 85 bacterial species, a dedicated survey of the literature was conducted (Table S2). 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 [46–48]. For *Parachlamydia acanthamoebae*, no clue about peptidoglycan thickness was found, so this trait was coded as "12", following the suggestion in BayesTraits manual (p. 9).

2.3.2. 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 (p. 13). We applied the steppingstone sampler, using 100 stones with 10,000 iterations per stone. As this procedure only allows for the comparison of two binary traits, and as 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.

2.3.3. 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 (p. 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. Ten 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 $10 \times 10{,}000$ samples. To investigate the sensitivity of the Bayesian inference of a monoderm LBCA to priors, one more 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.

2.3.4. Comparison of the Selected Models

Building on the steppingstones 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 and the peptidoglycan data (used for the ancestral state reconstruction) using Bayes Factors. We selected the steppingstones files from the runs with the tree rooted on the Terrabacteria. As above, the steppingstone sampler used 100 stones with 10,000 iterations per stone.

2.4. Evolution of the dcw Cluster

2.4.1. 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 allowed 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) needed 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 in the folder ProCARs. The orthologous groups 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 orthologous group was found for each gene. For *ftsL* and, to a much lesser extent, *ftsQ*, several orthologous groups 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 just before *ftsA* for *ftsQ*. Moreover, HMM profiles (pHMM) [49,50] (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 orthologous groups (many having only 2–3 sequences), respectively, whereas *mraW*, *mraZ* and *ftsA* were spread over 2, 3 and 4 orthologous groups, respectively.

2.4.2. Ancestral Gene Order Reconstruction

To reconstruct the evolution of the *dcw* cluster, we used the program ProCARs [51], 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 of 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.

2.4.3. Phylogenetic Analyses

For the single-gene analyses of the *dcw* cluster in the 85 genomes of interest, we used the 17 identified orthologous groups (possibly merged; 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 $\geq$50% missing character states, thereby yielding a sparser supermatrix of 85 species × 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.

2.5. Evolution of the Genes Related to the Outer Membrane

2.5.1. Homology Searches in Complete Proteomes

For our broader study of the taxonomic distribution of 16 genes involved in synthesis and in maintaining the integrity of the outer membrane across the 903 selected genomes (including previously discarded organisms like Thermotogae), we did not rely on synteny as those were not part of a single cluster in any organism. Instead, we searched for the orthologous groups containing unambiguous reference sequences for these genes. For each set of orthologous groups potentially corresponding to a gene of interest (merging from one to nine orthologous groups per gene), 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 [49,50]. 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, accessed on 16 February 2022). In contrast, when the alignment of all sequences was too poor, we focused on the original orthologous group containing the *E. coli* sequence and tried to build a profile by adding up to 6 (for *lolB* and *lptC*) of the additional orthologous groups 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, 16 February 2022). Then, we followed the same route as if the pHMM had been computed from a "good-enough" alignment. For the specific case of the *bamA* gene, we first collected 28 orthologous groups containing proteins annotated as BamA, Omp85 and/or TspB, then we used InterProScan v.5.48-83.0 with default parameters and disabled use of the precalculated match lookup [52] to determine the number of POTRA domains [53] in the 1425 individual sequences. Two curated alignments based on preliminary ML trees (see below) were built: one from the five orthologous groups where the sequences mostly had 4 or 5 POTRA domains (Table S4), which we considered as the orthologues of the genuine BamA protein of true diderms-LPS, and one from five orthologous groups having 2 or 3 POTRA domains, which included the BamA "4" sequences of Cyanobacteria, as well as related proteins (i.e., BamA-like/Lipo/TamA) [54]. By "curation", we mean elimination of incomplete and/or divergent individual sequences but without discarding representatives of scarcer groups. Finally, these two alignments were used to build two pHMM profiles and perform HMMER searches as described above.

2.5.2. Taxonomic and Phylogenetic Analyses

For each gene of the 16 genes, we retrieved the list of genomes containing 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 atypical diderms group. In each case, only one of the expected phyla of the atypical diderms 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 atypical diderms group (when they had copies of the gene). To decide 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 outer membrane following the same method. All trees were formatted as above, with unresolved nodes (BP < 25%) turned to multifurcations.

3. Results

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 [55]. Phylogenomic trees can either be inferred from supermatrices of concatenated genes [56] or through combination of single-gene trees into supertrees [57]. 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 (CATegories + Generalised Time-Reversible + Gamma) of sequence evolution [38–41]. Congruence analyses were run on the 117 individual genes using Phylo-MCOA [44] 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 S5).

Our unrooted tree is in good agreement with most recent concatenating phylogenomic studies aimed at resolving bacterial evolution [58–68]. In particular, we robustly recovered a bipartition of the bacterial lineages composing the Terrabacteria and the "Hydrobacteria" (=Gracilicutes sensu [69]). Within these "megaphyla" first defined by Hedges and Battistuzzi [58], resolution was weaker, as reflected in the lower posterior probabilities at medium phylogenetic depth, whereas phyla and known superphyla (e.g., FBC, for Fibrobacteres-Bacteroidetes-Chlorobi, and PVC, for Planctomycetes-Verrucomicrobia-Chlamydia) were always clearly resolved. In the Terrabacteria, relationships between member lineages slightly varied from run to run (we ran a total of six independent chains, Figure S3), while in the Hydrobacteria (e.g., FBC, PVC, Proteobacteria), Epsilonproteobacteria were occasionally separated from other groups of Proteobacteria (Figures S4 and S5). 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 [70–72]). Likewise, we further discarded the Thermotogae, which are also chimeras [70], even though their toga might be akin to a modified outer membrane [73,74] (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) [72,75–82].

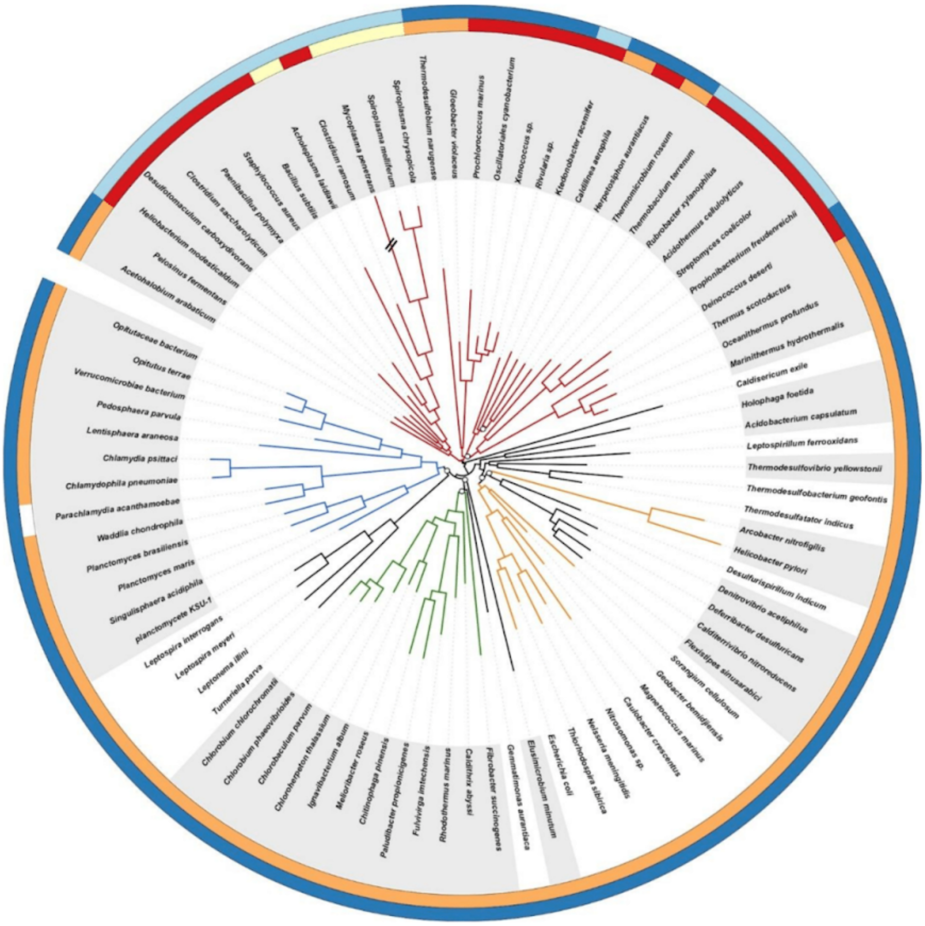


Figure 1. Phylogenomic tree of the bacterial domain based on a supermatrix concatenating 117 single-copy orthologous genes chosen for their broad conservation across Bacteria. The tree was rooted on Terrabacteria. The supermatrix had 85 species and 19,959 unambiguously aligned amino-acid positions (<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 (posterior probabilities 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. Colour key is red = Terrabacteria, orange = Proteobacteria, green = FBC superphylum, blue = PVC superphylum. Outer circles stand for the status of the peptidoglycan (PG) and of the outer membrane in the organisms, according to our literature survey: red = thick PG, orange = thin PG, yellow = no PG, dark blue = diderm, light blue = monoderm, white = no information. Alternating white and grey backgrounds highlight the alternance between differentially coloured groups or phyla.

Rooting the different domains of Life is not an easy issue [82]. In Figure 1, we chose to set the root of Bacteria between Terrabacteria and Hydrobacteria/Gracilicutes, following studies having included Archaea as an outgroup [25,41]. Remarkably, 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, i.e., Deinococcus-Thermus, Cyanobacteria, Synergistetes and Thermotogae, they compose Terrabacteria [58]. 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" in this study. Over time, several positions for the bacterial root have been proposed (Table S6). In the following, because our Bayesian analyses required a rooted tree, we tested several of them, yet excluding roots lying within the true diderms-LPS, 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.

3.2. Evolution of the Cell-Wall Architecture

To study the evolution of the cell-wall architecture, we carried out a thorough literature survey on all the bacteria kept 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 remodelling processes [13,14]. However, preliminary analyses showed 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 true diderms-LPS lineages. For the organisms belonging to the PVC superphylum, this architecture might be slightly different [83]. 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 [84] (Halanaerobiales and Thermoanaerobacteriales) and the Negativicutes have two membranes (some with lipopolysaccharides in the outer membrane) and a relatively thin peptidoglycan layer [16,85,86]. Regarding the status of the Chloroflexi cell-wall architecture, it is still controversial [68,87,88]. 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 outer membrane, but their cell walls differ from those of the true diderms-LPS by having a thick peptidoglycan instead of a thin layer (Table S2).

Consequently, the number of membranes observed in the extant organisms is either one (state 0) or two (i.e., there is an outer membrane, 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 true diderms-LPS (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 outer membrane (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).

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 favour a diderm LBCA (see Materials and Methods for details).

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

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).

In a second step, we used BayesTraits to reconstruct the state of the characters for the Last Common Ancestor (LCA) of every one 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 true diderms-LPS 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).

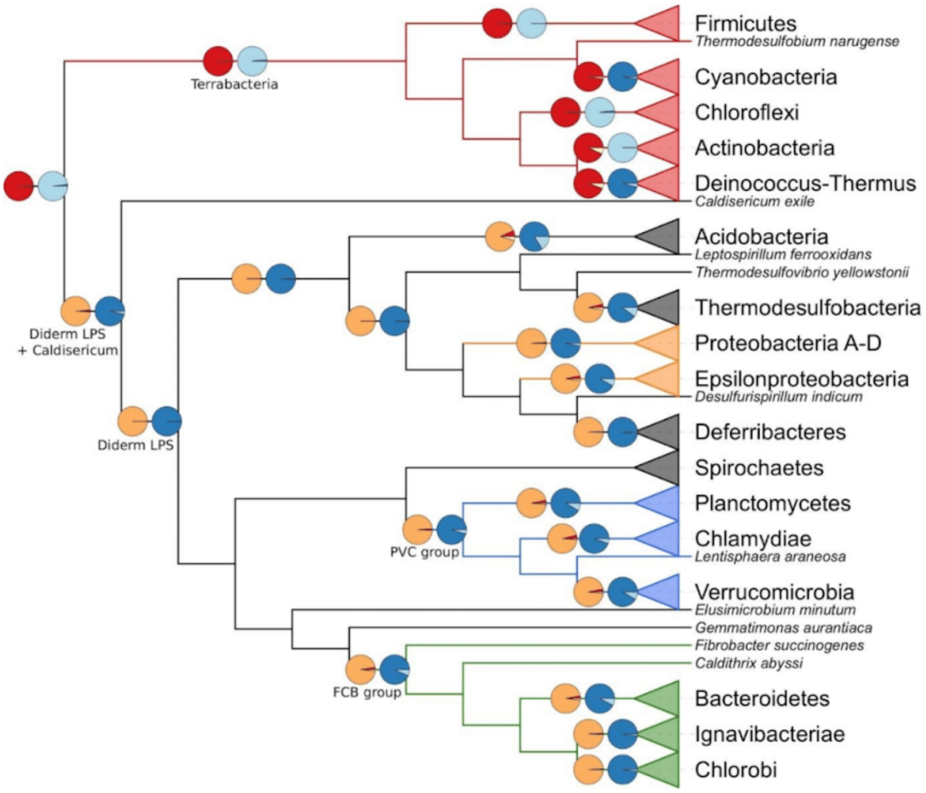


Figure 2. Cladogram derived from the tree of Figure 1 featuring the cell-wall architecture inferred for selected last common ancestors among Bacteria. Colour key is red = Terrabacteria, orange = Proteobacteria, green = FBC superphylum, blue = PVC superphylum Branches ending with a triangle represent collapsed groups (for details, see Figure 1 or Table S3). The pie chart sectors correspond to the posterior probabilities of the model reverse-jump hyperprior exponential 0 to 100 (R2). Colour key is red = thick PG, orange = thin PG, yellow = no PG, dark blue = diderm, light blue = monoderm.

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 favour of the more complex model [89].

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

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 outer membrane (see *Parachlamydia acanthamoebae* in Figure 1), whereas organisms with a peptidoglycan layer often have an outer membrane. 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 outer membrane, 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 outer membrane.

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 visualization software showing orthologous gene groups in their syntenic context (see Materials and Methods for details and "synteny_85_dcw.pdf" available in the folder ProCARs from our Figshare, for the status of the *dcw* cluster in the 85 bacteria of our phylogenomic tree). This approach led us to identify the orthologous groups 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 in the 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 seen 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 [90]. However, *murA* is also present in sub-clusters in Cyanobacteria, Planctomycetes, Lentisphaerae and *Caldithrix abyssi*. Otherwise, if present in the genome, *murA* is usually outside of the *dcw* cluster. Beside this specific gene and particular phyla, several true diderms-LPS 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) (see .xlsx file available in the folder ProCARs).

Taking the rooted phylogenomic tree of Figure 1 as an evolutionary framework and the orthologous groups identified just above as input extant data, we used a new variant of a homology-based reconstruction method (ProCARs) [51] 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 true diderms-LPS group included 16 genes, with the *murA* gene 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 true diderms-LPS (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 seen in the LCA of the FBC superphylum where *murA* and *murB* are in tandem. A parsimonious way to explain these observations would be that the *murA* gene has left the *dcw* gene cluster (but persisted in the genome) of the LCA of true diderms-LPS and the LCA of Actinobacteria, Deinococcus-Thermus and Chloroflexi (assuming these three phyla share a common ancestor). Alternatively, it was lost independently in the three latter phyla. 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 in the 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.

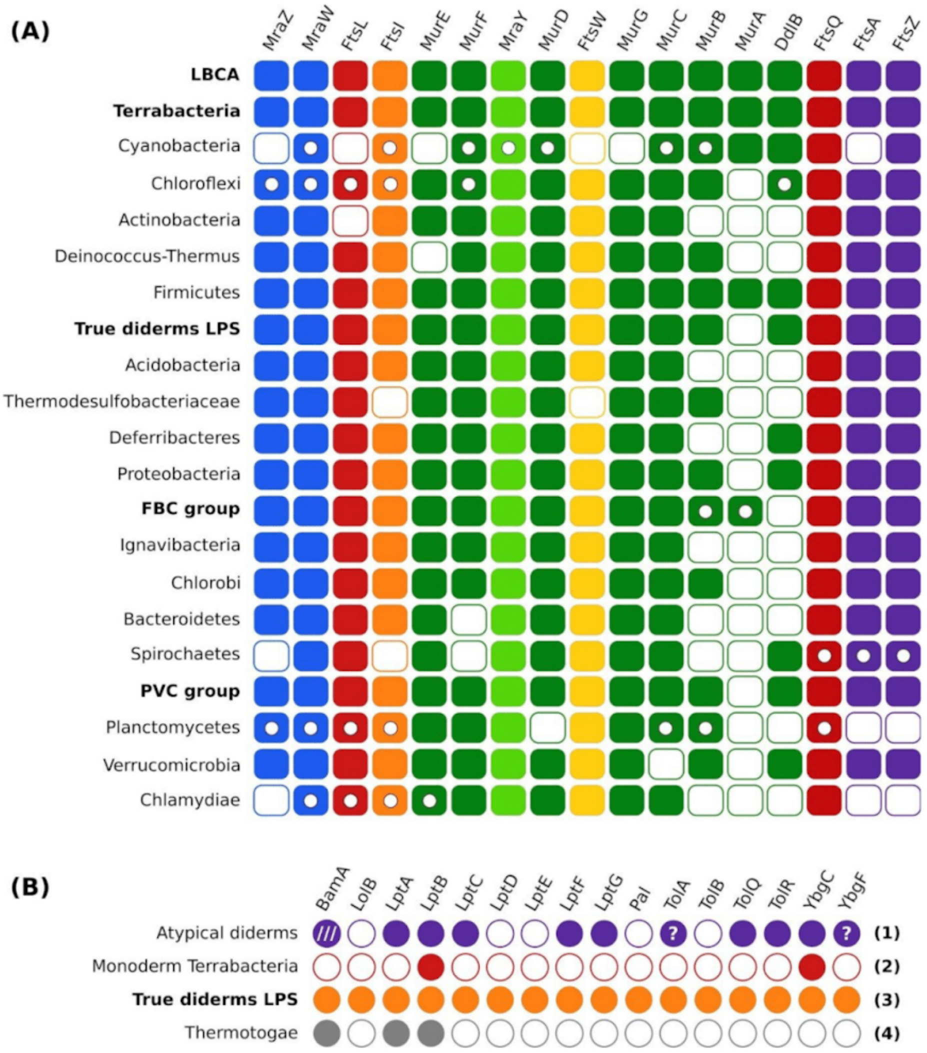


Figure 3. Overview of gene distribution and synteny analyses. (**A**) ProCARs results for *dcw* cluster

organization in selected LCA among Bacteria. Full rectangle = gene present and in the main cluster; empty circle in rectangle = gene present but 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 outer membrane. 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; /// = presence in a sub-group only (i.e., Deinococcus-Thermus). 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 in the folder Trees). Known phyla can be supported by low to high bootstrap proportions (BP: 9–100%) and posterior probabilities (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 true diderms-LPS (Firmicutes–Chloroflexi–Actinobacteria–Deinococcus-Thermus vs. Proteobacteria–FBC–PVC), are observable in several single-gene trees. Moreover, trees inferred from genes often found outside of the *dcw* cluster (e.g., *murC*, *murB* and *ddlB*) are blurrier than those computed from genes kept in the cluster. Finally, the trees of the genes *ftsQ* and *ftsL*, for which the orthologous groups 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%; posterior probabilities: 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 true diderms-LPS, 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.

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 atypical diderms lineages within Terrabacteria (Cyanobacteria, Deinococcus-Thermus and some Firmicutes, i.e., the Halanaerobiales and the Negativicutes) had to acquire their outer membrane independently and in distinct events from the event at the origin of true diderms-LPS. At face value, this inference might seem less parsimonious than hypothesizing a diderm LBCA and multiple independent outer membrane losses over the evolution of the bacterial domain, as suggested repeatedly [5,25,68]. To decide whether the outer membrane could indeed have evolved several times independently, we studied the taxonomic distribution of 16 genes involved in outer membrane 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 outer membrane [91]. LolB is the only outer membrane-anchored protein of the Lol pathway, which delivers lipoproteins to the outer membrane [3]. The Lpt system (LptA to LptG) ensures the transport of the lipopolysaccharides from the cytoplasm to the outer membrane [92]. 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 outer membrane [93].

The distribution of these genes was examined across our first selection of 903 bacterial genomes (all genomes even the previously discarded ones) using curated Hidden Markov Model (HMM) profiles built from orthologous groups 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 and "OM_genes_presence-hmms.csv" available in the folder Outer_membrane, for details), while single-gene trees inferred from the corresponding protein sequences are available ("LBCA_OM_16_SG.pdf" available in the 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" (see references in Table S2), i.e., Cyanobacteria, Deinococcus-Thermus and diderm Firmicutes; (2) "monoderm Terrabacteria", i.e., Chloroflexi, of which some may be monoderms but all are devoid of lipopolysaccharides [68,87], Actinobacteria, and monoderm Firmicutes; (3) "true diderms with LPS" (TDL = typical Gram–bacteria); (4) Thermotogae, in which the outer membrane has been replaced by a toga made of structural proteins and polysaccharide hydrolases (xylanases) [73,74,94]. Below, we briefly comment on these gene distributions from a functional perspective.

First, according to our comprehensive homology searches, *bamA* is exclusive to true diderms-LPS, Deinococcus-Thermus and Thermotogae, even though the latter lack nearly all other outer membrane-related genes studied here. This result suggests a true diderms-LPS origin for Thermotogae, which are now considered as chimeras partly derived from (or at least related to) Aquificales [70,72,95]. This chimerical nature of Thermotoga is the reason we did not include them in our phylogenomic tree (see above). Regarding the presence of the *bamA* gene in the atypical diderms of the group Deinococcus-Thermus, it has already been reported [96] and this result appears less compatible with a monoderm LBCA. However, in other atypical diderms, we could not find a genuine BamA protein. Instead, Cyanobacteria and diderm Firmicutes feature proteins that have a quite different domain architecture (see BamA4 and BamA-like in Heinz et al., 2014 [54]) and for which the orthology (i.e., overall sequence similarity due to vertical descent only) with the typical BamA is at best dubious. Therefore, we currently disagree with the idea that BamA per se would be common outside true diderms-LPS [97]. Nonetheless, BamA, taken as a family regrouping the typical BamA, "BamA4" and "BamA-like" proteins, might indeed be an essential family (each sub-group sharing a similar function) to all diderm (i.e., featuring an outer membrane) but its members do not necessarily share a vertical transmission from a single ancestral protein. To verify this hypothesis would require a whole new study and is thus not expanded in the current article. Second, *lolB* is exclusive to Proteobacteria, a member of true diderms-LPS, 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 outer membrane. It is likely that these two genes have function(s) outside their respective system, functions that could be unrelated to the outer membrane. This has already been proposed for *ybgF*, which might be part of a protein network involved in phospholipid biosynthesis [98]. On the opposite, the LptB protein is known to assemble with LptF and LptG to form an ABC transporter for lipopolysaccharides [92,99], but the two corresponding genes are apparently lacking in Acidobacteria (true diderms-LPS), Tenericutes and Chloroflexi. Perhaps unexpectedly, this is also the case for Actinobacteria, these monoderm bacteria further sharing with Chloroflexi the same distribution pattern for the 16 genes involved with the outer membrane.

Beyond *lptB* and *ybgC*, the Lpt and Tol-Pal systems are found in both atypical diderms and true diderms-LPS but to a different extent. Indeed, both systems are present in atypical diderms, albeit only in a largely reduced form, whereas in true diderms-LPS, 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) true diderms-LPS genomes, *lptD* and *lptE* on one side, *pal* and *tolB* on the other side, whereas all four genes are never found in atypical diderms genomes. Regarding *tolA* and *ybgF*, they may or may not be exclusive to true diderms-LPS, depending on the biological reality of their scarce occurrence in some organisms belonging to atypical diderms (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 in the folder Outer_membrane) are sisters to Bacteroidetes (member of true diderms-LPS) 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 in the folder Trees). Therefore, provided they are not the product of genome contamination [100], these genes are unlikely to have been vertically inherited.

From a functional point of view, the genes retained by atypical diderms for the Lpt system (*lptA*, *lptB*, *lptC*, *lptF* and *lptG*) are involved in the transport of the lipopolysaccharides from the cytoplasm to the outer membrane and thus are not directly associated to the outer membrane itself, contrarily to *lptD* and *lptE*, which form a complex at the outer membrane that may serve as the recognition site for the lipopolysaccharides [101]. Similarly, for the Tol-Pal system, atypical diderms genomes lack *pal* and *tolB*, two genes encoding proteins located in the periplasm and therefore directly associated to the outer membrane [102,103]. Overall, the Lpt and Tol-Pal systems in atypical diderms are thus restricted to components that might have a function in the absence of an outer membrane.

Remarkably, 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 in the folder Outer_membrane sub-folder synteny_output). As all these lineages belong to the true diderms-LPS, we cannot exclude that the conservation of the Tol-Pal cluster appears patchier than it really is, owing to uneven levels of genome assembly. Regarding the genes of the Lpt system, they 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 in the folder Outer_membrane sub-folder synteny_output). Nevertheless, as the synteny of the genes of both Lpt and Tol-Pal systems was only studied in the 85 genomes of our phylogenomic tree, we may have missed non-Proteobacterial genomes in which some of the *lpt* genes are indeed clustered, as reported in the recent study of Taib et al. [17].

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 [104] makes it elusive to decide 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 have been a monoderm bacterium featuring a complete 17-gene *dcw* cluster, two genes more than in the model *E. coli* cluster. This result was also supported by the recent study of [105], in which the authors found 146 protein families that formed a predicted core for the metabolic network of the LBCA. From these families, phylogenetic trees were produced and the divergence of the modern genomes from the root to the tips was analysed. It appears that the Clostridia (a class of Firmicutes) are the least diverged of the modern genomes and thus the first lineage to diverge from the predicted LBCA were similar to the modern Clostridia. Based on these results, the authors suggested that the LBCA could have been a monoderm bacteria.

As diderm bacteria are not monophyletic, whatever the root used 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 [68]. Indeed, acquiring an outer membrane 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 "diderm-first" hypothesis, in which the outer membrane is an ancestral feature having evolved only once in the LBCA and later lost in monoderms [5,17,25,68]. However, such an observation can be made in Archaea, where most of the studied organisms have a monoderm cell wall featuring a S-layer and/or pseudomurein, methanochondroitin and protein sheaths. In this context, some diderm Archaea have been reported in different distant phyla, like the Crenarchaeon *Ignicoccus hospitalis*, the Euryarchaeon ARMAN (Archaeal Richmond Mine Acidophilic Nanoorganisms) or the *Candidatus Altiarchaeum hamiconnexum* (SM1 Euryarchaeum) in the DPANN group [106]. Although it has not been proved that a monoderm cell wall is the general architecture in Archaea, the discovery of diderm Archaea within different phyla shows that acquisition of a second membrane has occurred multiple times during archaeal evolution. Moreover, 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 [5,68] 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. [25]. The latter study featured a reconciliation tree and infered the diderm state of the LBCA based on the genes involved in lipopolysaccharides synthesis and the flagellar subunits, notably PilQ, which is part of the Type IV pili. While the approach of Coleman and co-workers was also model-based, it differed 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 studied the underlying gene distribution patterns to corroborate our inference. It is of note that the Type IV pili is also present in monoderm bacteria [107], 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 lipopolysaccharides), we identified 16 candidates: BamA, which is part of a complex assembling the proteins in the outer membrane [91], LolB, which is part of the proteins inserting the lipopolysaccharides in the outer membrane [3], the Lpt proteins, which serve as a transport chain from the inner, i.e., cytoplasmic [108], membrane (IM) to the outer membrane [92], and the Tol-Pal system, the exact function of which is still unknown but important to the integrity of the outer membrane [93]. Then, we studied the distribution of the 16 corresponding genes in 903 broadly sampled bacterial genomes. Four recurring patterns of outer membrane gene distribution were identified (Figure 3B): (1) atypical diderms (Deinococcus-Thermus, Cyanobacteria and diderm Firmicutes), (2) monoderm Terrabacteria (Actinobacteria, Chloroflexi and monoderm Firmicutes), (3) true diderms-LPS, and (4) Thermotogae. Thermotogae have chimerical genomes [70] and are likely derived with respect to other bacteria; thus, their cell-wall architecture is of secondary origin. Therefore, 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 positioned deeply within Actinobacteria [109], which again implies a secondary origin for their peculiar cell-wall architecture.

From these patterns, it appears that even monoderm Terrabacteria share some genes involved with the outer membrane despite their lack of an outer membrane. It implies that these genes provide at best circumstantial evidence concerning the presence or the absence of an outer membrane. Thus, solely relying on their detection to infer the presence of an outer membrane would be hazardous. In the study of Coleman et al. [25], the authors build upon two types of genes to justify their inference of a diderm LBCA: the genes involved with the lipopolysaccharides synthesis and the genes involved with the pili type IV. However, our results show that the mere presence of lipopolysaccharides genes is an unreliable feature to infer the presence of an outer membrane, given that even monoderm bacteria can carry some of them. Similarly, the study of [107] showed that the type IV pili is not exclusive to the diderm bacteria. Therefore, the inference of a diderm LBCA by Coleman et al. was based on genes that only provide ambiguous evidence for the outer membrane.

Pattern 2 shows that Chloroflexi share the same gene distribution as monoderm Terrabacteria, despite being mostly considered as diderms (3 out of 4 genomes) in our reconstruction of the cell wall. Currently, there is still debate on whether Chloroflexi are monoderm or diderm organisms, microscopical observations having been inconclusive so far but hinting at the presence of an outer membrane in some Chloroflexi [87,88]. The fact that they share the same outer membrane gene distribution pattern as monoderm Terrabacteria is a clue in favour 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, despite that unwarranted handicap, we still recovered the LBCA as a monoderm bacterium. In our opinion, this result can be taken as more 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 outer membrane (i.e., linked to the outer membrane), whereas pattern 1 only includes proteins indirectly involved with the outer membrane (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 outer membrane and located in the cytoplasm. Although we know (some of) the outer membrane pathways functioning in true diderms-LPS, for atypical diderms, we only identified the common parts between their pathways and the true diderms-LPS pathways. The rest of the true diderms-LPS pathways should have an equivalent in the atypical diderms 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, as some of the functions performed by the genes present in pattern 3 (true diderms-LPS) but absent in pattern 1 (atypical diderms) should be carried out in one way or another (e.g., the maintenance of the outer membrane or the outer membrane invagination during cell division) [110]. In this case, the common base would be the partial Lpt and Tol-Pal systems, upon which at least two different systems for handling the outer membrane would have built in the true diderms-LPS and (all or some) atypical diderms. 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).

Assuming a monoderm LBCA, single-gene trees might suggest that some outer membrane genes found in atypical diderms (e.g., LptF and LptG) stem from horizontal transfer from true diderms-LPS, 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 at best. Based on a parsimony reasoning, the exclusivity of pattern 3 to true diderms-LPS and the fact that it is shared between all of them suggest, alongside their well-supported branch in our phylogenomic tree, the monophyly of the true diderms-LPS group. Indeed, if all current genomes of a group have the same subset of genes, the LCA of the group is likely to have had these genes (in a form or another). If correct, the bacterial root cannot lie within true diderms-LPS and as already mentioned, a root on (or within) Terrabacteria implies that the diderm cell-wall architecture appeared at least on two separate occasions. The latter inference is necessary to account for diderms other than true diderms-LPS in Firmicutes, Cyanobacteria, Chloroflexi and Deinococcus-Thermus, which then raises the issue of how the lipopolysaccharides are transported from the IM to the outer membrane for these atypical diderms nested within Terrabacteria. Indeed, they do not share the same Lpt system as true diderms-LPS as 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 [62,111]) to have been described to feature a monoderm cell-wall architecture [112]. 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 [64,65].

However, in [25,113], the CPR subtree is found within the Terrabacteria with strong support. Consequently, depending on the accepted topology, the CPR could either be 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. As it is shared by both monoderm and diderm organisms, the *dcw* cluster does not give 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 true diderms-LPS 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 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 true diderms-LPS 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 outer membrane in the lineage leading to true diderms-LPS, but such a scenario exists for the appearance of diderms in Firmicutes: it is the failed endospore origin [11,13,15,114]. 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 sporulation, the prespore engulfed in the bacterial mother cell has 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 outer membrane. Although not yet a diderm-LPS architecture, a cortex-less spore could be a starting point for the emergence of diderm bacteria in the specific case of Firmicutes. In 2016, Tocheva [14] 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 outer membrane in atypical diderms other than Firmicutes, we have already mentioned that Chloroflexi might be monoderms, based on their shared pattern (pattern 2) with monoderm Terrabacteria. This leaves us with Cyanobacteria and Deinococcus-Thermus, along with the large true diderms-LPS group. Because pattern 3 looks like a complexification of pattern 1, the origin of didermia in true diderms-LPS might come from one of these atypical diderms phyla by horizontal gene transfer of outer membrane genes, followed by complexification in an ancestor of true diderms-LPS. Alternatively, true diderms-LPS ancestors might have transferred outer membrane genes to distinct ancestors of atypical diderms 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. However, it is of note that the failed sporulation scenario is compatible with the inferences of [105].

5. Conclusions

Our results suggest that the LBCA might have been, against familiar parsimony reasoning, a monoderm bacteria with a thick peptidoglycan layer, which is also supported by the recent study of [105]. 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 outer membrane 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 methods would require a more accurate description of the cell-wall architecture of the extant organisms, notably the presence or absence of the lipopolysaccharides, an information which, in our experience, is often lacking. When available, it is concentrated in the older literature, when organisms were cultivated and characterized before being sequenced, in contrast to the numerous candidate bacterial phyla that populate recent phylogenomic trees [66,67]. Nevertheless, even older genomes do not guarantee an exploitable description, like *Rivularia* sp. (Table S2: 38). Moreover, we observe that some outer membrane genes involved with the precursors of lipopolysaccharides synthesis are also present in genomes of bacteria that does not have lipopolysaccharides on their outer membrane (or even an outer membrane), thus relying solely on the presence of specific genes to determine the presence or absence of lipopolysaccharides is not adequate.

One could argue that the current study does not concern the LBCA but the LCA of cultured (and characterized) Bacteria and we would not completely disagree as we ourselves see it as a proof of concept of the method. A follow-up would be interesting to carry out once accurate information for the cell wall of more phyla are available. In such a follow-up study, it could be interesting to add supplementary genomes such as the "rogue" lineages (e.g., Aquificae and Thermotogae), additional phyla of uncertain phylogenetic position (e.g., basal Terrabacteria), completely new genomes (e.g., CPR) or even an outgroup to root the tree (e.g., Archaea). Aquifex being "just another" group of diderms and Thermotogae being a chimera with a specific diderm architecture, their inclusion would only provide a limited amount of information compared to considering additional Terrabacteria genomes or representatives of the recently discovered CPR. Regarding the difficulty to place such lineages accurately in a phylogenomic tree, it could be overcome by adding genes that are not single copy but at the expense of more work to sort out the orthologous copies. The CPR group would be a particularly welcome addition, provided a useful description of their cell wall could be obtained. Concerning the addition of an outgroup, the question of how it will be used should be answered first: will it be included in the cell-wall reconstruction analyses or will it only be used to root the bacterial subtree. Indeed, if it is not used for reconstruction, any slow evolving fully sequenced Archaea would be usable. On the other hand, if we are interested in reconstructing their cell wall too, we would need to select them very carefully, just as we did for Bacteria. In this respect, the cell-wall diversity of Archaea is as complicated as the bacterial one, if not more, which would add another level of difficulty, and thus uncertainty, to the inferred results.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes13020376/s1. Figure S1: Unrooted phylogenomic tree of the bacterial domain based on a supermatrix concatenating 117 single-copy orthologous genes chosen for their broad conservation across Bacteria. Figure S2: Unrooted phylogenomic tree of the bacterial domain based on a supermatrix concatenating 117 single-copy orthologous genes chosen for their broad conservation across Bacteria. Figure S3: Evolution of the log likelihood of six PhyloBayes MCMC chains running under the CAT+GTR+Γ model of sequence evolution. 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. Figure S5: Trees inferred by the six individual MCMC chains running under the CAT+GTR+Γ model of sequence evolution. 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. Figure S7: Posterior probabilities for a monoderm LBCA according to five different models and six possible roots for the bacterial domain. 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. 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. Figure S11: Posterior probabilities for the peptidoglycan (PG) and membrane traits in the LCA of four bacterial groups. Figure S12: MraY tree inferred using RAxML under the LG+F+Γ model of sequence evolution. Figure S13: Phylogenomic tree based on a supermatrix of 85 species × 4571 unambiguously aligned amino-acid positions (8.47% missing character states) using 15 of the *dcw* cluster genes. Figure S14: Unrooted TolA tree inferred using RAxML under the LG+F+Γ model. Figure S15: Unrooted YbgF tree inferred using RAxML under the LG+F+Γ model. Figure S16: Schema of the MySQL database used by the synteny tool. Table S1: List of the 117 genes used for the phylogenomic tree of Figure 1. Table S2: List of references used to determine the cell-wall architecture for the 85 representative organisms of Figure 1. Table S3: Details of the data given to BayesTraits for the ancestral trait reconstruction. Table S4: Number of POTRA domains predicted by InterProScan in the majority of the sequences composing each orthologous group (OG) identified as a member of the Omp85/TpsB family. Table S5: Results of the cross-validation procedure comparing four different models of sequence evolution available in PhyloBayes MPI. Table S6: Possible roots for the bacterial domain reported in the phylogenomic literature since 2006.

Author Contributions: R.R.L. 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. E.S., F.K. and D.B. conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper and approved the final draft. V.L. prepared figures, reviewed drafts of the paper and approved the final draft. A.P. developed a specific version of ProCARs for this study, reviewed drafts of the paper and approved the final draft. D.S. provided technical support for the high-performance computing cluster, reviewed drafts of the paper and approved the final draft. P.C. reviewed drafts of the paper and approved the final draft. All authors have read and agreed to the published version of the manuscript.

Funding: 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, partially funded by 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). 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).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and analysed for this study can be found in the FigShare repository available here: https://doi.org/10.6084/m9.figshare.14932386.v2 (16 February 2022). Similarly, the database schema and corresponding table dump are available at: https://doi.org/10.6084/m9.figshare.17102651.v1 (16 February 2022).

Acknowledgments: 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.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schleifer, K.H.; Kandler, O. Peptidoglycan Types of Bacterial Cell Walls and Their Taxonomic Implications. *Bacteriol. Rev.* **1972**, *36*, 407–477. [CrossRef] [PubMed]
2. Coico, R. Gram Staining. *Curr. Protoc. Microbiol.* **2006**, A-3C.
3. Silhavy, T.J.; Kahne, D.; Walker, S. The Bacterial Cell Envelope. *Cold Spring Harb. Perspect. Biol.* **2010**, *2*, a000414. [CrossRef] [PubMed]
4. Woese, C.R. Bacterial Evolution. *Microbiol. Rev.* **1987**, *51*, 221–271. [CrossRef]
5. Megrian, D.; Taib, N.; Witwinowski, J.; Beloin, C.; Gribaldo, S. One or Two Membranes? Diderm Firmicutes Challenge the Gram-Positive/Gram-Negative Divide. *Mol. Microbiol.* **2020**, *113*, 659–671. [CrossRef] [PubMed]
6. Cavalier-Smith, T. The Origin of Eukaryote and Archaebacterial Cells. *Ann. N. Y. Acad. Sci.* **1987**, *503*, 17–54. [CrossRef] [PubMed]
7. Cavalier-Smith, T. Deep Phylogeny, Ancestral Groups and the Four Ages of Life. *Philos. Trans. R. Soc. B Biol. Sci.* **2010**, *365*, 111–132. [CrossRef]
8. Koch, A.L. Were Gram-Positive Rods the First Bacteria? *TRENDS Microbiol.* **2003**, *11*, 166–170. [CrossRef]

9. Sutcliffe, I.C. A Phylum Level Perspective on Bacterial Cell Envelope Architecture. *Trends Microbiol.* **2010**, *18*, 464–470. [CrossRef]
10. 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.* **2011**, *100*, 171–182. [CrossRef]
11. Errington, J. L-Form Bacteria, Cell Walls and the Origins of Life. *Open Biol.* **2013**, *3*, 120143. [CrossRef]
12. Lake, J.A. Evidence for an Early Prokaryotic Endosymbiosis. *Nature* **2009**, *460*, 967–971. [CrossRef] [PubMed]
13. Tocheva, E.I.; Matson, E.G.; Morris, D.M.; Moussavi, F.; Leadbetter, J.R.; Jensen, G.J. Peptidoglycan Remodeling and Conversion of an Inner Membrane into an Outer Membrane during Sporulation. *Cell* **2011**, *146*, 799–812. [CrossRef]
14. Tocheva, E.I.; Ortega, D.R.; Jensen, G.J. Sporulation, Bacterial Cell Envelopes and the Origin of Life. *Nat. Rev. Microbiol.* **2016**, *14*, 535–542. [CrossRef]
15. Vollmer, W. Bacterial Outer Membrane Evolution via Sporulation? *Nat. Chem. Biol.* **2011**, *8*, 14–18. [CrossRef]
16. Antunes, L.C.S.; Poppleton, D.; Klingl, A.; Criscuolo, A.; Dupuy, B.; Brochier-Armanet, C.; Beloin, C.; Gribaldo, S. Phylogenomic Analysis Supports the Ancestral Presence of LPS-Outer Membranes in the Firmicutes. *eLife* **2016**, *5*, 1–21. [CrossRef]
17. Taib, N.; Megrian, D.; Witwinowski, J.; Adam, P.; Poppleton, D.; Borrel, G.; Beloin, C.; Gribaldo, S. Genome-Wide Analysis of the Firmicutes Illuminates the Diderm/Monoderm Transition. *Nat. Ecol. Evol.* **2020**, *4*, 1661–1672. [CrossRef]
18. Mingorance, J.; Tamames, J. The Bacterial Dcw Gene Cluster: An Island in the Genome? In *Molecules in Time and Space*; Springer: Boston, MA, USA, 2004; pp. 249–271, ISBN 978-0-306-48579-4.
19. Tamames, J. Evolution of Gene Order Conservation in Prokaryotes. *Genome Biol.* **2001**, *2*, 1–11. [CrossRef]
20. Real, G.; Henriques, A.O. Localization of the Bacillus Subtilis MurB Gene within the Dcw Cluster Is Important for Growth and Sporulation. *J. Bacteriol.* **2006**, *188*, 1721–1732. [CrossRef] [PubMed]
21. Barloy-Hubler, F.; Lelaure, V.; Galibert, F. Ribosomal Protein Gene Cluster Analysis in Eubacterium Genomics: Homology between Sinorhizobium Meliloti Strain 1021 and Bacillus Subtilis. *Nucleic Acids Res.* **2001**, *29*, 2747–2756. [CrossRef] [PubMed]
22. Nikolaichik, Y.A.; Donachie, W.D. Conservation of Gene Order amongst Cell Wall and Cell Division Genes in Eubacteria, and Ribosomal Genes in Eubacteria and Eukaryotic Organelles. *Genetica* **2000**, *108*, 1–7. [CrossRef]
23. Eraso, J.M.; Markillie, L.M.; Mitchell, H.D.; Taylor, R.C.; Orr, G.; Margolin, W. The Highly Conserved MraZ Protein Is a Transcriptional Regulator in Escherichia Coli. *J. Bacteriol.* **2014**, *196*, 2053–2066. [CrossRef]
24. Pilhofer, M.; Rappl, K.; Eckl, C.; Bauer, A.P.; Ludwig, W.; Schleifer, K.H.; Petroni, G. 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.* **2008**, *190*, 3192–3202. [CrossRef]
25. Coleman, G.A.; Davín, A.A.; Mahendrarajah, T.A.; Szánthó, L.L.; Spang, A.; Hugenholtz, P.; Szöllősi, G.J.; Williams, T.A. A Rooted Phylogeny Resolves Early Bacterial Evolution. *Science* **2021**, *372*. [CrossRef] [PubMed]
26. Kersey, P.J.; Allen, J.E.; Christensen, M.; Davis, P.; Falin, L.J.; Grabmueller, C.; Hughes, D.S.T.; Humphrey, J.; Kerhornou, A.; Khobova, J.; et al. Ensembl Genomes 2013: Scaling up Access to Genome-Wide Data. *Nucleic Acids Res.* **2014**, *42*, D546–D552. [CrossRef] [PubMed]
27. Moreno-Hagelsieb, G.; Wang, Z.; Walsh, S.; ElSherbiny, A. Phylogenomic Clustering for Selecting Non-Redundant Genomes for Comparative Genomics. *Bioinformatics* **2013**, *29*, 947–949. [CrossRef] [PubMed]
28. Léonard, R.R.; Leleu, M.; Vlierberghe, M.V.; Cornet, L.; Kerff, F.; Baurain, D. ToRQuEMaDA: Tool for Retrieving Queried Eubacteria, Metadata and Dereplicating Assemblies. *PeerJ* **2021**, *9*, e11348. [CrossRef]
29. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]
30. Sayers, E.W.; Barrett, T.; Benson, D.A.; Bolton, E.; Bryant, S.H.; Canese, K.; Chetvernin, V.; Church, D.M.; DiCuccio, M.; Federhen, S.; et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2011**, *39*, D38–D51. [CrossRef]
31. R Core Team, Rf. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
32. Edgar, R.C. Search and Clustering Orders of Magnitude Faster than BLAST. *Bioinformatics* **2010**, *26*, 2460–2461. [CrossRef]
33. Li, L.; Stoeckert, C.J.; Roos, D.S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* **2003**, *13*, 2178–2189. [CrossRef]
34. Van Vlierberghe, M.; Philippe, H.; Baurain, D. Broadly Sampled Orthologous Groups of Eukaryotic Proteins for the Phylogenetic Study of Plastid-Bearing Lineages. *BMC Res. Notes* **2021**, *14*, 143. [CrossRef] [PubMed]
35. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
36. Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.* **2000**, *17*, 540–552. [CrossRef] [PubMed]
37. Roure, B.; Rodriguez-Ezpeleta, N.; Philippe, H. SCaFoS: A Tool for Selection, Concatenation and Fusion of Sequences for Phylogenomics. *BMC Evol. Biol.* **2007**, *7* Suppl. S1, S2. [CrossRef]
38. Lartillot, N.; Rodrigue, N.; Stubbs, D.; Richer, J. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Syst. Biol.* **2013**, *62*, 611–615. [CrossRef]

39. Lartillot, N.; Philippe, H. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Mol. Biol. Evol.* **2004**, *21*, 1095–1109. [CrossRef]
40. Lartillot, N.; Philippe, H. Computing Bayes Factors Using Thermodynamic Integration. *Syst. Biol.* **2006**, *55*, 195–207. [CrossRef]
41. Lartillot, N.; Brinkmann, H.; Philippe, H. Suppression of Long-Branch Attraction Artefacts in the Animal Phylogeny Using a Site-Heterogeneous Model. *BMC Evol. Biol.* **2007**, *7* Suppl. S1, S4. [CrossRef]
42. Lartillot, N.; Lepage, T.; Blanquart, S. PhyloBayes 3: A Bayesian Software Package for Phylogenetic Reconstruction and Molecular Dating. *Bioinformatics* **2009**, *25*, 2286–2288. [CrossRef] [PubMed]
43. Letunic, I.; Bork, P. Interactive Tree Of Life (ITOL) v5: An Online Tool for Phylogenetic Tree Display and Annotation. *Nucleic Acids Res.* **2021**, *49*, W293–W296. [CrossRef] [PubMed]
44. De Vienne, D.M.; Ollier, S.; Aguileta, G. Phylo-MCOA: A Fast and Efficient Method to Detect Outlier Genes and Species in Phylogenomics Using Multiple Co-Inertia Analysis. *Mol. Biol. Evol.* **2012**, *29*, 1587–1598. [CrossRef] [PubMed]
45. Stamatakis, A. RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef] [PubMed]
46. Pagel, M.; Meade, A.; Barker, D. Bayesian Estimation of Ancestral Character States on Phylogenies. *Syst. Biol.* **2004**, *53*, 673–684. [CrossRef] [PubMed]
47. Pagel, M.; Meade, A. Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. *Am. Nat.* **2015**, *167*, 808–825. [CrossRef]
48. Meade, A.; Pagel, M. BayesTraits V3. 0.1. 2017. Available online: http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.1/BayesTraitsV3.0.1.html(accessed on 9 February 2022).
49. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195. [CrossRef]
50. Mistry, J.; Finn, R.D.; Eddy, S.R.; Bateman, A.; Punta, M. Challenges in Homology Search: HMMER3 and Convergent Evolution of Coiled-Coil Regions. *Nucleic Acids Res.* **2013**, *41*, e121-e121. [CrossRef]
51. Perrin, A.; Varré, J.-S.; Blanquart, S.; Ouangraoua, A. ProCARs: Progressive Reconstruction of Ancestral Gene Orders. *BMC Genomics* **2015**, *16 Suppl 5*, S6. [CrossRef]
52. Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-Scale Protein Function Classification. *Bioinformatics* **2014**, *30*, 1236–1240. [CrossRef]
53. Sánchez-Pulido, L.; Devos, D.; Genevrois, S.; Vicente, M.; Valencia, A. POTRA: A Conserved Domain in the FtsQ Family and a Class of β-Barrel Outer Membrane Proteins. *Trends Biochem. Sci.* **2003**, *28*, 523–526. [CrossRef] [PubMed]
54. Heinz, E.; Lithgow, T. A Comprehensive Analysis of the Omp85/TpsB Protein Superfamily Structural Diversity, Taxonomic Occurrence, and Evolution. *Front. Microbiol.* **2014**, *5*, 370. [CrossRef]
55. Delsuc, F.; Brinkmann, H.; Philippe, H. Phylogenomics and the Reconstruction of the Tree of Life. *Nat. Rev. Genet.* **2005**, *6*, 361–375. [CrossRef]
56. Philippe, H.; de Vienne, D.M.; Ranwez, V.; Roure, B.; Baurain, D.; Delsuc, F. Pitfalls in Supermatrix Phylogenomics. *Eur. J. Taxon.* **2017**, *283*, 1–25. [CrossRef]
57. 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*; Anisimova, M., Ed.; Humana: New York, NY, USA, 2019; pp. 211–239.
58. Battistuzzi, F.U.; Hedges, S.B. A Major Clade of Prokaryotes with Ancient Adaptations to Life on Land. *Mol. Biol. Evol.* **2009**, *26*, 335–343. [CrossRef]
59. Wu, D.; Hugenholtz, P.; Mavromatis, K.; Pukall, R.; Dalin, E.; Ivanova, N.N.; Kunin, V.; Goodwin, L.; Wu, M.; Tindall, B.J.; et al. A Phylogeny-Driven Genomic Encyclopaedia of Bacteria and Archaea. *Nature* **2009**, *462*, 1056–1060. [CrossRef] [PubMed]
60. Yutin, N.; Puigbo, P.; Koonin, E.V.; Wolf, Y.I. Phylogenomics of Prokaryotic Ribosomal Proteins. *Curr. Sci.* **2012**, *101*, 1435–1439. [CrossRef] [PubMed]
61. Lasek-nesselquist, E.; Gogarten, J.P. The Effects of Model Choice and Mitigating Bias on the Ribosomal Tree of Life. *Mol. Phylogenet. Evol.* **2013**, *69*, 17–38. [CrossRef]
62. Rinke, C.; Schwientek, P.; Sczyrba, A.; Ivanova, N.N.; Anderson, I.J.; Cheng, J.-F.; Darling, A.E.; Malfatti, S.; Swan, B.K.; Gies, E.A.; et al. Insights into the Phylogeny and Coding Potential of Microbial Dark Matter. *Nature* **2013**, *499*, 431–437. [CrossRef] [PubMed]
63. 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. USA* **2015**, *112*, 6670–6675. [CrossRef] [PubMed]
64. Hug, L.A.; Baker, B.J.; Anantharaman, K.; Brown, C.T.; Probst, A.J.; Castelle, C.J.; Butterfield, C.N.; Hernsdorf, A.W.; Amano, Y.; Kotaro, I.; et al. A New View of the Tree of Life. *Nat. Microbiol.* **2016**, *1*, 16048. [CrossRef]
65. Castelle, C.J.; Banfield, J.F. Major New Microbial Groups Expand Diversity and Alter Our Understanding of the Tree of Life. *Cell* **2018**, *172*, 1181–1197. [CrossRef]
66. Parks, D.H.; Chuvochina, M.; Waite, D.W.; Rinke, C.; Skarshewski, A.; Chaumeil, P.-A.; Hugenholtz, P. A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life. *Nat. Biotechnol.* **2018**, *36*, 996–1004. [CrossRef]
67. Zhu, Q.; Mai, U.; Pfeiffer, W.; Janssen, S.; Asnicar, F.; Sanders, J.G.; Belda-ferre, P.; Al-ghalith, G.A.; Kopylova, E.; Mcdonald, D.; et al. Phylogenomics of 10,575 Genomes Reveals Evolutionary Proximity between Domains Bacteria and Archaea. *Nat. Commun.* **2019**, *10*, 1–14. [CrossRef]

68. Cavalier-Smith, T.; Ema, E.; Chao, Y. Multidomain Ribosomal Protein Trees and the Planctobacterial Origin of Neomura (Eukaryotes, Archaebacteria). *Protoplasma* **2020**, 1–133. [CrossRef] [PubMed]
69. Cavalier-Smith, T. Rooting the Tree of Life by Transition Analyses. *Biol. Direct* **2006**, *1*, 19. [CrossRef]
70. Zhaxybayeva, O.; Swithers, K.S.; Lapierre, P.; Fournier, G.P.; Bickhart, D.M.; DeBoy, R.T.; Nelson, K.E.; Nesbø, C.L.; Doolittle, W.F.; Gogarten, J.P.; et al. On the Chimeric Nature, Thermophilic Origin, and Phylogenetic Placement of the Thermotogales. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 5865–5870. [CrossRef]
71. 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.* **2012**, *2*, 1–14. [CrossRef]
72. 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.* **2013**, *5*, 2478–2497. [CrossRef] [PubMed]
73. Rachel, R.; Wildhaber, I.; Stetter, K.O.; Baumeister, W. The Structure of the Surface Protein of Thermotoga Maritima. In *Crystalline Bacterial Cell Surface Layers*; Springer: Berlin/Heidelberg, Germany, 1988; pp. 83–86.
74. Rachel, R.; Engel, A.M.; Huber, R.; Stetter, K.-O.; Baumeister, W. A Porin-Type Protein Is the Main Constituent of the Cell Envelope of the Ancestral Eubacterium Thermotoga Maritima. *FEBS Lett.* **1990**, *262*, 64–68. [CrossRef]
75. Bapteste, E.; Boucher, Y.; Leigh, J.; Doolittle, W.F. Phylogenetic Reconstruction and Lateral Gene Transfer. *Trends Microbiol.* **2004**, *12*, 406–411. [CrossRef]
76. Mira, A.; Pushker, R.; Legault, B.A.; Moreira, D.; Rodríguez-Valera, F. Evolutionary Relationships of Fusobacterium Nucleatum Based on Phylogenetic Analysis and Comparative Genomics. *BMC Evol. Biol.* **2004**, *4*, 50. [CrossRef]
77. Beiko, R.G.; Harlow, T.J.; Ragan, M. a Highways of Gene Sharing in Prokaryotes. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 14332–14337. [CrossRef]
78. Koonin, E.V. Orthologs, Paralogs, and Evolutionary Genomics. *Annu. Rev. Genet.* **2005**, *39*, 309–338. [CrossRef]
79. Koonin, E.V. Horizontal Gene Transfer: Essentiality and Evolvability in Prokaryotes, and Roles in Evolutionary Transitions. *F1000Research* **2016**, *5*, F1000 Faculty Rev-1805. [CrossRef]
80. Boussau, B.; Guéguen, L.; Gouy, M. Accounting for Horizontal Gene Transfers Explains Conflicting Hypotheses Regarding the Position of Aquificales in the Phylogeny of Bacteria. *BMC Evol. Biol.* **2008**, *8*, 1–18. [CrossRef]
81. Philippe, H.; Brinkmann, H.; Lavrov, D.V.; Littlewood, D.T.J.; Manuel, M.; Wörheide, G.; Baurain, D. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biol.* **2011**, *9*. [CrossRef] [PubMed]
82. Gouy, R.; Baurain, D.; Philippe, H. Rooting the Tree of Life: The Phylogenetic Jury Is Still Out. *Philos. Trans. R. Soc. B Biol. Sci.* **2015**, *370*, 20140329. [CrossRef] [PubMed]
83. Rivas-Marín, E.; Canosa, I.; Devos, D.P. Evolutionary Cell Biology of Division Mode in the Bacterial Planctomycetes-Verrucomicrobia- Chlamydiae Superphylum. *Front. Microbiol.* **2016**, *7*. [CrossRef] [PubMed]
84. Cruz-Morales, P.; Orellana, C.A.; Moutafis, G.; Moonen, G.; Rincon, G.; Nielsen, L.K.; Marcellin, E. Revisiting the Evolution and Taxonomy of Clostridia, a Phylogenomic Update. *Genome Biol. Evol.* **2019**, *11*, 2035–2044. [CrossRef]
85. Mavromatis, K.; Ivanova, N.; Anderson, I.; Lykidis, A.; Hooper, S.D.; Sun, H.; Kunin, V.; Lapidus, A.; Hugenholtz, P.; Patel, B.; et al. Genome Analysis of the Anaerobic Thermohalophilic Bacterium Halothermothrix Orenii. *PLoS ONE* **2009**, *4*, e4192. [CrossRef]
86. Kivistö, A.T.; Karp, M.T. Halophilic Anaerobic Fermentative Bacteria. *J. Biotechnol.* **2011**, *152*, 114–124. [CrossRef] [PubMed]
87. Sutcliffe, I.C. Cell Envelope Architecture in the Chloroflexi: A Shifting Frontline in a Phylogenetic Turf War. *Environ. Microbiol.* **2011**, *13*, 279–282. [CrossRef] [PubMed]
88. Gaisin, V.A.; Kooger, R.; Grouzdev, D.S.; Gorlenko, V.M.; Pilhofer, M. Cryo-Electron Tomography Reveals the Complex Ultrastructural Organization of Multicellular Filamentous Chloroflexota (Chloroflexi) Bacteria. *Front. Microbiol.* **2020**, *11*, 1373. [CrossRef] [PubMed]
89. Gilks, W.R.; Richardson, S.; Spiegelhalter, D. *Markov Chain Monte Carlo in Practice*; CRC Press: Boca Raton, FL, USA, 1995; ISBN 978-1-4822-1497-0.
90. Yutin, N.; Galperin, M.Y. A Genomic Update on Clostridial Phylogeny: Gram-Negative Spore Formers and Other Misplaced Clostridia. *Environ. Microbiol.* **2013**, *15*, 2631–2641. [CrossRef] [PubMed]
91. Hagan, C.L.; Silhavy, T.J.; Kahne, D. β-Barrel Membrane Protein Assembly by the Bam Complex. *Annu. Rev. Biochem.* **2011**, *80*, 189–210. [CrossRef]
92. 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.* **2011**, *404*, 1093–1098. [CrossRef]
93. 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. *Mol. Microbiol.* **2002**, *44*, 695–708. [CrossRef]
94. Ranjit, C.; Noll, K.M. Distension of the Toga of Thermotoga Maritima Involves Continued Growth of the Outer Envelope as Cells Enter the Stationary Phase. *FEMS Microbiol. Lett.* **2016**, 363.
95. Bernard, G.; Chan, C.X.; Ragan, M.A. Alignment-Free Microbial Phylogenomics under Scenarios of Sequence Divergence, Genome Rearrangement and Lateral Genetic Transfer. *Sci. Rep.* **2016**, *6*, 1–12. [CrossRef]
96. Yu, J.; Lu, L. BamA Is a Pivotal Protein in Cell Envelope Synthesis and Cell Division in Deinococcus Radiodurans. *Biochim. Biophys. Acta BBA-Biomembr.* **2019**, *1861*, 1365–1374. [CrossRef]
97. Voulhoux, R.; Bos, M.P.; Geurtsen, J.; Mols, M.; Tommassen, J. Role of a Highly Conserved Bacterial Protein in Outer Membrane Protein Assembly. *Science* **2003**, *299*, 262–265. [CrossRef]

98. Gully, D.; Bouveret, E. A Protein Network for Phospholipid Synthesis Uncovered by a Variant of the Tandem Affinity Purification Method in Escherichia Coli. *Proteomics* **2006**, *6*, 282–293. [CrossRef]
99. Narita, S.; Tokuda, H. Biochemical Characterization of an ABC Transporter LptBFGC Complex Required for the Outer Membrane Sorting of Lipopolysaccharides. *FEBS Lett.* **2009**, *583*, 2160–2164. [CrossRef] [PubMed]
100. Cornet, L.; Meunier, L.; Van Vlierberghe, M.; Léonard, R.R.; Durieu, B.; Lara, Y.; Misztak, A.; Sirjacobs, D.; Javaux, E.J.; Philippe, H.; et al. Consensus Assessment of the Contamination Level of Publicly Available Cyanobacterial Genomes. *PLoS ONE* **2018**, *13*, e0200323. [CrossRef]
101. Wang, Z.; Xiang, Q.; Zhu, X.; Dong, H.; He, C.; Wang, H.; Zhang, Y.; Wang, W.; Dong, C. Structural and Functional Studies of Conserved Nucleotide-Binding Protein LptB in Lipopolysaccharide Transport. *Biochem. Biophys. Res. Commun.* **2014**, *452*, 443–449. [CrossRef] [PubMed]
102. Rigal, A.; Bouveret, E.; Lloubes, R.; Lazdunski, C.; Benedetti, H. The TolB Protein Interacts with the Porins of Escherichia Coli. *J. Bacteriol.* **1997**, *179*, 7274–7279. [CrossRef]
103. Ray, M.-C.; Germon, P.; Vianney, A.; Portalier, R.; Lazzaroni, J.C. Identification by Genetic Suppression OfEscherichia Coli TolB Residues Important for TolB-Pal Interaction. *J. Bacteriol.* **2000**, *182*, 821–824. [CrossRef] [PubMed]
104. Demoulin, C.F.; Lara, Y.J.; Cornet, L.; François, C.; Baurain, D.; Wilmotte, A.; Javaux, E.J. Cyanobacteria Evolution: Insight from the Fossil Record. *Free Radic. Biol. Med.* **2019**, *140*, 206–223. [CrossRef]
105. Xavier, J.C.; Gerhards, R.E.; Wimmer, J.L.; Brueckner, J.; Tria, F.D.; Martin, W.F. The Metabolic Network of the Last Bacterial Common Ancestor. *Commun. Biol.* **2021**, *4*, 1–10. [CrossRef]
106. Kuhn, A. The Bacterial Cell Wall and Membrane—A Treasure Chest for Antibiotic Targets. In *Bacterial Cell Walls and Membranes*; Kuhn, A., Ed.; Subcellular Biochemistry; Springer International Publishing: Cham, Switzerland, 2019; pp. 1–5. ISBN 978-3-030-18768-2.
107. Melville, S.; Craig, L. Type IV Pili in Gram-Positive Bacteria. *Microbiol. Mol. Biol. Rev.* **2013**, *77*, 323–341. [CrossRef]
108. 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.* **2016**, *8*, 325–326. [CrossRef]
109. Verma, M.; Lal, D.; Kaur, J.; Saxena, A.; Kaur, J.; Anand, S.; Lal, R. Phylogenetic Analyses of Phylum Actinobacteria Based on Whole Genome Sequences. *Res. Microbiol.* **2013**, *164*, 718–728. [CrossRef]
110. 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. USA* **2020**, *117*, 6777–6783. [CrossRef]
111. Parks, D.H.; Rinke, C.; Chuvochina, M.; Chaumeil, P.; Woodcroft, B.J.; Evans, P.N.; Hugenholtz, P.; Tyson, G.W. Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life. *Nat. Microbiol.* **2017**, *2*. [CrossRef]
112. Luef, B.; Frischkorn, K.R.; Wrighton, K.C.; Holman, H.-Y.N.; Birarda, G.; Thomas, B.C.; Singh, A.; Williams, K.H.; Siegerist, C.E.; Tringe, S.G.; et al. Diverse Uncultivated Ultra-Small Bacterial Cells in Groundwater. *Nat. Commun.* **2015**, *6*, 1–8. [CrossRef] [PubMed]
113. Martinez-Gutierrez, C.A.; Aylward, F.O. Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. *Mol. Biol. Evol.* **2021**, *38*, 5514–5527. [CrossRef] [PubMed]
114. Dawes, I.W. Sporulation in Evolution. *Mol. Cell. Asp. Microb. Evol. Camb. Univ. Press N. Y.* **1981**, 85–130.

4.2. De Novo Transcriptome Meta-Assembly of the Mixotrophic Freshwater Microalga Euglena gracilis

genes MDPI

Article

De Novo Transcriptome Meta-Assembly of the Mixotrophic Freshwater Microalga *Euglena gracilis*

Javier Cordoba [1], Emilie Perez [1,2], Mick Van Vlierberghe [2], Amandine R. Bertrand [2], Valérian Lupo [2], Pierre Cardol [1] and Denis Baurain [2,*]

[1] InBioS—PhytoSYSTEMS, Laboratoire de Génétique et Physiologie des Microalgues, ULiège, B-4000 Liège, Belgium; j.cordoba@outlook.es (J.C.); emilie.perez@alumni.uliege.be (E.P.); pierre.cardol@uliege.be (P.C.)

[2] InBioS—PhytoSYSTEMS, Unit of Eukaryotic Phylogenomics, ULiège, B-4000 Liège, Belgium; mvanvlierberghe@doct.uliege.be (M.V.V.); amandine.bertrand@doct.uliege.be (A.R.B.); valerian.lupo@doct.uliege.be (V.L.)

* Correspondence: denis.baurain@uliege.be; Tel.: +32-4-366-3864

Abstract: *Euglena gracilis* is a well-known photosynthetic microeukaryote considered as the product of a secondary endosymbiosis between a green alga and a phagotrophic unicellular belonging to the same eukaryotic phylum as the parasitic trypanosomatids. As its nuclear genome has proven difficult to sequence, reliable transcriptomes are important for functional studies. In this work, we assembled a new consensus transcriptome by combining sequencing reads from five independent studies. Based on a detailed comparison with two previously released transcriptomes, our consensus transcriptome appears to be the most complete so far. Remapping the reads on it allowed us to compare the expression of the transcripts across multiple culture conditions at once and to infer a functionally annotated network of co-expressed genes. Although the emergence of meaningful gene clusters indicates that some biological signal lies in gene expression levels, our analyses confirm that gene regulation in euglenozoans is not primarily controlled at the transcriptional level. Regarding the origin of *E. gracilis*, we observe a heavily mixed gene ancestry, as previously reported, and rule out sequence contamination as a possible explanation for these observations. Instead, they indicate that this complex alga has evolved through a convoluted process involving much more than two partners.

Keywords: transcriptome assembly; gene expression; transcriptional regulation; ontology network; co-expression network; taxonomic analysis; database contamination; kleptoplastidy

check for updates

Citation: Cordoba, J.; Perez, E.; Van Vlierberghe, M.; Bertrand, A.R.; Lupo, V.; Cardol, P.; Baurain, D. *De Novo* Transcriptome Meta-Assembly of the Mixotrophic Freshwater Microalga *Euglena gracilis*. *Genes* **2021**, *12*, 842. https://doi.org/10.3390/genes12060842

Academic Editor: Jose M. Requena

Received: 8 April 2021
Accepted: 27 May 2021
Published: 29 May 2021

1. Introduction

Euglena gracilis is a secondary green alga that can grow in a wide variety of environments. *E. gracilis* belongs to the euglenids, a monophyletic group of free-living, single-celled flagellates that inhabit aquatic ecosystems. Euglenids are distinguished mainly by their unique type of cell covering, the pellicle. The latter is a complex structure composed of proteinaceous strips covered by a cell membrane and underlain by the microtubule system and the cisternae of the endoplasmic reticulum [1]. Together, euglenids, symbiontids (free-living flagellates living in low-oxygen marine sediments), diplonemids (free-living marine flagellates) and kinetoplastids (free-living and parasitic flagellates, e.g., *Trypanosoma*) form the monophyletic group of Euglenozoa [2–5]. Euglenids are early diverged members of the *Euglenozoa* and distant relatives to the kinetoplastids [6]. Thus, analysing *E. gracilis* genomic information is a way to approach the evolution of parasitism, due to their common ancestry with kinetoplastids [7,8]. For example, it has been shown that many additional subunits of the mitochondrial respiratory chain previously considered exclusive to kinetoplastids are shared with *E. gracilis*, and therefore cannot be associated with the parasitic lifestyle [9]. Yet, it is worth mentioning that free-living bodonids (e.g., *Bodo saltans*) are better comparators for parasitism [10,11]. The relationship between euglenids and kinetoplastids has been first proposed by T. Cavalier-Smith based on ultrastructural similarities (e.g., "mitochondrial cristae shaped like a flattened disc with a narrow neck") [12], then supported by other lines of evidence, such as alignments of nuclear rRNA [13], the addition of a leader sequence to nuclear pre-mRNAs [14] and the presence of trypanothione reductase in *E. gracilis*, previously found only in kinetoplastids [15].

E. gracilis bears a complex plastid [16], derived from a green alga belonging to *Pyramimonadales*, and acquired by a free living phagotrophic eukaryovorous euglenid ancestor [17–19]. As the result of a so-called "secondary" endosymbiosis, this chloroplast is bound by three membranes, whereas primary plastids only have two membranes [20,21]. Whatever the specific event, endosymbiosis is accompanied by massive gene loss and gene transfer from the genome of the symbiont to the nuclear genome of the host (Endosymbiotic Gene Transfer or EGT) [22]. Moreover, there can be gene transfers from sources other than the symbiont giving rise to the observed plastid [Horizontal (or Lateral) Gene Transfer or HGT/LGT], for example, over (more or less cryptic) transient endosymbioses (e.g., "shopping bag" [23–25] and "red carpet" [26] hypotheses). Alternatively, HGT can occur in a, possibly ulterior, "non-endosymbiotic context" [27,28] (e.g., "limited transfer window" hypothesis" [29]), because it may be easier to duplicate or recruit a foreign gene for servicing the nascent plastid than to get it from the symbiont itself [30]. In any case, both EGT and HGT have shaped the nuclear genome of photosynthetic euglenids, leading to heavy genetic mosaicism (e.g., [7,31,32]).

Due to its great metabolic flexibility, a large number of culture media and growing conditions have been used to study *E. gracilis* over the past 60 years [33–37]. Commonly, the mineral composition remains similar from one medium to another, but three parameters vary greatly: the pH (which can be acidic or neutral), the source of organic carbon (e.g., acetate, ethanol, and succinate) and the concentration of the carbon source (from 10 mM to more than 150 mM). *E. gracilis* can therefore exploit a variety of organic carbon sources, as well in the dark (heterotrophic conditions) as in the light (mixotrophic conditions), where a high concentration of organic carbon leads to a decrease in photosynthesis by repressing chlorophyll biosynthesis, reflecting the fact that this organism switches between nutritional modes and combines them readily [38–40]. *E. gracilis* is also known for its atypical metabolic pathways, some of them producing compounds of commercial interest. In photosynthetic euglenoids, carbon reserves are stored in the cytoplasm in the form of paramylon (β-1,3-glucan), in place of the starch (α-1,4 and α-1,6-glucan) typical of the green line [41,42]. Paramylon can be used to produce bioplastics [43] and, similarly to other β-glucans, has been reported to display some anti-tumoural activity [44]. In anoxic (fermentative) conditions, *E. gracilis* has the unique ability among microalgae to convert paramylon into wax ester compounds suitable for drop-in jet biofuels conversion because of their low freezing point [45–47]. *E. gracilis* is also used as a source of dietary supplements (e.g., the most bioactive form of vitamin E, α-tocopherol, is present in *E. gracilis* biomass in a relatively high amount) [48].

Due to its evolutionary and biotechnological interests, *E. gracilis* is the best studied member of the euglenids. Its chloroplast genome (143 kb) was among the first plastid genomes ever sequenced [49], while its tiny mitochondrial genome has been recently resolved [50,51]. To date, few studies have used high throughput sequencing technologies to publish Omics information on *E. gracilis* [7,52,53]. In this respect, attempts to sequence its nuclear genome are also very recent (initially estimated between 1 Gb to 9 Gb; see [54] for a review). These efforts have culminated with the release of a very large (500 Mb) and highly fragmented draft genome, as authors recalled, due to gapped contigs or unknown base representation in half of the genome [7].

In this work, we have assembled a consensus transcriptome taking advantage of the raw read data publicly available, including newly generated transcriptomic libraries, for a total of five different data sources. Our assembly protocol was very thorough, with a special emphasis on potential contaminant sequences, resulting in the most complete transcriptome released to date for *E. gracilis*, according to a systematic comparison with the two other public transcriptomes [7,53]. After functional and taxonomic annotation of the predicted coding sequences, we performed a comparative study of their expression level across a range of culture conditions and studies, which allowed us to build an information-rich network of co-expressed genes. However, these results confirm that transcriptional control is not the primary level of genetic regulation in euglenozoans, while our taxonomic analyses point to highly mixed gene ancestry, compatible with a kleptoplastidic phase of plastid acquisition.

2. Materials and Methods

2.1. Data Collection

2.1.1. Public Repositories

Searching for public RNA-Seq data for *E. gracilis* in the International Nucleotide Sequence Database Collaboration (INSDC) returned eight studies. We further recovered an additional dataset, produced and submitted to the European Nucleotide Archive (ENA) repositories by ourselves (see Section 2.1.2 for details). Of these nine studies, only five short read datasets (5 experiments/23 samples) that used Illumina technology to analyse whole transcriptomes were exploitable. Among the discarded experiments, PRJEB4713 contained 454 GS FLX Titanium long reads, a size that is difficult to handle by the chosen assembler, while PRJEB21674 only included a single euglenid sample (among 1179), yet labelled as "Euglena sp.", PRJNA294935 primarily contained mitochondrial sequences, and PRJNA12797 (built out of ESTs) was not accessible from public repositories. At last, PRJDB4781 was not included because our meta-assemblies had been completed by the date of its release (October 2019). The data files from the five retained experiments were downloaded using fastq-dump utility from the SRA Toolkit with -I and –split-file arguments to divide files into forward and reverse paired reads. We also collected the two transcriptome assemblies hitherto available, GEFR01 and GDJR01. The former was encoded under study accession PRJNA298469, which corresponds to experiments B and C, and the latter, which corresponds to experiment D, was encoded as study PRJNA289402. For further details on experimental design or/and samples, see Table 1.

2.1.2. In-House Experiments, Cell Culture and Sequencing

The strain of *E. gracilis* (1224-5/25) was obtained from SAG (Sammlung von Algenkulturen Göttingen, Germany). Cells were cultured in liquid mineral medium tris-minimum-phosphate (TMP) at pH 7.0 and 25 °C, supplemented with a mixture of vitamins (vitamin B1 2·10-2 mM, vitamin B8 10-4 mM and vitamin B12 10-4 mM). In three samples, acetate (60 mM) was added as a carbon source, under different photosynthetic photon flux densities (PPFD, T8 fluorescent neon tubes) (in the dark, at low PPFD (50 µE m^{-2} s^{-1}) or at medium PPFD (200 µE m^{-2} s^{-1}), while in a fourth sample, acetate was not supplied and light was set to low PPFD (50 µE m^{-2} s^{-1}). For each sample, the cells in the exponential phase (1–2 × 10^{-6} cells/mL) were recovered by centrifugation, 10 min at 500 g. Total RNA was extracted with the protocol outlined in [55], then fragmented and retro-transcribed before standardization using the Duplex-Specific Nuclease kit (Evrogen, Russia). Each library was prepared using the Illumina total mRNA kit (Illumina, San Diego, CA, USA) and quantified by qPCR using the KAPA Library Quantification Kit (Roche, Switzerland). Subsequently, samples were sequenced in both reading directions (paired-end 2 × 100 nt) on four separate tracks of a high-speed sequencer Illumina HiSeq 2000, yielding on average ca. 235 million reads per sample. Library preparation, DSN normalization and high-throughput sequencing by Illumina technology were carried out by the GIGA genomics platform (https://www.gigagenomics.uliege.be (accessed on 23 July 2014)). Raw reads have been deposited at the ENA database under the study accession number PRJEB38787 (Table 1).

2.2. Data Assembly

A schematic representation of the de novo transcriptome reconstruction and analysis pipeline is given in Figure 1. All computations were performed on a grid computer.

Table 1. Representation of the collected data and overview of the experimental design. Exp. Code: letter assigned to each experiment (one letter per study). Study Acc.: public accession number of the BioProject. Sample Code: first letter corresponds to the experiment, first digit to experimental conditions of the samples, and second digit (if any) to the replicates. Run Acc.: public accession number of read FASTQ files. Temp.: estimated Celsius degrees of cell culture temperature. Medium: type of cell culture medium, rich (R) or mineral (M) plus carbon source (+C). Light: estimated light experimental conditions, darkness (D), low-light (LL) and high-light (HL). Shaking: rpm of shaker incubator. Cult. Cond.: trophic regime, fermentative (F), heterotrophic (H), phototrophic (P) or mixotrophic (M). Harvest Phase: development stage of the culture when collected, exponential phase (Exp) or stationary phase (Stat).

Exp. Code	Study Acc.	Sample Code	Run Acc.	Temp.	Medium	Light	Shaking	Cult Cond.	Harvest Phase	Reference
A	PRJNA310762	A.1.1	SRR3159774	25	R + C	D	0	H	Exp	[7]
		A.1.2	SRR3159775	25	R + C	D	0	H	Exp	
		A.1.3	SRR3159776	25	R + C	D	0	H	Exp	
		A.2.1	SRR3159777	25	R + C	LL	0	M	Exp	
		A.2.2	SRR3159778	25	R + C	LL	0	M	Exp	
		A.2.3	SRR3159779	25	R + C	LL	0	M	Exp	
B	PRJEB10085	B.1	ERR974915	21	M + C	LL	0	M	Stat	[52]
		B.2	ERR974916	30	R+C	D	200	H	Stat	
C	PRJNA298469	C.0	SRR2628535	25	M	LL	0	M	Stat	[7]
D	PRJNA289402	D.0	SRR3195326	26	R+C	HL	120	M	Stat	[53]
		D.1.1	SRR3195327	26	R+C	HL	120	M	Stat	
		D.1.2	SRR3195329	26	R+C	HL	120	M	Stat	
		D.1.3	SRR3195331	26	R+C	HL	120	M	Stat	
		D.2.1	SRR3195332	26	R+C	HL	120	F	Stat	
		D.2.2	SRR3195334	26	R+C	HL	120	F	Stat	
		D.2.3	SRR3195335	26	R+C	HL	120	F	Stat	
		D.3.1	SRR3195338	26	R+C	HL	120	F	Stat	
		D.3.2	SRR3195339	26	R+C	HL	120	F	Stat	
		D.3.3	SRR3195340	26	R+C	HL	120	F	Stat	
E	PRJEB38787	E.1	ERR4227585	25	M	LL	100	P	Exp	This study
		E.2	ERR4227586	25	M+C	D	100	H	Exp	
		E.3	ERR4227587	25	M+C	LL	100	M	Exp	
		E.4	ERR4227588	25	M+C	HL	100	M	Exp	

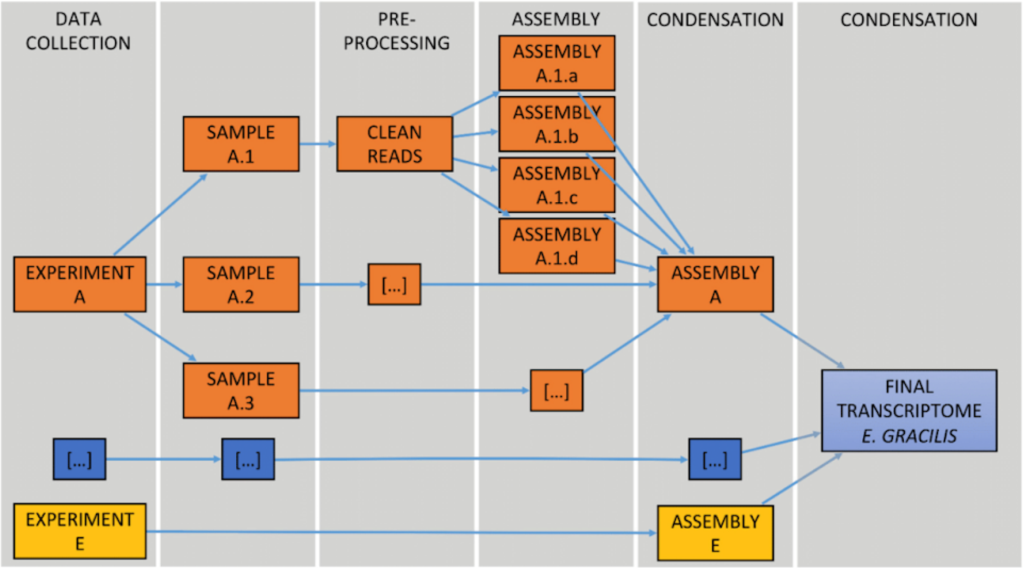


Figure 1. Schematic representation of our de novo transcriptome meta-assembly pipeline.

2.2.1. Data Pre-Processing

Every raw read file (run accessions SRR/ERR) was treated as one sample, even if two or more files were replicates of the same experimental condition. Once collected and transformed into fastq files, all samples were treated separately. Raw reads were analysed with FastQC v0.11.6 to assess the quality of the data [56]. PRINSEQ-lite.pl v0.20.4 was used to remove reads that contained more than one ambiguous nucleotide [57]. Then, Trimmomatic v0.32 was used with the following parameters (ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 SLIDINGWINDOW: 4:25 LEADING: 3 TRAILING: 3 MINLEN: 25) to truncate the low quality regions of certain sequences and cut adapters and other Illumina-specific sequences from the reads [58]. Output data was sorted into three different batches as paired, unpaired and singleton reads. Finally, read quality was re-assessed using FastQC, and the resulting plots visually compared to those obtained in the beginning to check the effect of the filtering procedure.

2.2.2. Transcriptome Assembly

Pre-processed reads (paired, unpaired and singleton reads) were assembled per experiment in two steps to yield five transcriptomes, one per experiment. We used Trinity v2.4.0 software [59] for de novo transcriptome assembly. During the first step, samples of each experiment were assembled four times, combining values (one/two) of minimum count for k-mers to be assembled (–min_kmer_cov) with normalization turned off (–no_normalize_reads) or on (default) to provide maximal sensitivity for reconstructing lowly expressed transcripts. In all cases, we used the default parameters with a minimum contig length (–min_contig_length) of 100 nt. Second, to reconstruct one single transcriptome per experiment, the four assembled transcriptome replicates were pooled together with the tr2aacds.pl script (using default parameters) from the EvidentialGene v2016.07.11 software package [60,61].

2.2.3. Transcriptome Decontamination

To ensure the purity of the five transcriptomes, we determined the guanine-cytosine (GC) content distribution across reconstructed transcripts. Furthermore, we explored the potential contamination of the five transcriptomes individually by comparing their transcripts against the NCBI nucleotide database (nt) using BLASTN v2.2.28 [62,63]. We used a conservative approach with an E-value threshold of 1×10^{-50} and an identity threshold of 90% to maximize the identification of true matches. The best hit for each query was selected, and the organism name (sscinames) of these top matches were collected, tabulated and quantified. Abundant organisms other than *Euglena* were flagged as putative contaminants. To obtain uncontaminated transcriptomes, the original reads were first aligned to the corresponding genomes (downloaded from Ensembl [64] using Bowtie 2 v2.2.6 in local mode (–local –no-unal)) [65,66]. Reads for which the alignment score exceeded the default minimal value of 20 + 8.0 * ln(L), where L is the read length, were removed. Then, the remaining (i.e., unaligned) reads were assembled again following the procedure described in Section 2.2.2.

2.2.4. Generation of a Consensus Transcriptome

The five resulting transcriptomes (one per experiment) were further combined and analysed with the tr2aacds.pl and evgmrna2tsa2.pl (-onlypubset) scripts from EvidentialGene to select the overall best candidate transcripts. The remaining reconstructed transcripts were discarded because they were classified either as redundant, fragmented or uninformative coding sequences, based on untranslated region (UTR) length, gaps, amino acid quality, and stop and start codon presence. After reducing redundancy, EvidentialGene clustered the best transcripts by groups of likely isoforms using CD-HIT v4.6.8 [67,68] and a similarity threshold of 90% on the amino-acid sequences. Sequences were considered as true isoforms (i.e., representing the same gene) when sharing high-identity (≥98%) exon-sized fragments, as determined with BLASTN v2.2.28 (E-value cut off of 1×10^{-19}).

Transcripts proposed by EvidentialGene as the most representative isoform for each gene were selected for annotation (see Sections 2.4 and 2.5) and for studying gene expression (Sections 2.6–2.8).

2.3. Assessment of Transcriptome Quality

Additional analyses were performed to determine the quality of the assembled transcripts. The same set of analyses was also performed on the two other transcriptomes publicly available (GEFR01 [7] and GDJR01 [53]) for comparison with the present study. First, basic statistics based on the length of transcripts and the number of ORFs were computed. Read representation was determined by mapping back the cleaned reads (see Section 2.2.1) to each of the three transcriptomes with the aligner Bowtie 2 v2.2.6 (–local, –no-unal) as described in [65]. Note that unpaired and singleton reads were excluded from all quality statistics. In parallel, we used two evaluation tools, Detonate v1.11 [69] and TransRate v1.0.3 [70], to get reference-free quality scores for the three transcriptomes.

To check the presence of the spliced leader (SL) sequence [14] in the three public transcriptomes, we used wordmatch from the EMBOSS software package [71] and three length thresholds (12, 14 and 24 nt) found in the literature [52,53]. Matches were only considered when falling at the 5′-end of a transcript, whether in forward or reverse orientation, as transcripts are not oriented in the transcriptomes. More precisely, each transcript was first reverse-complemented, and both versions (forward and reverse) were truncated at 40 nt before running wordmatch. Besides, transcripts actually corresponding to rRNA sequences were identified by combining RNAmmer v1.2 [72] and MegaBLAST v2.2.28 [62] searches (E-value cut-off of 1×10^{-50}, the latter using accessions X12890.1 (*E. gracilis* rrnC operon), M12677.1 (SSU rRNA 18S) and X53361.2 (LSU rRNA 28S) as queries. Regarding coding sequences, we estimated the numbers of putative genes with GeneMarkS-T (beta version) [73] and measured transcriptome completeness with BUSCO v.3.0.1 [74,75] using both "Eukaryota" and "Protists *ensembl*" datasets.

Lastly, we used CD-HIT-2D v4.6.8 [67,68] to identify similar predicted protein sequences between transcriptomes with our transcriptome as a reference. We explored different word sizes (2 to 5) at several thresholds of sequence identity (ranging from 0.5 to 0.9). Sequences from the other two public transcriptomes that could not be clustered with sequences of our consensus transcriptome were tentatively aligned using BLASTP v2.2.28 instead [62]. We further calculated the expression of presumably "missing" sequences in GDJR01 (D) and GEFR01 (B-C), respectively, following the procedure described in Section 2.5. The sequence was deemed invalid and not considered missing if its expression was below one transcript per kilobase million (TPM) in the transcriptome from which it had been identified. In a complementary analysis, highly similar nucleotide sequences from the three transcriptomes were clustered all together at once using CD-HIT-EST (identity threshold of 0.9, word size of 8, coverage of the shorter sequence of 0.9). Within each cluster, transcripts were pooled per transcriptome and their properties used to compare the three transcriptomes over all clusters, in terms of redundancy, length and identity. Analyses were performed either on all clusters or only on clusters shared across the three transcriptomes.

2.4. Transcript Annotation

The annotation procedure was carried out in three steps. First, assembled transcripts (i.e., the EvidentialGene representative isoforms) were annotated with EggNOG-mapper v1 [76,77]. We used HMMER to compare our data with the eukaryotic database of EggNOG, prioritizing coverage. Second, we annotated our transcripts by similarity using PSI-BLAST v2.2.28 searches [62] (E-value cut-off of 0.001) against Swiss-Prot [78]. Third, we aligned the assembled transcripts to the NCBI protein (*nr*) database [63] using TBLASTN v2.2.28 [62] (same E-value cut-off). We recovered Gene Ontology terms (GO) [79] and Kyoto Encyclopedia of Genes and Genomes Orthologs terms (KO) [80] of each transcript for further term enrichment analysis and network representation (see Section 2.7 for details). For that purpose, EggNOG features were assigned when possible to a transcript; if annotation was missing, PSI-BLAST v2.2.28 annotation was provided instead, or even TBLASTN v2.2.28 features whenever the two first previous methods failed. For mitochondrion and plastid-specific analyses, the components of the photosynthetic and respiratory electron transport chains were identified by BLASTP v2.2.28 searches [62] (E-value cut-off of 0.001) against reference proteins described in the literature. Hence, respiratory subunits were taken from [9,81,82], whereas subunits of photosystem I, photosystem II, cytochrome b6f complex, cF1Fo ATP-synthase were sourced from [83], and LHC polyproteins from [84].

2.5. Taxonomic Analyses

Taxonomic affinities were determined based on BLASTX v2.2.28 [62] searches against a broadly sampled proteome database, composed of 73 manually selected eukaryotes [85] and 19,802 representative prokaryotes subsampled from a curated database of 27,762 genomes [86]. For each assembled transcript, a last common ancestor (LCA) was computed based on their closest relatives (best hits, if any) in the database, provided they had a bit-score $\geq$80 and were within 95% of the bit-score of the first hit (MEGAN-like algorithm [86,87]). Organellar (plastid and mitochondrion) encoded proteins were distinguished from nuclear-encoded proteins by querying (BLASTP) two *E. gracilis* organelle databases assembled from the NCBI RefSeq “Proteins” portal [63]. To identify with certainty an organelle-encoded protein, only hits with a minimum percentage identity of 99% and a strictly identical length were considered. Such organelle-encoded sequences were expected at least from our own reads, which were generated in the absence of poly-A selection.

In parallel, tetranucleotide frequencies (TNFs) were computed for individual transcripts using the default settings of compseq from the EMBOSS software package [71]. Then, assembled transcripts for which a taxonomic affiliation had been obtained were ranked following their GC content and split into four partitions of equal size in terms of number of transcripts. Finally, ten principal component analyses (PCAs) were computed on TNFs, each one based on 1000 randomly chosen transcripts, using the prcomp function of the STATS v3.4.3 R base package [88]. For each PCA, two different colour schemes were applied on data points: the broad taxonomic affiliation of the transcript LCA (divided into four groups: Viridiplantae, Kinetoplastida, other Eukaryota and Bacteria), and the GC-content partition of the transcript.

2.6. Expression Quantification

The abundance of assembled transcripts was estimated by using RSEM v1.2.31 [89] and Bowtie2 v2.2.6 aligner [65,66]. Specifically, we used the align_and_estimate_abundance.pl Perl script wrapped in the Trinity v2.4.0 software package [59]. Data was then processed with abundance_estimates_to_matrix.pl Perl script without normalization parameters to generate the final expression matrix. Expression values are provided in transcripts per kilobase million (TPM) and pooled per gene (i.e., gene-level counts) [90].

Each count value was log2-transformed and converted to a Z-score to make samples comparable (sample mean was subtracted from each sample observation and divided by sample standard deviation). Batch effects were tentatively removed with the help of the SVA v.3.26.0 R package [91], so as to adjust data for unwanted sources of variation. However, such correction proved to be ineffective and thus abandoned (see Results and Discussion). For downstream analyses, only the 2500 most variable genes were retained (based on their expression variance across the 23 samples).

2.7. Gene Clustering Based on Expression Profiles

The 2500 most variable genes were clustered using the Partitioning around medoids (PAM) algorithm (from the CLUSTER v.2.0.7 R package) [92], which creates a fixed number of clusters (k) by minimizing the sum of the dissimilarities of the observations to their closest representative object (medoid). To capture both positive and negative relationships between gene pairs, we used a dissimilarity matrix of expression based on the squared Pearson correlation ($d = 1 - r^2$). The optimal cluster segregation was selected by cycling through the number of potential solutions, ranging from k = 5 to 75. In each solution, an average of maximal absolute correlations within-cluster (w-k cor_{max}) and an average of minimum absolute correlations between-cluster medoids (b-k cor_{min}) were computed. To intercept the point where optimal cluster segregation occurred, a reinterpretation of the Dunn index was used, and we computed the b-k cor_{min} and w-k cor_{max} ratio, choosing the solution with the minimal ratio value. At this optimal point, decreasing or increasing the number of cluster solutions would not better explain the data [93]. Heat map and hierarchical clustering analyses (correlation was used as the distance and centroid linkage clustering as the method) of expression data were carried out using the pheatmap function from the pheatmap v1.0.12 R package [94] and, when necessary, row-wise data (gene expression of the transcripts) was aggregated using k-means clustering to facilitate visual inspection of expression across conditions.

2.8. Gene Ontology (Enrichment) Analyses

The clusters based on the 2500 most variable genes were further analysed to visualize overrepresented biological terms using the whole GO and KEGG term space from Section 2.4 as a background. We explored enriched pathways within the expression clusters using ClueGo v2.5.0 tool [95], a visualization plug-in implemented in the Cytoscape v3.6.0 environment [96]. Term overrepresentation was estimated by an enrichment test based on the hypergeometric distribution followed by Benjamini–Hochberg adjustment for multiple testing. An annotation network was built with the ClueGo plug-in from kappa scores, which reflect the associations between genes and GO and KEGG terms. Network specificity was set between 3 and 12 GO hierarchy levels, and term selection was set to a minimum of 3% genes per cluster. Kappa score threshold was set to 0.3, and we allowed GO parent-child term fusion. Moreover, we explored the network with the MCODE algorithm [97], implemented as a Cytoscape plug-in, to detect densely connected regions or hubs in the network. Those hubs were found in the network establishing a degree cut-off of 2 for network scoring criteria, without including loops. Option Fluff was selected and parameters for Cluster Finding panel were set at 0.1 and 0.2 for node density and node score cut-off, respectively, a minimum of 2 edges per node of cluster cores (K-Core) and a maximum depth of 100.

3. Results and Discussion

3.1. Data Collection/Datasets

Out of the eight datasets publicly available for *E. gracilis*, only four [PRJNA310762 (A), PRJEB10085 (B), PRJNA298469 (C), PRJNA289402 (D)], were retained to assemble our consensus transcriptome, along with our own experiment PRJEB38787 (E; Table 1), which used Duplex-Specific thermostable nuclease (DSN) normalization to avoid poly-A selection. These five datasets totalled circa 2.6 billion raw Illumina reads (100-nt long), of which 70% belong to our experiment. After quality treatment, between 5 and 7% of reads were lost in experiments PRJNA310762 (A), PRJNA298469 (C) and PRJNA289402 (D), whereas the rejection of reads was more important in experiments PRJEB10085 (B) and PRJEB38787 (E). In PRJEB10085 (B), 19% of reads were truncated as a consequence of low-quality regions, whereas in PRJEB38787 (E), 50% of reads were discarded because of the high number of ambiguous nucleotides, especially in reverse reads. Hence, we got 57.8 million of good quality reads out of 62 after pre-processing of experiment PRJNA310762 (A) [7], 310 million reads out of 383 for experiment PRJEB10085 (B) [52], and 267.7 million from experiment PRJNA289402 (D). In the latter case, we used all samples as input, whereas Yoshida et al. (2016) only used the reads from cells grown in mixotrophic conditions to build their assembly [53]. Finally, Ebenezer et al. (2019) used 410 million reads as input for their transcriptome assembly, probably as the result of combining reads from PRJEB10085 (B) and PRJNA298469 [7].

After quality filtering, ca. 1.5 billion reads were retained, pre-processed read files of each individual experiment were assembled in four replicates using Trinity and then condensed into one individual transcriptome per experiment using EvidentialGene, which served as the basis for creating the consensus transcriptome (see Materials and Methods for details). Overall, PRJEB38787 (E), PRJEB10085 (B), PRJNA289402 (D), PRJNA310762 (A) and PRJNA298469 (C) experiments accounted for 55, 20, 17, 4, and 2% of the pre-processed reads used for the individual assemblies, respectively.

3.2. De Novo Assembly Evaluation

3.2.1. Individual Assemblies

The presence of sequences within a data set that originate from sources other than the sequenced sample is a known limitation of RNA-Seq experiments (e.g., [98,99] in human datasets). For some studies, such as large-scale phylogenomics, contaminants can be very problematic and must be dealt with using an array of different approaches [100]. Thus, before combining the individual five transcriptomes into a final consensus transcriptome, all assembled sequences were BLASTed against the NCBI nucleotide (*nt*) database [63] to identify possible contaminants. Using stringent thresholds, we found in the five transcriptomes only 948 unique hits of reconstructed transcripts that matched organisms other than *E. gracilis*. These organisms were considered as possible contaminants. Among them, we selected the five organisms whose abundance was the greatest (*Homo sapiens*, *Saccharomyces cerevisiae*, *Escherichia coli*, *Ovis aries* and *Caenorhabditis elegans*). It is noteworthy that sheep (and cow) DNA is commonly sequenced on our genomic platform. By mapping all pre-processed reads to the nuclear genome of these five species, we found that contaminants were less than 0.01% of the reads matching one of the contaminant genomes. In comparison, it has been shown that 0.13% of contaminant reads were present on average in a subset of 150 sequencing data files from the 1000 Genomes Project [101]. In the case of PRJNA298469 (C), we flagged as contaminants 68 reads per million reads (RPM), a larger proportion compared to the other experiments, which varied between 2 and 29 RPM (Table 2). Contaminant reads were removed and new assemblies of each experiment were generated anew from decontaminated reads, following the same procedure as above (see Section 2.2.2 for details). Afterwards, a new BLAST analysis was performed to quantify whether the contamination level was reduced. As expected, hits matching to *C. elegans*, *Escherichia coli*, *H. sapiens*, *O. aries* and *Saccharomyces cerevisiae* decreased, while hits matching to *Euglena* remained similar (Supplementary Figure S1). Besides, we traced the non-*Euglena* sequences that persisted in the final consensus transcriptome presented just below (see Section 3.2.2). Overall, from 716 unique hits of non-*Euglena* sequences identified with the latter BLAST analysis, only 64 were still present in the final consensus transcriptome (see Section 3.3.2 for details on the contamination sources). As a case in point, the complex genetic makeup of *E. gracilis* (e.g., [52]) makes it difficult to determine when a sequence, even if very peculiar, has been acquired from a very distantly related species or whether it can be a contaminant (see also Section 3.3.2 for an attempt to differentiate the two cases). For example, the glyoxylate cycle is localized within the mitochondria in *E. gracilis* and isocitrate lyase and malate synthase form only one bifunctional enzyme, called EgGCE [102,103]. A bifunctional enzyme for the glyoxylate cycle is also found in the worm *C. elegans* (opisthokonts), revealing an independent acquisition of the bifunctional enzyme by convergent evolution in these two organisms [104].

The five decontaminated individual transcriptomes were then evaluated with TransRate to check their uniformity. Four transcriptomes yielded ca. 42,342 (±6159) transcripts on average, whilst the number of reconstructed sequences in experiment PRJEB10085 (B) was more than twice the average, 95,490 sequences (Table 2). In addition, the computed GC content was 58% for experiment PRJEB10085 (B), a lower percentage compared to the other assembled transcriptomes, which was around 64%. Finally, we discovered a high frequency of sequences under 500 nt and characterized by a lower GC content (Supplementary Figure S2). After those small sequences were removed (representing 62% of the transcripts), TransRate statistics were recomputed and yielded values more in line with other experiments, both in terms of number of sequences (36,287) and GC content (62%).

We could not determine what the removed sequences were by similarity searches. They might represent some sort of artefact, contamination, or even be the result of a specific feature of experiment PRJEB10085 (B), for example the sequencing of a different strain, i.e., *E. gracilis* var. saccharophila Klebs (SAG 1224/7a) [52], whereas the other four experiments all used the Z strain (SAG 1224-5/25).

Table 2. Basic statistics based on transcript properties of reconstructed transcriptomes from collected data. ACC: study accession, REF: bibliographic reference, RAW: number of downloaded reads, PRE: number of good reads after pre-processing, CNT: number of reads removed after pre-processing considered as contamination (reads per million; rpm), SEQ: number of transcripts, MIN: minimal sequence length, MAX: maximal sequence length, MEAN: mean sequence length, TOTAL: combined sequence length, SEQ < 200: number of transcripts under 200 n, SEQ > 1 k: number of transcripts over 1000 nt, SEQ > 10 k: number of transcripts over 10,000 nt, ORF: number of sequences with a predicted open reading frame, ORF (%): for contigs with an ORF, the mean % of the contig covered by the ORF, N[z]: minimum contig length needed to cover [z]% of the transcriptome. GC (%): percentage of guanine-cytosine content, PART and PART (%): number and percentage of sequences contributed to the final consensus transcriptome (see below). In PRJEB10085 (B) (filtered), sequences <500 nt were further discarded (see text).

Statistic	A	B	B (Filtered)	C	D	E
ACC	PRJNA310762	PRJEB10085	PRJEB10085	PRJNA298469	PRJNA289402	PRJEB38787
REF	[7,52,53]					This study
RAW	61,531,862	383,416,636	383,416,636	27,096,926	285,148,782	1,902,226,200
PRE	57,862,467	310,302,570	310,302,570	25,244,887	267,779,751	875,299,135
CNT	740 (12 rpm)	9080 (29 rpm)	9080 (29 rpm)	1750 (68 rpm)	1191 (4 rpm)	2403 (2 rpm)
SEQ	38,559	95,490	36,287	42,363	37,425	51,021
MIN	101	101	500	101	101	101
MAX	13,929	21,744	21,744	11,354	26,839	10,795
MEAN	1043	647	1312	810	1120	610
TOTAL	40,861,413	64,426,688	47,615,807	34,438,742	42,382,170	31,671,589
SEQ < 200	4330	17,074	0	782	3051	3989
SEQ > 1 k	16,289	18,638	18,638	10,932	17,048	7104
SEQ > 10 k	4	15	15	1	13	1
ORF	24,757	29,060	27,842	27,063	24,817	26,882
ORF (%)	88%	82%	83%	89%	87%	93%
N90	576	347	654	419	606	367
N70	1140	667	1101	686	1187	528
N50	1607	1282	1574	1014	1658	753
N30	2257	2033	2243	1452	2318	1090
N10	3600	4026	3707	2358	3812	1850
GC (%)	64%	58%	62%	64%	64%	64%
PART	22,234	-	27,730	10,129	19,663	11,602
PART (%)	24.3%	-	30.3%	11.1%	21.5%	12.7%

3.2.2. Final Consensus Transcriptome

To obtain our final transcriptome, we combined the individual five decontaminated transcriptomes into a consensus transcriptome. Regardless of the aforementioned differences in the amount of pre-processed reads per dataset, the contribution of transcripts from each study in the final consensus transcriptome was rather balanced, where PRJEB10085 (B), PRJNA310762 (A), PRJNA289402 (D), PRJEB38787 (E), and PRJNA298469 (C) accounted for 30.3%, 24.3%, 21.5%, 12.7%, and 11.1%, respectively (Table 2). The resulting transcripts were classified into non-redundant protein-encoding genes, and one representative isoform was selected for each gene. Our new transcriptome was then compared with the other two publicly available transcriptomes, GDJR01 (D) [53] and GEFR01 (B-C) [7] (Table 3). Ebenezer et al. (2019) [7] used a combination of in-house generated sequences (PRJNA298469 (C)) and publicly available data from O'Neill et al. (2015) [52] (PRJEB10085 (B)) to assemble a transcriptome. Assembly transcriptome statistics were computed with TransRate. The overall number of sequences reported in the present work is 91,040, with N50 of 1432 nt, whereas in GDJR01 (D), it was 113,152 (N50 1604), and 72,506 (N50 1242) in GEFR01 (B-C).

The mean length of our transcripts was 1096 nt, a value closer to GDJR01 (D) than GEFR01 (B-C), which was ca. 200 nt smaller. The number of protein coding regions predicted by GeneMarkS-T (58,542) and the number of open reading frames (ORF) found with TransRate (62,287) are slightly smaller than in GDJR01 (D), but about twice greater than in GEFR01 (B-C). Our own sequences were classified into 49,922 predicted non-redundant protein-encoding genes, which is comparable to GDJR01 (D), but almost eighteen thousand genes more than in GEFR01 (B-C). As expected, these recomputed numbers are similar to those reported in the original publications of Yoshida et al. (2016) [53] and Ebenezer et al. (2019) [7]. Additionally, O'Neill et al. (2015) [52] found over 32,000 unique components for their *E. gracilis* transcriptome. The total size of our consensus transcriptome is 100 Mb, whilst the size of GDJR01 (D) is 122 Mb, 63 Mb for GEFR01 (B-C) and 38.4 Mb for O'Neill et al. (2015) [52] transcriptome. Overall, the genome size of *E. gracilis* has been estimated from total DNA content to range between 1 Gbp to 9 Gbp [54]. In contrast, the most recent estimation based on high throughput sequencing data was 332–500 Mb in size for the whole haploid genome [7] but, because half of the genome is gapped or has unknown base representation, the authors pointed out that this latter estimation was likely to be approximate.

Table 3. Basic statistics of transcript properties computed for the three public transcriptome assemblies, including the consensus transcriptome generated in the present work, and completed with data retrieved from the publications of Ebenezer et al. (2019) [7] and Yoshida et al. (2016) [53]. Row titles are as in Table 2, except for CDS: number of unique coding sequences (i.e., ORFs or UNIGENEs), GMS-T and GMS-T (%): number and percentage of predicted protein coding regions calculated by GeneMarkS-T.

Statistic	GEFR01	GDJR01	HBDM01
REF	[7,53]		This study
SEQ	72,506	113,152	91,040 [1]
MIN	202	201	201
MAX	25,763	21,553	26,839
MEAN	869	1087	1096
TOTAL	63,049,595	122,976,775	100,187,451
SEQ < 200	0 [1]	0 [1]	0 [1]
SEQ > 1 k	19,740	49,277	37,294
SEQ > 10 k	25	27	24
ORF [2]	30,467	65,943	62,287
ORF (%)	79%	73%	85%
N90	374	523	545
N70	704	1130	965
N50	1242	1604	1432
N30	1916	2181	2049
N10	3344	3347	3410
GC (%)	61%	63%	63%
CDS	32,128	49,826	49,922
GMS-T	35,929	63,432	58,542
GMS-T (%)	49%	56%	64%

[1] Submission tools for sequence repositories do not accept transcripts $\leq$ 200 nt. Hence, the number of sequences in the public version of HBDM01 is lower than reported elsewhere in this work. [2] ORFs were determined with TransDecoder, whereas CDS were determined with EvidentialGene (or a similar tool, depending on the study).

The pre-processed reads from the five experiments were aligned back to the three public transcriptomes as a metric of completeness. In most cases, the percentage of mapping was over 80%, reaching even more than 90%, with the exception of reads produced by ourselves PRJEB38787 (E), which had a representation of ~75% and ~50% in GEFR01 (B-C) and GDJR01 (D), respectively (Table 4). It is probable that our reads have a lower mapping percentage because they were generated from DSN-normalized total RNA samples, for which analyses of a preliminary sequencing lane revealed many reads corresponding to non-mRNA sequences (e.g., rRNA). However, the specifically low mapping to GDJR01

(D) cannot be explained easily because "transcripts" matching to rRNA sequences were identified in all three public transcriptomes (Supplementary Archive File S1).

Table 4. Mapping fraction of pre-processed reads from each collected dataset (rows) to the three public transcriptome assemblies (columns), GEFR01 [7], GDJR01 [53] and HBDM01 (this study).

Code	Accession	Reference	GEFR01	GDJR01	HBDM01
A	PRJNA310762	[7]	87.40%	92.51%	93.38%
B	PRJEB10085	[52]	84.68%	90.13%	91.49%
C	PRJNA298469	[7]	80.26%	91.66%	90.39%
D	PRJNA289402	[53]	85.25%	95.04%	94.28%
E	PRJEB38787	This study	75.28%	51.39%	80.76%

Using BUSCO on our predicted proteins, we found that the consensus transcriptome contained 84.8% of complete eukaryotic orthologs and half of them were duplicated, while 10.6% were missing (Figure 2). In comparison, we estimated the completeness of GDJR01 (D) at 80.8% of complete orthologs, of which a fifth were duplicated, and completeness of GEFR01 (B-C) at 76.9%, with only 4% of them duplicated. Moreover, we observed that lower percentages of complete orthologs were accompanied by higher numbers of fragmented and missed sequences. Overall, our consensus transcriptome appears to be the most complete, GEFR01 (B-C) being the least. Ebenezer et al. (2019) [7] also determined BUSCO completeness in GDJR01 (D) and GEFR01 (B-C) transcriptomes in addition to the original transcriptome presented by O'Neill et al. (2015) [52] and similarly concluded that GEFR01 (B-C) was the least complete transcriptome. Beyond transcripts missing due to low expression, discrepancies in the number of complete orthologs predicted by the different studies may also be due to the use of different tools for protein prediction. Whereas we used cdna_bestorf.pl script from EvidentialGene, the other studies used TransDecoder [59], which, reportedly, tends to predict larger amounts of proteins, but performs worse for true transcripts [105]. Despite these differences, the general representation scores of the reads in the assembled transcripts were similar across the three public transcriptomes, even if depending on the exact evaluation software used (Table 5).

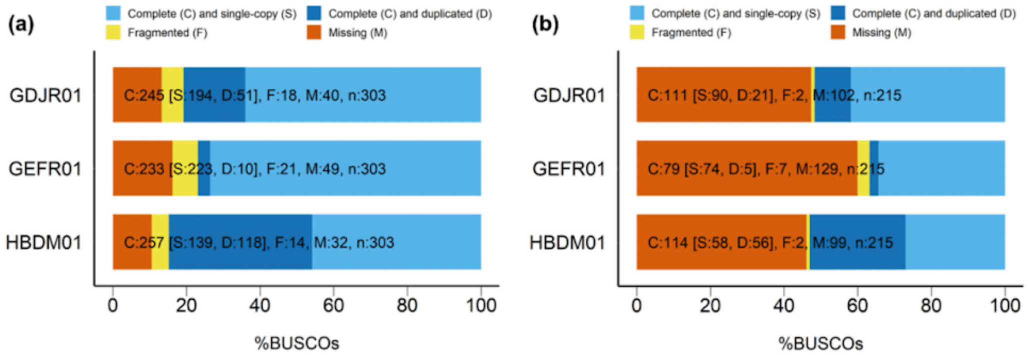


Figure 2. BUSCO-generated charts showing the relative completeness of the three public transcriptome assemblies, GEFR01 [7], GDJR01 [53] and HBDM01 (this study). BUSCO datasets were based on odb9. (**a**) "Eukaryota" (303 BUSCOs); (**b**) "Protists *ensembl*" (215 BUSCOs).

Table 5. TransRate and Detonate assembly scores for the three public transcriptome assemblies, GEFR01 [7], GDJR01 [53] and HBDM01 (this study). Scores indicate how well transcripts are supported by the RNA-Seq data.

Assembly Score	GEFR01	GDJR01	HBDM01
TransRate Score	0.1789	0.0304	0.0430
TransRate Optimal Score	0.2051	0.1729	0.0764
Detonate Score	$-97{,}461 \times 10^6$	$-97{,}561 \times 10^6$	$-97{,}459 \times 10^6$

As already mentioned, one evidence supporting the evolutionary relationship between trypanosomatids and euglenids are trans-splicing mechanisms [14]. We found that the SL-sequence was present in no more than 10.8% of transcripts in our transcriptome, far from the approximately 53–60% prevalence reported before [14,53], and closer to the 16% found by [52]. However, when performing the exact same analysis on the other two public transcriptomes, we find contrasting results, with SL-sequence matches recovered in at most of 2% and 30.3% of GEFR01 and GDJR01, respectively (Table 6). This indicates that the transcriptome of Yoshida et al. (2016) [53] has the most complete transcripts in 5-end, even though our own assembly includes 200 transcripts with a full-length perfect match to the 24-nt SL-sequence (vs. 45 and 5 for GEFR01 and GDJR01, respectively). Comparison of the mapping coverage for the three public transcriptomes shows that partial matches (12–14 nt) are much more numerous than full-length matches, as expected, but that the former are concentrated at the very beginning of the transcripts, which suggests that they are genuine SL-sequences (Supplementary Figure S3).

Table 6. SL-sequence related statistics for the three public transcriptome assemblies, GEFR01 [7], GDJR01 [53] and HBDM01 (this study). These correspond to exact matches limited to the first 40 nucleotides of each transcript.

Threshold (nt)	Statistic	GEFR01	GDJR01	HBDM01
24	Forward matches	24	5	86
	Reverse matches	21	0	114
	Total matches	45	5	200
	Average length (nt)	24.00	24.00	24.00
14	Forward matches	176	16,580	3370
	Reverse matches	200	12,999	3265
	Total matches	376	29,579	6635
	Average length (nt)	16.28	15.57	15.59
12	Forward matches	749	18,322	4403
	Reverse matches	766	16,016	5397
	Total matches	1515	34,338	9800
	Average length (nt)	13.37	15.19	14.68

Finally, we determined whether sequences of the other two available transcriptomes were present in our consensus transcriptome through two complementary approaches: one pairwise, sensitive and based on protein sequences, and one global, conservative and based on nucleotide sequences (Supplementary Table S1b). First, when using CD-HIT-2D with our transcriptome as a reference, a word size of 2 and an identity threshold of 0.4, 26.1% (34,490) of total sequences from GDJR01 (D) were missing and 37.6% (28,552) of total sequences from GEFR01 (B-C). Missing sequences were BLASTed (TBLASTN E-value cut-off of 0.001) against our transcriptome, and 20.5% (27,152) of total sequences of GDJR01 (D) were recaptured and 24.8% (18,870) of GEFR01 (B-C) (Supplementary Table S1a). After computing TPM values using the pre-processed reads generated in this study, we found that only 518 missing sequences of GDJR01 (D) were expressed above 1 TPM and 1595 in GEFR01 (B-C), which means that potentially 0.5% and 2% of the truly expressed sequences from GDJR01 (D) and GEFR01 (B-C), respectively, are missing from our consensus transcriptome. Hence, these sensitive analyses suggest that we captured more than 98% of the sequences produced in the other transcriptomes hitherto published. Second, CD-HIT-EST was used to compute clusters of related transcripts at an identity threshold of 90%. We recovered 121,851 clusters, in which the three transcriptomes had very similar patterns of presence and representation (Supplementary Table S1b). Hence, each transcriptome had at least one transcript in 60,220 to 66,041 clusters, whereas they each provided the representative (longest) sequence in 39,434 to 41,610 clusters. Singleton cluster statistics were slightly different, with GEFR01 having 29,997 specific clusters, followed by GDJR01 (27,058) and then our own transcriptome (19,028). When focusing on the 24,164 clusters shared between the three transcriptomes, we see that our transcriptome contributes the highest number of representative sequences, which confirms that they are generally longer than their homologues in the other two transcriptomes. This is also visible in a direct comparison of the mean an maximum transcript length across the three transcriptomes, whether on the 121,851 or the 24,164 clusters (Supplementary Figure S4). In contrast, comparison of the median and max identity between transcripts of the three datasets reveals that GEFR01 sequences are the most similar on average to the sequences from the two other transcriptomes. They are also the less redundant, with the lowest number of transcripts per cluster.

Altogether, these comparative analyses indicate that the three publicly available transcriptomes each have a distinct edge on the other two: Ebenezer et al. (2019) [7] assembled a compact set of sequences nonetheless providing a large fraction of unique transcripts, whereas Yoshida et al. (2016) [53] obtained a more redundant transcriptome, but with many transcripts complete at their 5-end, as evidenced by the detection of SL-sequences, and for our part, we generated the longest transcripts on average, including a few hundred featuring a full-length SL-sequence, with moderate redundancy.

3.3. Global (Transcriptome) Annotation

3.3.1. Functional Annotation of Transcripts

The combination of annotation strategies in our 49,922 predicted non-redundant protein-encoding genes yielded 9916 sequences with GO terms, 7775 KEGG orthologs, 13,298 sequences with a functional annotation and 13,850 with a taxonomic affiliation (Supplementary Table S2; see also Section 3.3.2). In the same way, O'Neill et al. (2015) [52] found 14,389 proteins with annotated functions out of the 32,128 predicted proteins of their transcriptome, whereas out of the 49,826 unique components reported by Yoshida et al. (2016) [53], approximately 11,314 were functionally annotated. Ebenezer et al. (2019) [7] annotated over 19,000 sequences, but without discerning what kind of attributes were associated in each case.

In comparison to the annotation performed in the other transcriptomes, we were able to find all the enzymes of the mevalonate pathway, including the diphosphomevalonate decarboxylase (EC 4.1.1.33), which was missing in the work of O'Neill et al. (2015) [52], thereby revealing that the last reaction is catalysed by a canonical enzyme. Regarding the carbohydrate-active enzymes, we found results similar to those outlined by O'Neill et al. (2015) [52]. Hence, we identified a great number of glycosyltransferases (311) and glycoside hydrolases (80), of which a quarter (19) were different types of glucanases (Supplementary Table S3). Corroborating the results of Yoshida et al. (2016) [53], we found two transcripts encoding glucan synthases, but could not identify transcripts encoding a 1,3-β-D-glucan phosphorylase, despite that such an enzyme has been previously characterised biochemically [106,107].

In *E. gracilis*, the photoreceptor is considered by some authors to be a rhodopsin-like protein where the retinal chromophore is a carotenoid [108]. We found five enzymes involved in retinol metabolism (EC 2.3.1.76; EC 3.1.1.64, EC 2.3.1.135; EC 1.1.1.105, EC 1.3.99.23) but, in line with Ebenezer et al.'s (2019) [7] findings, we could not find any rhodopsin-like protein candidates. Instead, we found 47 genes involved in visual perception processes (GO:0007601) and, more broadly, 333 genes related to photoresponse (Supplementary Table S4), including 13 cAMP/cGMP phosphodiesterases involved in amplification of luminous signal, 15 GTPase regulators, nine arrestins, which are important for regulating signal transduction at G protein-coupled receptors, eight cryptochromes, and three cyclic nucleotide-gated channels of rod photoreceptors. In addition, we found 13 proteins of the paraflagellar rod, a structure observed in euglenids, kinetoplastids and dinoflagellates [109–111]. Such a structure is associated with the paraflagellar body (also called paraxonemal body, PAB) in *E. gracilis* [112]. We also found 49 transcripts coding for photoactivated adenylate cyclases (PAC), which are light-sensitive proteins of PAB [113]. Of these, 43 clearly show a bacterial affinity in our analyses, whereas two are highly similar two trypanosomatid sequences [114].

To better understand the general functionality of the consensus transcriptome, we reported the GO annotation results as high-level terms of the three ontologies without the detail of the specific fine-grained terms. For such a task, we used the generic GO Slim Mapper tool of The Saccharomyces Genome Database [115], and the list of summarized GO terms (GO slim) can be found in Supplementary Table S5. As we used a compendium of culture conditions, we expected to capture the sum of functionalities represented by the studies individually. We found a total number of 164 GO terms after GO slim analysis, represented by core metabolism (41), transport (13), cell organization (15) and maintenance (25), nucleotide metabolism (35) and protein synthesis (17), vesicle or cilium organization (15) among others. The annotation from O'Neill et al. (2015) [52] was classified into 157 GO categories while Yoshida et al. (2016) [53] determined, under mixotrophic conditions, that the main functional categories were genetic information processing (399 components), translation (291 components), and energy metabolism (239 components). Besides, genes belonging to the latter three categories were generally down-regulated during anaerobic treatment [53]. In the same way, Ebenezer et al. (2019) [7] indicated that major categories were dominated by core metabolic, structural and informational process supergroups, consistent with the current work and previous studies [52,53].

3.3.2. Taxonomic Annotation of Transcripts

As a complex alga resulting from a secondary endosymbiosis between a euglenozoan host and a chlorophyte alga, *E. gracilis* bears genes from multiple origins [16,25]. In terms of sequence similarity (and depending on the current sampling in reference organisms), its nuclear genome is expected to be composed of four main gene classes: (i) *Euglena*-specific genes, (ii) kinetoplastid-specific genes, (iii) eukaryotic genes (i.e., widespread in other eukaryotes), and (iv) (green) genes acquired during the secondary endosymbiosis [31]. Over the last fifteen years, this issue has been extensively studied, both using similarity [52,53] and phylogenetic [7,9,31,32,116–119] approaches, either at small (i.e., targeted subsets) [9,116–118] or larger (i.e., transcriptomic) scales and, when at larger scale, either by focusing on the chloroplast [119] or by surveying "unbiased" transcript collections [7,31,32,52,53]. All these studies have revealed that *E. gracilis* display sequence similarities to a panel of organisms that is larger than predicted by a simple theory of secondary symbiogenesis [120,121]. Unsurprisingly, our large-scale similarity analyses of the consensus transcriptome confirm the results of these previous works (Figure 3). A first observation is that only 28% of the predicted non-redundant protein-encoding genes (13,850 out of 49,922) bear any exploitable similarity with sequences in reference databases. Among those, 937 (7%) correspond to organisms to which we could not assign a specific taxon, whereas 4054 (29%) were only identified as "Eukaryota". The remaining gene similarities are distributed among kinetoplastids (1364, 10%), green plants (977, 7%) and other subgroups of eukaryotes, whether photosynthetic, such as cryptophytes (530, 4%) and haptophytes (468, 3%), or not, e.g., opisthokonts (947, 7%). Bacterial groups account for 1690 transcripts (12%), among which the most prominent are proteobacteria (34% of bacteria) and cyanobacteria (212, 13%). Only 40 (2%) and 15 (0.9%) transcripts are affiliated to the PVC group or Chlamydiae, respectively [122]. As expected [31], focusing on 119 nuclear-encoded genes involved in mitochondrial and photosynthetic electron transfer chains increases the similarity signal in favour of kinetoplastids (20 out of 86, 22%) and green plants (20 out of 33, 58%), respectively (Supplementary Figure S5; see also HTML Supplementary Files S2 and S3).

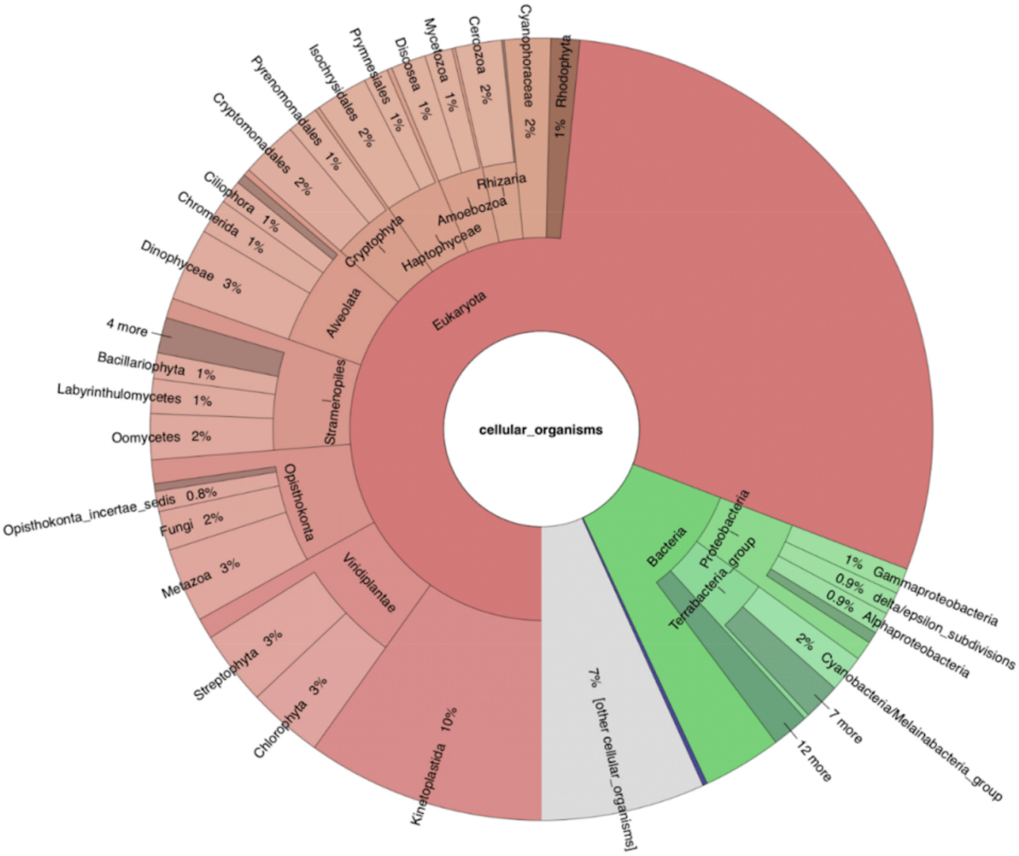


Figure 3. Taxonomic analysis of reconstructed transcripts (BLASTX MEGAN-like affiliations). The Krona chart is a zoom on the 13,850 transcripts to which a taxonomy could be associated, i.e., 28% of the 49,922 reconstructed transcripts. Among this classified fraction, 937 (7%) correspond to organisms to which we cannot assign a specific taxon ("other cellular organisms"). The thin blue slice is labelled "Archaea" (0.2%). The interactive chart is available as HTML Supplementary File S1.

Similarly to other complex algae (e.g., cryptophytes and chlorarachniophytes [123], ochrophytes and haptophytes [124,125]), *E. gracilis* transcriptomes show a heavily mixed ancestry in terms of gene donor lineages. However, it is a known (yet somewhat neglected) issue that publicly available transcriptomes can be contaminated by foreign sequences because of ecology (e.g., predator–prey, host–parasite or symbiotic relationships), or due to cross-contamination (either in the lab or on sequencing platforms) (see [126] and references therein). That is why we exerted special care to avoid including non-*Euglena* transcripts when assembling the five individual transcriptomes (see Section 3.2.1). In our final consensus transcriptome, we still identified 64 sequences as contaminants, of which 23 are false positives, owing to strong sequence similarity with different kinetoplastids (9 transcripts), green plants or algae (7), or non-green microalgae (7). Since the transcriptome had already been publicly released at the time, the other 41 remaining sequences were retained in subsequent analyses, but tagged as contaminants (Supplementary Table S6). Moreover, we used the taxonomic annotation of the 13,850 annotated transcripts to determine whether contaminants could be identified by their base composition pattern (see [127] and references therein). To this end, PCA plots were computed based on transcript tetranucleotide frequencies. Two types of colour annotation were then applied: one following a scale of GC-content and one following the taxonomy (Supplementary Figure S6). It appears that the taxonomic signal is mixed throughout these PCAs, whereas GC-content clearly corresponds to the PC1 axis. Thus, it was not possible in our case to identify and sort out contaminated transcripts (if any) from *Euglena* transcripts with this approach.

3.4. Systematic Functional Annotation of Top Differentially Expressed Genes

To better understand the functional organization of the most relevant *E. gracilis* genes under the assayed culture conditions, we computed a network of ontologies, based on transcript expression levels across all samples and studies (Supplementary Table S7). For this purpose, we only selected GO and KEEG terms that corresponded to the 2500 most variable genes (in terms of expression) to determine which biological functions were represented and how they were related to each other. The resulting organized network contained 119 nodes, with an average of nine neighbours per node, and 436 genes from the initial 2500 genes were retained (some genes being part of multiple hubs). We then used the MCODE algorithm to find evidence of higher order organization (Figure 4). The network was composed of nine modules (or hubs), each defined by one ontological category (Supplementary Table S8). Hub number 1 (72 transcripts) reflects “regulation of DNA damage checkpoint”, with transcripts involved in apoptosis, control of transcription and other developmental processes. Unlike hub number 7 (see below), hub 1 has a stress response component. Hub 2 (191 transcripts) is the largest hub, and comprises genes involved in translational initiation and termination, or protein targeting to a membrane, and is thus defined by “ribosome” terms. Hub 2 is connected to hubs 3, 5 and 6 in the network. Categorized as a “thylakoid” hub, hub 3 (133 transcripts) is the second largest hub. It mainly comprises photosynthetic electron transport chain transcripts and other components that respond to light stimuli. According to taxonomic annotation, the majority of the genes represented in this hub come from green organisms. Transcripts involved in protein kinase activity were found in Hub 4 (23 transcripts), defined as “cyclin-dependent protein serine/threonine kinase regulator activity”. Hub 5 (25 transcripts) corresponded mainly to processes involved in genetic information processing, such as spliceosome, exosome, chromosome-associated proteins, or chaperones. Hub 6 (79 transcripts) is defined by several categories related to mitochondrial protein complexes and mitochondria transport, and has a central position in the network (connections to hubs 1, 2, 3 and 8). Hub 7 (46 transcripts) was defined by “DNA integrity checkpoint” ontology terms and consisted of cell cycle processes, such as transition from G1 phase to S or the previously mentioned DNA integrity checkpoint. Hub 8 (53 transcripts) was categorized as “response to temperature stimulus” and was composed mainly of transcripts that encode heat shock proteins. Components of hub 9 (22 transcripts) were related to “negative regulation of translation”. Overall, our 2500 most relevant genes appear to be distributed around the central role of the mitochondrion, whose origin traces back to the euglenozoan host cell [31]. In this respect, our taxonomic analysis specifically revealed that more than 10% of genes are related to kinetoplastids (the closest available proxy for the host cell) in all hubs, except for hub 3, categorized as “thylakoid” (Supplementary Table S9).

3.5. Cluster Annotation Enrichment Analysis and Gene Co-Expression

From the same top 2500 variable genes, we identified positive and negative relationships between pairs of genes based on gene expression. We tried to capture genes that behave conjointly across the various experimental conditions and group them into clusters. According to our expectations where a gene would be binary regulated (up or down), the optimal k solution should range between 2^5 (32) and 2^{13} (8192) (accounting for 5 to 13 distinct experimental conditions with a total sample number of 23; see Table 1). We computed the optimal number of clusters and determined that 36 clusters was the most suitable solution for the selected genes (Supplementary Figure S7). To better understand the underlying biological processes inside the clusters, ontologies that were overrepresented were extracted and analysed. Only five out of the 36 clusters were characterized by significantly overrepresented ontological terms (Supplementary Table S10). In total, those five clusters were composed of 631 transcripts out of the 2500 initially used for clustering, and 52% of them had at least one annotation attribute. Their expression can be visualized in hierarchically clustered heat maps (Figure 5).

Results from the enrichment tests revealed that "nucleosome category" was overrepresented in cluster 1, which contains transcripts of the "DNA damage checkpoint" and "ribosome" hubs of the ontological network, hub 1 and 2, respectively (see above). These transcripts encode histones, and core components of "nucleosome", that participate in wrapping and compacting DNA into chromatin. The observation that DNA packaging, transcription and translation shared the same gene expression pattern may be relevant because in euglenids, as well as in dinoflagellates, chromosomes are permanently condensed [128]. Furthermore, transcripts encoding different components of the chloroplast reaction centres of hub 3 were also found in this cluster. This cluster was characterized by a larger down-regulated expression in PRJEB38787 (E), while other experiments were slightly over and under zero. Cluster 4 was enriched in "photosynthetic electron transport" and "DNA damage checkpoint" related terms mainly present in hub 3, with several transcripts encoding ATP synthase subunits in the former and cell cycle and apoptosis regulator proteins in the latter. Gene expression in cluster 4 was homogeneous with values ranging between one or minus one, except for a group of genes greatly down-regulated in studies PRJNA310762 (A), PRJEB10085 (B), PRJNA298469 (C), and likely to be not expressed in such experiments. About a third of the transcripts from cluster 19 encode different types of serine/threonine proteins and are ontologically typified by "cyclin-dependent protein serine/threonine kinase regulator activity", which are processes closely related to cell cycle regulation. Their expression was slightly negative in the experiment PRJEB38787 (E) and positive in PRJEB10085 (B) while it remained unaltered in the rest of the experiments. "Neuroblast proliferation" and "neuroblast division" categories illustrated cluster 24, which, considering the unicellular nature of *E. gracilis*, was more likely to be related to cytoskeletal structure of eukaryotic cells formed during cell division or cell polarity than regulation of neurogenesis. In study PRJNA289402 (D), ABC transporters, fatty acid and polyketide synthesis were more down-regulated than in the remaining studies. Lastly, cluster 25 was enriched in "positive regulation of mitochondria organization" due to the presence of putative mitochondrial heat shock proteins that were co-regulated across studies. Besides, expression of cluster 25 was disparate for PRJNA289402 (D), compared with the other studies. A main difference was a group of transcripts largely downregulated in the PRJNA289402 (D) experiment, while they were upregulated in the remaining studies. Those transcripts putatively encode different components of the nitrogen metabolism, some chloroplastic electron transport chain components and ATP-dependent RNA helicase. A few transcripts related to cell cycle and translation, present in the annotation network, were found in cluster 25.

The cluster patterns reported above show that expression is driven by study rather than experimental conditions of the studies. Even if disappointing, these findings were similar after the tentative SVA correction of the batch effect present in the studies (Supplementary Figure S8). Presumably, our approach was not able to properly capture the batch effect, maybe due to an unbalanced batch-group design of the studies [129]. Nonetheless, we observed that a selection of 133 genes, coding for the components of the photosynthetic and respiratory electron transport chains, were grouped together. This subset of genes, located in the chloroplast and in the mitochondrion, respectively, was selected because most of the experimental conditions (light/dark, presence or absence of acetate in the medium, oxic/anoxic environment) of the studies were expected to affect respiration and photosynthesis. As illustrated in Figure 6, the expression of these genes is also driven by the study rather than by the reported physico-chemical parameters of each experiment. Yet, most components of the mitochondrial electron transport chain among the 133 selected genes were grouped together after hierarchical clustering of their expression, while chloroplastic components exploded into different subgroups. Concretely, genes coding for light-harvesting complexes grouped together distantly from other chloroplastic components. These transcripts are nuclear-encoded and showed a taxonomic affinity to Streptophyta (Supplementary Table S11).

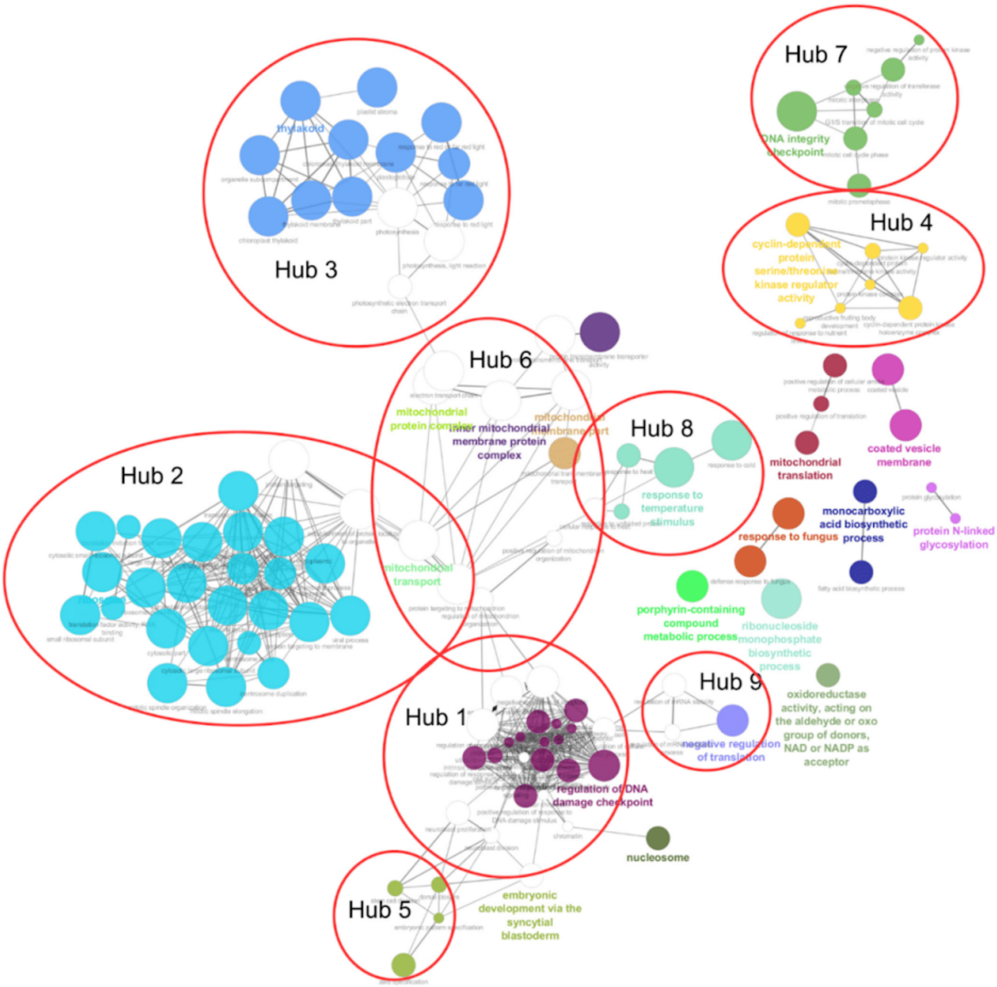

Hub 7
negative regulation of protein kinase activity
negative regulation of transferase activity
mitotic interphase
G1/S transition of mitotic cell cycle
DNA integrity checkpoint
mitotic cell cycle phase
mitotic prometaphase
Hub 4
cyclin-dependent protein serine/threonine kinase regulator activity
protein kinase regulator activity
cyclin-dependent protein serine/threonine kinase activity
protein kinase complex
reproductive fruiting body development
cyclin-dependent protein kinase holoenzyme complex
regulation of response to nutrient levels
Hub 3
thylakoid
plastid stroma
response to red or far red light
photosynthesis
response to red light
photosynthesis, light reaction
photosynthetic electron transport chain
chloroplast thylakoid
thylakoid part
Hub 6
protein transmembrane transporter activity
electron transport chain
mitochondrial protein complex
inner mitochondrial membrane protein complex
mitochondrial membrane part
mitochondrial transmembrane transport
mitochondrial transport
positive regulation of mitochondrion organization
protein targeting to mitochondrion
regulation of mitochondrion organization
Hub 8
response to heat
response to cold
response to temperature stimulus
cellular response to heat
positive regulation of cellular amide metabolic process
coated vesicle
positive regulation of translation
mitochondrial translation
coated vesicle membrane
protein glycosylation
monocarboxylic acid biosynthetic process
protein N-linked glycosylation
response to fungus
defense response to fungus
fatty acid biosynthetic process
porphyrin-containing compound metabolic process
ribonucleoside monophosphate biosynthetic process
Hub 2
cytosolic small ribosomal subunit
translation factor activity, RNA binding
small ribosomal subunit
cytosolic part
viral process
protein targeting to membrane
centrosome duplication
mitotic spindle organization
mitotic spindle elongation
Hub 1
regulation of DNA damage checkpoint
positive regulation of response to DNA damage stimulus
neuroblast proliferation
neuroblast division
chromatin
nucleosome
Hub 9
regulation of mRNA stability
negative regulation of translation
oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor
Hub 5
stem cell division
dorsal closure
embryonic pattern specification
axis specification
embryonic development via the syncytial blastoderm

Figure 4. Annotation network of ontological terms showing the functional organization and relationships between the 2500 most variable genes. GO and KEGG terms were considered as a large pool in which the genes could be associated with 0 to N terms. Such associations served as the basis to infer the network (see text). Colours correspond to ontological terms (or groups of related ontological terms).

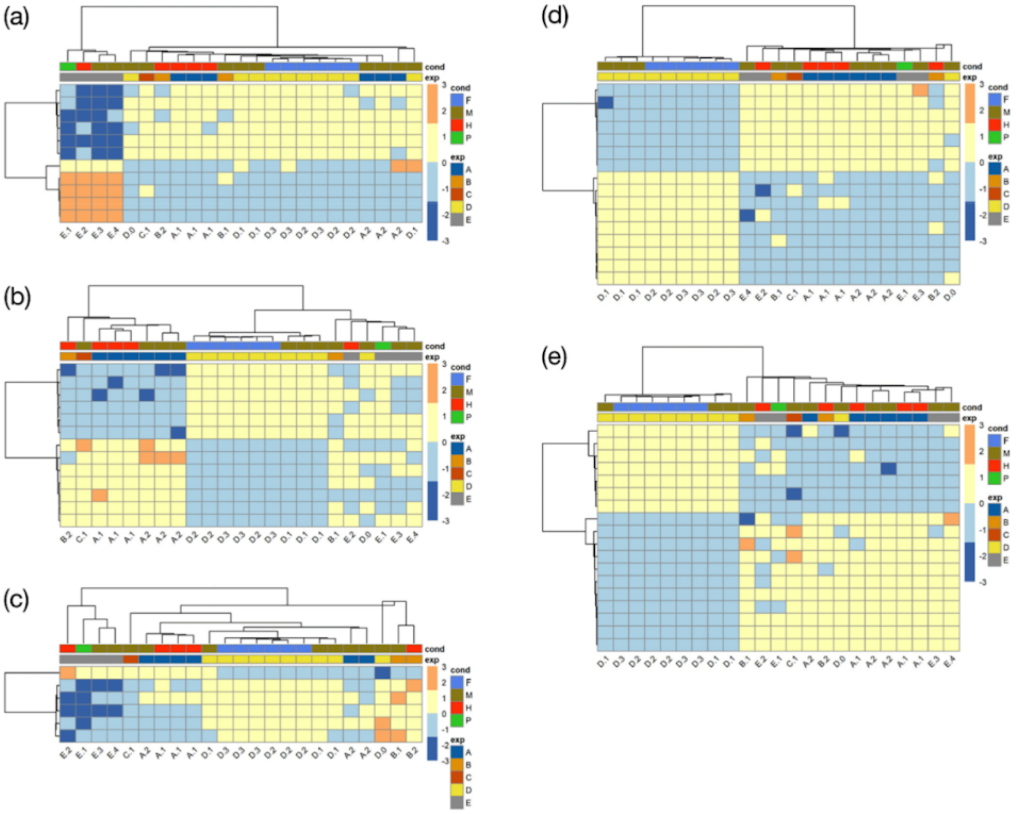


Figure 5. Selected co-expression clusters computed on the 2500 most variable genes. Only the five clusters characterized by significantly overrepresented ontological terms (featuring 631 transcripts) are shown. Heat maps and trees regroup samples behaving similarly across genes on the horizontal axis and genes behaving similarly across samples on the vertical axis; gene expression is vertically clustered to facilitate visualization (see text). Samples are colour-coded both by condition (F = fermentative, M = mixotrophic, H = heterotrophic, P = phototrophic) and by study (A = PRJNA310762, B = PRJEB10085, C = PRJNA298469, D = PRJNA289402, E = PRJEB38787). (**a**) Cluster 1; (**b**) cluster 4; (**c**) cluster 19; (**d**) cluster 24; (**e**) cluster 25.

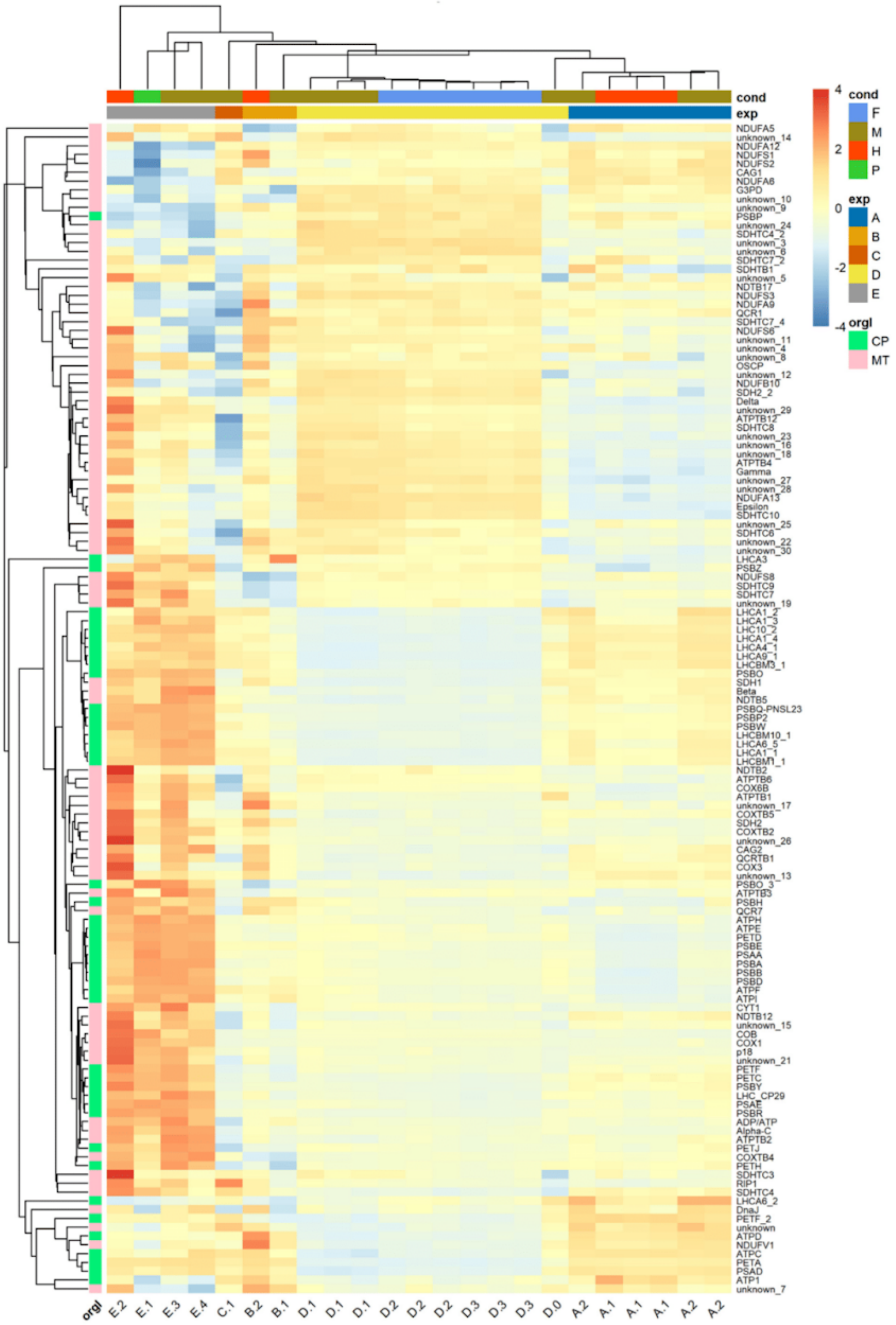


Figure 6. Expression heat map of 133 genes involved in electron transport chains. Heat maps and trees regroup samples behaving similarly across genes on the horizontal axis and genes behaving similarly across samples on the vertical axis (see

text). Samples are colour-coded both by condition (F = fermentative, M = mixotrophic, H = heterotrophic, P = phototrophic) and by study (A = PRJNA310762, B = PRJEB10085, C = PRJNA298469, D = PRJNA289402, E = PRJEB38787). Genes are colour-coded by organelle (CP = chloroplast; MT = mitochondrion).

Overall, our last analysis indicates that genes that share common metabolic functions are packed together, as would be expected, even though the expression is driven by study rather than culture condition. Beyond the technical issues that may have contributed to a loss of exploitable signal (e.g., heterogeneous experimental "design", see Table 1, uncorrected batch effects), these negative results can also be interpreted as additional evidence for the idea that, similar to what is known in trypanosomatids, nuclear gene expression in *E. gracilis* is not primarily regulated at the transcriptional level. In these parasites, gene regulation mostly occurs at the post-transcriptional level, through stabilization/degradation of mRNA molecules and control of mRNA translation (see [8] for a recent review of the issue). While the former mechanism should in principle change transcript abundance, the latter one might not be visible in comparative transcriptomics. For example, Yoshida et al. (2016) observed little change at the transcriptomic level following anaerobic treatment. Moreover, these changes in gene expression were inconsistent with respect to the activation of paramylon degradation and wax ester production [53]. In a more systematic investigation, Ebenezer et al. (2019) reported a striking lack of correlation between transcriptomic and proteomic data when comparing light and dark conditions [7]. As already mentioned, the raw transcriptomic data from these two studies were included in the present work (along with those of O'Neill et al. (2015) [52] and our own data), which allowed us to compare gene expression across a wider range of culture conditions at once. A few meaningful clusters of genes (i.e., following functional term enrichment) could be identified based on shared expression patterns across samples, which suggests that there is some biological signal in transcript abundance. However, the dominance of batch effects on these levels further questions the usefulness of transcriptomics for functional studies in *E. gracilis*.

4. Conclusions

Owing to its singular evolutionary origin, a merger between a chlorophyte alga and a phagotrophic unicellular belonging to a non-model eukaryotic group [20], *E. gracilis* is a fascinating, multifaceted chimeric organism, whose significance is constantly growing in domains as varied as the production of bio-based products [43], the treatment of wastewater ([130]), the provision of food supplements for space exploration [131], or the elucidation of mechanisms it shares with its parasitic trypanosome cousins [8,9,15] (see also the other articles of the present Special Issue).

By building a consolidated transcriptome of this photosynthetic eukaryote, we aimed at providing a solid resource to the community, taking into account previous work [7,52,53], yet enriched with unreleased data (obtained back in 2012–2014; Supplementary Figure S9) [132]. Our final consensus transcriptome comprises 91,040 unique transcripts and 49,922 predicted non-redundant protein-encoding genes. It appears to be the most complete up-to-date, at least according to sequence metrics, the number of universal orthologs found, read percentages supporting the assembly, and the fact that most of the *E. gracilis* sequences available to date have been included. Hence, we have been able to capture more than 98% of the sequences produced in the other transcriptomes hitherto published, while the number of predicted genes is in the same range [7,53]. This suggests that there was still some room for improvement, contrary to expectations for the opposite [7], and it might be related to the inclusion of reads obtained without poly-A selection, but following DSN normalization.

Annotating these transcripts, whether from a functional or taxonomic point of view, remains a challenge, notably because of the lack of well-characterized closely related organisms, the trypanosomes being relatively derived parasites [133]. This results in a mere 26–27% of our predicted genes annotated by sequence similarity, above the 23% of Yoshida et al. (2016) [53], but below the 45% of O'Neill et al. 2015 [52] and the 52–55% of

Ebenezer et al. (2019) [7], who further considered orthogroup sharing as annotation. In principle, this should encourage more large-scale studies, e.g., comparative transcriptomics performed in a wide range of culture conditions and stresses, in order to build a reliable gene expression network from co-expression data, and thereby provide alternative means for annotating genes of unknown function. Alas, as it now appears quite clearly, gene expression is mostly controlled at the post-transcriptional level in euglenozoans [7,8], including the regulation of chloroplast development in photosynthetic euglenids [134]. This implies that functional studies in *E. gracilis* have to be carried out through proteomics rather than transcriptomic approaches (e.g., [119,135]). This is fully possible considering the availability of several high-quality transcriptome assemblies to feed reference databases for proteomic fragment identification, including the one presented in this work. In this respect, the unfortunate lack of a complete genome beyond the draft level, even if frustrating, is not an insuperable issue [7].

Regarding the highly mixed taxonomic affinities of *Euglena* transcripts, our similarity searches yielded proportions in line with previous studies, even when those studies were based on more reliable phylogenetic approaches [136], such as the comprehensive work of Ebenezer et al. (2019) [7]. Altogether, the current knowledge points to the "shopping bag" [23–25] (or "red-carpet" [26]) model for the evolutionary origin of *Euglena*, i.e., transient endosymbioses during which multiple rounds of HGT/EGT have progressively shaped the plastid proteome. Yet, it is noteworthy that such a gene mixture would also be compatible with a kleptoplastidic origin for photosynthetic euglenids, in which the transient "endosymbioses" would actually imply stolen plastids and not intact symbionts. Moreover, some predatory euglenids, such as *Peranema trichophorum*, can feed either by phagocytosis of whole cells or by drilling a hole in their prey and then sucking up its cellular contents [137], a process known as myzocytosis [138]. Beyond providing a selective force for transferring genes to the host nucleus to service the ingested plastids, as in the recently characterized ARS (Antarctic Ross Sea) dinoflagellate bearing haptophyte-derived kleptoplastids [139], a kleptoplastidic model would also better fit the three membranes of the euglenid chloroplasts [20,140] and the presence of kleptoplastids acquired by myzocytosis in the early branching *Rapaza viridis* [141].

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes12060842/s1. Figure S1: Taxonomic distribution of best BLAST hits before and after decontamination. Figure S2: GC-content distribution across reconstructed transcripts and in function of transcript length. Figure S3: Mapping coverage analysis for the 24-nt SL-sequence on the 5-end of the transcripts. Figure S4: Comparison of transcript count, length and identity over clusters of highly similar transcripts. Figure S5: Taxonomic analysis of reconstructed transcripts corresponding to mitochondrial and photosynthetic electron transfer chains. Figure S6: PCA plots computed on the tetranucleotide frequencies of taxonomically annotated reconstructed transcripts. Figure S7: Correlation values for a range of cluster solutions. Figure S8: PCA plots computed on gene expression before and after SVA batch effect correction. Figure S9: Quality-control of the total RNA prepared in our lab. Table S1a: Pairwise overlap between the new consensus transcriptome and two publicly available transcriptomes. Table S1b: Global overlap between the three public transcriptomes. Table S2: Annotation of the 49,922 predicted non-redundant protein-encoding genes. Table S3: List of 392 genes corresponding to carbohydrate-active enzymes. Table S4: List of 380 genes involved in visual perception processes and photoresponse. Table S5: List of 164 GO slim terms generated by the Slim Mapper tool. Table S6: List of 64 possibly contaminant transcripts persisting in the final consensus transcriptome. Table S7: Expression values in transcripts per kilobase million (TMP) for the 49,922 genes. Table S8: Composition of the 9 hubs in the ontology network. Table S9: Taxonomic analysis of the 9 hubs in the ontology network. Table S10: Composition of the 5 clusters in the gene co-expression network. Table S11: Expression values (in TPM) of 133 genes involved in photosynthetic and respiratory electron transfer chains. Archive file S1: RNAmmer and MegeBLAST reports for rRNA sequences. HTML file S1: Interactive Krona chart for the taxonomic affiliations of the 49,922 genes. HTML file S1: Krona chart for the nuclear genes involved in the mitochondrial electron transfer chain. HTML file S3: Krona chart for the nuclear genes involved in the photosynthetic electron transfer chain.

Author Contributions: Conceptualization, P.C. and D.B.; methodology, J.C. and D.B.; software, J.C., M.V.V., A.R.B., V.L. and D.B.; formal analysis, J.C., M.V.V., A.R.B. and V.L.; investigation, J.C. and E.P.; resources, P.C. and D.B.; data curation, J.C., M.V.V., V.L. and D.B.; writing—original draft preparation, J.C. and D.B.; writing—review and editing, J.C., P.C. and D.B.; visualization, J.C., E.P., M.V.V. and A.R.B.; supervision, P.C. and D.B.; project administration, J.C., P.C. and D.B.; funding acquisition, P.C. and D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Belgian FRS-FNRS (Fonds de la Recherche Scientifique), grant number PDR T.0032, and the European Research Council (ERC), H2020-EU BEAL project 682580, both awarded to P.C. Computational resources were provided through two grants to D.B. (University of Liège "Crédit de démarrage 2012" SFRD-12/04 and FRS-FNRS "Crédit de recherche 2014" CDR J.0080.15). E.P., M.V.V., A.R.B. and V.L. were (or still are) FRIA fellows (Fonds pour Formation à la Recherche dans l'Industrie et dans l'Agriculture) of the FRS-FNRS. P.C. is a Senior Research Associate from the FRS-FNRS.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were generated in this study. These can be found at the European Nucleotide Archive (ENA), under study accession PRJEB38787 and TSA project accession HBDM01000000. Most custom analysis scripts have been deposited on GitHub (https://github.com/microalgues/clustering (accessed on 22 August 2019)).

Acknowledgments: We are grateful to Patrick E. Meyer for useful advice on network analyses.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Leander, B.S.; Farmer, M.A. Comparative morphology of the euglenid pellicle. II. Diversity of strip substructure. *J. Eukaryot. Microbiol.* **2001**, *48*, 202–217. [CrossRef]
2. Adl, S.M.; Simpson, A.G.; Lane, C.E.; Lukes, J.; Bass, D.; Bowser, S.S.; Brown, M.W.; Burki, F.; Dunthorn, M.; Hampl, V.; et al. The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* **2012**, *59*, 429–493. [CrossRef]
3. Breglia, S.A.; Yubuki, N.; Hoppenrath, M.; Leander, B.S. Ultrastructure and molecular phylogenetic position of a novel euglenozoan with extrusive episymbiotic bacteria: Bihospites bacati n. gen. et sp. (Symbiontida). *BMC Microbiol.* **2010**, *10*, 145. [CrossRef]
4. Burki, F. The Eukaryotic Tree of Life from a Global Phylogenomic Perspective. *Cold Spring Harb. Perspect. Biol.* **2014**, *6*, a016147. [CrossRef] [PubMed]
5. Zakrys, B.; Milanowski, R.; Karnkowska, A. Evolutionary Origin of Euglena. *Adv. Exp. Med. Biol.* **2017**, *979*, 3–17. [CrossRef]
6. Butenko, A.; Opperdoes, F.R.; Flegontova, O.; Horak, A.; Hampl, V.; Keeling, P.; Gawryluk, R.M.R.; Tikhonenkov, D.; Flegontov, P.; Lukes, J. Evolution of metabolic capabilities and molecular features of diplonemids, kinetoplastids, and euglenids. *BMC Biol.* **2020**, *18*, 23. [CrossRef]
7. Ebenezer, T.E.; Zoltner, M.; Burrell, A.; Nenarokova, A.; Novak Vanclova, A.M.G.; Prasad, B.; Soukal, P.; Santana-Molina, C.; O'Neill, E.; Nankissoor, N.N.; et al. Transcriptome, proteome and draft genome of Euglena gracilis. *BMC Biol.* **2019**, *17*, 11. [CrossRef]
8. Vesteg, M.; Hadariova, L.; Horvath, A.; Estrano, C.E.; Schwartzbach, S.D.; Krajcovic, J. Comparative molecular cell biology of phototrophic euglenids and parasitic trypanosomatids sheds light on the ancestor of Euglenozoa. *Biol. Rev. Camb. Philos. Soc.* **2019**, *94*, 1701–1721. [CrossRef] [PubMed]
9. Perez, E.; Lapaille, M.; Degand, H.; Cilibrasi, L.; Villavicencio-Queijeiro, A.; Morsomme, P.; Gonzalez-Halphen, D.; Field, M.C.; Remacle, C.; Baurain, D.; et al. The mitochondrial respiratory chain of the secondary green alga Euglena gracilis shares many additional subunits with parasitic Trypanosomatidae. *Mitochondrion* **2014**, *19 Pt B*, 338–349. [CrossRef] [PubMed]
10. Jackson, A.P.; Quail, M.A.; Berriman, M. Insights into the genome sequence of a free-living Kinetoplastid: Bodo saltans (Kinetoplastida: Euglenozoa). *BMC Genom.* **2008**, *9*, 594. [CrossRef] [PubMed]
11. Deschamps, P.; Lara, E.; Marande, W.; Lopez-Garcia, P.; Ekelund, F.; Moreira, D. Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids. *Mol. Biol. Evol.* **2011**, *28*, 53–58. [CrossRef]
12. Cavalier-Smith, T. Eukaryote kingdoms: Seven or nine? *Biosystems* **1981**, *14*, 461–481. [CrossRef]
13. Sogin, M.L.; Elwood, H.J.; Gunderson, J.H. Evolutionary diversity of eukaryotic small-subunit rRNA genes. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 1383–1387. [CrossRef] [PubMed]
14. Tessier, L.H.; Keller, M.; Chan, R.L.; Fournier, R.; Weil, J.H.; Imbault, P. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in Euglena. *EMBO J.* **1991**, *10*, 2621–2625. [CrossRef] [PubMed]
15. Montrichard, F.; Le Guen, F.; Laval-Martin, D.L.; Davioud-Charvet, E. Evidence for the co-existence of glutathione reductase and trypanothione reductase in the non-trypanosomatid Euglenozoa: Euglena gracilis Z. *FEBS Lett.* **1999**, *442*, 29–33. [CrossRef]

16. Sibbald, S.J.; Archibald, J.M. Genomic Insights into Plastid Evolution. *Genome Biol. Evol.* **2020**, *12*, 978–990. [CrossRef] [PubMed]
17. Rogers, M.B.; Gilson, P.R.; Su, V.; McFadden, G.I.; Keeling, P.J. The complete chloroplast genome of the chlorarachniophyte Bigelowiella natans: Evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Mol. Biol. Evol.* **2007**, *24*, 54–62. [CrossRef] [PubMed]
18. Turmel, M.; Gagnon, M.C.; O'Kelly, C.J.; Otis, C.; Lemieux, C. The chloroplast genomes of the green algae Pyramimonas, Monomastix, and Pycnococcus shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. *Mol. Biol. Evol.* **2009**, *26*, 631–648. [CrossRef]
19. Jackson, C.; Knoll, A.H.; Chan, C.X.; Verbruggen, H. Plastid phylogenomics with broad taxon sampling further elucidates the distinct evolutionary origins and timing of secondary green plastids. *Sci. Rep.* **2018**, *8*, 1523. [CrossRef]
20. Gibbs, S.P. The chloroplasts of Euglena may have evolved from symbiotic green algae. *Can. J. Bot.* **1978**, *56*, 2883–2889. [CrossRef]
21. Cavalier-Smith, T. Membrane heredity and early chloroplast evolution. *Trends Plant Sci.* **2000**, *5*, 174–182. [CrossRef]
22. Timmis, J.N.; Ayliffe, M.A.; Huang, C.Y.; Martin, W. Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **2004**, *5*, 123–135. [CrossRef]
23. Larkum, A.W.; Lockhart, P.J.; Howe, C.J. Shopping for plastids. *Trends Plant. Sci.* **2007**, *12*, 189–195. [CrossRef] [PubMed]
24. Howe, C.J.; Barbrook, A.C.; Nisbet, R.E.; Lockhart, P.J.; Larkum, A.W. The origin of plastids. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2008**, *363*, 2675–2685. [CrossRef] [PubMed]
25. Keeling, P.J. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annu. Rev. Plant Biol.* **2013**, *64*, 583–607. [CrossRef]
26. Ponce-Toledo, R.I.; Lopez-Garcia, P.; Moreira, D. Horizontal and endosymbiotic gene transfer in early plastid evolution. *New Phytol.* **2019**, *224*, 618–624. [CrossRef]
27. Teich, R.; Zauner, S.; Baurain, D.; Brinkmann, H.; Petersen, J. Origin and distribution of Calvin cycle fructose and sedoheptulose bisphosphatases in plantae and complex algae: A single secondary origin of complex red plastids and subsequent propagation via tertiary endosymbioses. *Protist* **2007**, *158*, 263–276. [CrossRef] [PubMed]
28. Petersen, J.; Ludewig, A.K.; Michael, V.; Bunk, B.; Jarek, M.; Baurain, D.; Brinkmann, H. Chromera velia, endosymbioses and the rhodoplex hypothesis–plastid evolution in cryptophytes, alveolates, stramenopiles, and haptophytes (CASH lineages). *Genome Biol. Evol.* **2014**, *6*, 666–684. [CrossRef]
29. Barbrook, A.C.; Howe, C.J.; Purton, S. Why are plastid genomes retained in non-photosynthetic organisms? *Trends Plant Sci.* **2006**, *11*, 101–108. [CrossRef]
30. Singer, A.; Poschmann, G.; Muhlich, C.; Valadez-Cano, C.; Hansch, S.; Huren, V.; Rensing, S.A.; Stuhler, K.; Nowack, E.C.M. Massive Protein Import into the Early-Evolutionary-Stage Photosynthetic Organelle of the Amoeba Paulinella chromatophora. *Curr. Biol.* **2017**, *27*, 2763–2773. [CrossRef] [PubMed]
31. Ahmadinejad, N.; Dagan, T.; Martin, W. Genome history in the symbiotic hybrid Euglena gracilis. *Gene* **2007**, *402*, 35–39. [CrossRef]
32. Maruyama, S.; Suzaki, T.; Weber, A.P.; Archibald, J.M.; Nozaki, H. Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in euglenids. *BMC Evol. Biol.* **2011**, *11*, 105. [CrossRef] [PubMed]
33. Cramer, M.; Myers, J. Growth and photosynthetic characteristics of Euglena gracilis. *Archiv. Für Mikrobiol.* **1952**, *17*, 384–402. [CrossRef]
34. Wilson, B.W.; Danforth, W.F. The extent of acetate and ethanol oxidation by Euglena gracilis. *J. Gen. Microbiol.* **1958**, *18*, 535–542. [CrossRef]
35. Buetow, D. Ethanol stimulation of oxidative metabolism in Euglena gracilis. *Nature* **1961**, *190*, 1196. [CrossRef]
36. Mego, J.L.; Farb, R.M. Alcohol dehydrogenases of Euglena gracilis, strain Z. *Biochim. Biophys. Acta* **1974**, *350*, 237–239. [CrossRef]
37. Sharpless, T.K.; Butow, R.A. An inducible alternate terminal oxidase in Euglena gracilis mitochondria. *J. Biol. Chem.* **1970**, *245*, 58–70. [CrossRef]
38. App, A.A.; Jagendorf, A.T. Repression of chloroplast development in Euglena gracilis by substrates. *J. Protozool.* **1963**, *10*, 340–343. [CrossRef]
39. Buetow, D.E. Acetate repression of chlorophyll synthesis in Euglena gracilis. *Nature* **1967**, *213*, 1127–1128. [CrossRef]
40. Vannini, G.L. Degeneration and regeneration of chloroplasts in Euglena gracilis grown in the presence of acetate: Ultrastructural evidence. *J. Cell Sci.* **1983**, *61*, 413–422. [CrossRef]
41. Calvayrac, R.; Laval-Martin, D.; Briand, J.; Farineau, J. Paramylon synthesis by Euglena gracilis photoheterotrophically grown under low O2 pressure: Description of a mitochloroplast complex. *Planta* **1981**, *153*, 6–13. [CrossRef]
42. Monfils, A.K.; Triemer, R.E.; Bellairs, E.F. Characterization of paramylon morphological diversity in photosynthetic euglenoids (Euglenales, Euglenophyta). *Phycologia* **2011**, *50*, 156–169. [CrossRef]
43. Shibakami, M.; Tsubouchi, G.; Hayashi, M. Thermoplasticization of euglenoid beta-1,3-glucans by mixed esterification. *Carbohydr. Polym.* **2014**, *105*, 90–96. [CrossRef]
44. Watanabe, T.; Shimada, R.; Matsuyama, A.; Yuasa, M.; Sawamura, H.; Yoshida, E.; Suzuki, K. Antitumor activity of the beta-glucan paramylon from Euglena against preneoplastic colonic aberrant crypt foci in mice. *Food Funct.* **2013**, *4*, 1685–1690. [CrossRef]
45. Matsuda, F.; Hayashi, M.; Kondo, A. Comparative profiling analysis of central metabolites in Euglena gracilis under various cultivation conditions. *Biosci. Biotechnol. Biochem.* **2011**, *75*, 2253–2256. [CrossRef]

46. Furuhashi, T.; Ogawa, T.; Nakai, R.; Nakazawa, M.; Okazawa, A.; Padermschoke, A.; Nishio, K.; Hirai, M.Y.; Arita, M.; Ohta, D. Wax ester and lipophilic compound profiling of Euglena gracilis by gas chromatography-mass spectrometry: Toward understanding of wax ester fermentation under hypoxia. *Metabolomics* **2015**, *11*, 175–183. [CrossRef]
47. Inui, H.; Ishikawa, T.; Tamoi, M. Wax Ester Fermentation and Its Application for Biofuel Production. *Adv. Exp. Med. Biol.* **2017**, *979*, 269–283. [CrossRef] [PubMed]
48. Ogbonna, J.C.; Tomiyamal, S.; Tanaka, H. Heterotrophic cultivation of Euglena gracilis Z for efficient production of α-tocopherol. *J. Appl. Phycol.* **1998**, *10*, 67–74. [CrossRef]
49. Hallick, R.B.; Hong, L.; Drager, R.G.; Favreau, M.R.; Monfort, A.; Orsat, B.; Spielmann, A.; Stutz, E. Complete sequence of Euglena gracilis chloroplast DNA. *Nucleic Acids Res.* **1993**, *21*, 3537–3544. [CrossRef] [PubMed]
50. Spencer, D.F.; Gray, M.W. Ribosomal RNA genes in Euglena gracilis mitochondrial DNA: Fragmented genes in a seemingly fragmented genome. *Mol. Genet. Genom.* **2011**, *285*, 19–31. [CrossRef] [PubMed]
51. Dobakova, E.; Flegontov, P.; Skalicky, T.; Lukes, J. Unexpectedly streamlined mitochondrial genome of the euglenozoan Euglena gracilis. *Genome Biol. Evol.* **2015**. [CrossRef]
52. O'Neill, E.C.; Trick, M.; Hill, L.; Rejzek, M.; Dusi, R.G.; Hamilton, C.J.; Zimba, P.V.; Henrissat, B.; Field, R.A. The transcriptome of Euglena gracilis reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. *Mol. Biosyst.* **2015**. [CrossRef]
53. Yoshida, Y.; Tomiyama, T.; Maruta, T.; Tomita, M.; Ishikawa, T.; Arakawa, K. De novo assembly and comparative transcriptome analysis of Euglena gracilis in response to anaerobic conditions. *BMC Genom.* **2016**, *17*. [CrossRef] [PubMed]
54. Ebenezer, T.E.; Carrington, M.; Lebert, M.; Kelly, S.; Field, M.C. Euglena gracilis Genome and Transcriptome: Organelles, Nuclear Genome Assembly Strategies and Initial Features. *Adv. Exp. Med. Biol.* **2017**, *979*, 125–140. [CrossRef] [PubMed]
55. Loppes, R.; Radoux, M. Identification of short promoter regions involved in the transcriptional expression of the nitrate reductase gene in Chlamydomonas reinhardtii. *Plant Mol. Biol.* **2001**, *45*, 215–227. [CrossRef] [PubMed]
56. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*; Babraham Bioinformatics, Babraham Institute: Cambridge, UK, 2010.
57. Schmieder, R.; Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **2011**, *27*, 863–864. [CrossRef]
58. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
59. Haas, B.J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P.D.; Bowden, J.; Couger, M.B.; Eccles, D.; Li, B.; Lieber, M.; et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **2013**, *8*, 1494–1512. [CrossRef]
60. Gilbert, D. Gene-omes built from mRNA-seq not genome DNA. *F1000Research* **2016**, *5*, 1695.
61. Gilbert, D.G. Genes of the pig, Sus scrofa, reconstructed with EvidentialGene. *PeerJ* **2019**, *7*, e6374. [CrossRef]
62. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef]
63. Sayers, E.W.; Beck, J.; Bolton, E.E.; Bourexis, D.; Brister, J.R.; Canese, K.; Comeau, D.C.; Funk, K.; Kim, S.; Klimke, W.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2021**, *49*, D10–D17. [CrossRef] [PubMed]
64. Howe, K.L.; Achuthan, P.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; Armean, I.M.; Azov, A.G.; Bennett, R.; Bhai, J.; et al. Ensembl 2021. *Nucleic Acids Res.* **2021**, *49*, D884–D891. [CrossRef] [PubMed]
65. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
66. Langmead, B.; Wilks, C.; Antonescu, V.; Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **2019**, *35*, 421–432. [CrossRef] [PubMed]
67. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef]
68. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef]
69. Li, B.; Fillmore, N.; Bai, Y.; Collins, M.; Thomson, J.A.; Stewart, R.; Dewey, C.N. Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biol.* **2014**, *15*, 553. [CrossRef] [PubMed]
70. Smith-Unna, R.; Boursnell, C.; Patro, R.; Hibberd, J.M.; Kelly, S. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* **2016**, *26*, 1134–1144. [CrossRef]
71. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]
72. Lagesen, K.; Hallin, P.; Rodland, E.A.; Staerfeldt, H.H.; Rognes, T.; Ussery, D.W. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **2007**, *35*, 3100–3108. [CrossRef] [PubMed]
73. Tang, S.; Lomsadze, A.; Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **2015**, *43*, e78. [CrossRef] [PubMed]
74. Simao, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef] [PubMed]

75. Waterhouse, R.M.; Seppey, M.; Simao, F.A.; Manni, M.; Ioannidis, P.; Klioutchnikov, G.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **2018**, *35*, 543–548. [CrossRef] [PubMed]
76. Huerta-Cepas, J.; Forslund, K.; Coelho, L.P.; Szklarczyk, D.; Jensen, L.J.; von Mering, C.; Bork, P. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **2017**, *34*, 2115–2122. [CrossRef]
77. Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernandez-Plaza, A.; Forslund, S.K.; Cook, H.; Mende, D.R.; Letunic, I.; Rattei, T.; Jensen, L.J.; et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **2019**, *47*, D309–D314. [CrossRef]
78. UniProt, C. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [CrossRef]
79. The Gene Ontology Consortium. The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Res.* **2021**, *49*, D325–D334. [CrossRef]
80. Kanehisa, M.; Furumichi, M.; Sato, Y.; Ishiguro-Watanabe, M.; Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res.* **2021**, *49*, D545–D551. [CrossRef] [PubMed]
81. Yadav, K.N.S.; Miranda-Astudillo, H.V.; Colina-Tenorio, L.; Bouillenne, F.; Degand, H.; Morsomme, P.; Gonzalez-Halphen, D.; Boekema, E.J.; Cardol, P. Atypical composition and structure of the mitochondrial dimeric ATP synthase from Euglena gracilis. *Biochim. Biophys. Acta Bioenerg.* **2017**, *1858*, 267–275. [CrossRef]
82. Miranda-Astudillo, H.V.; Yadav, K.N.S.; Colina-Tenorio, L.; Bouillenne, F.; Degand, H.; Morsomme, P.; Boekema, E.J.; Cardol, P. The atypical subunit composition of respiratory complexes I and IV is associated with original extra structural domains in Euglena gracilis. *Sci. Rep.* **2018**, *8*, 9698. [CrossRef] [PubMed]
83. Sobotka, R.; Esson, H.J.; Konik, P.; Trskova, E.; Moravcova, L.; Horak, A.; Dufkova, P.; Obornik, M. Extensive gain and loss of photosystem I subunits in chromerid algae, photosynthetic relatives of apicomplexans. *Sci. Rep.* **2017**, *7*, 13214. [CrossRef] [PubMed]
84. Koziol, A.G.; Durnford, D.G. Euglena Light-Harvesting Complexes Are Encoded by Multifarious Polyprotein mRNAs that Evolve in Concert. *Mol. Biol. Evol.* **2008**, *25*, 92–100. [CrossRef] [PubMed]
85. Van Vlierberghe, M.; Philippe, H.; Baurain, D. Broadly sampled orthologous groups of eukaryotic proteins for the phylogenetic study of plastid-bearing lineages. *BMC Res. Notes* **2021**, *14*. [CrossRef] [PubMed]
86. Cornet, L.; Meunier, L.; Van Vlierberghe, M.; Leonard, R.R.; Durieu, B.; Lara, Y.; Misztak, A.; Sirjacobs, D.; Javaux, E.J.; Philippe, H.; et al. Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLoS ONE* **2018**, *13*, e0200323. [CrossRef]
87. Huson, D.H.; Auch, A.F.; Qi, J.; Schuster, S.C. MEGAN analysis of metagenomic data. *Genome Res.* **2007**, *17*, 377–386. [CrossRef]
88. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
89. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef] [PubMed]
90. Soneson, C.; Love, M.I.; Robinson, M.D. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research* **2015**, *4*, 1521. [CrossRef]
91. Leek, J.T. Svaseq: Removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **2014**, *42*. [CrossRef]
92. Maechler, M.; Rousseeuw, P.; Struyf, A.; Hubert, M.; Hornik, K. Cluster: Cluster Analysis Basics and Extensions, R Package Version 2.0.7-1. 2018. Available online: https://cran.r-project.org/web/packages/cluster/index.html (accessed on 28 May 2021).
93. Dunn, J.C. Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **1974**, *4*, 95–104. [CrossRef]
94. Kolde, R.; Kolde, M.R. Pheatmap: Pretty Heatmaps, R Package Version 1.0.12. 2019. Available online: https://rdrr.io/cran/pheatmap/ (accessed on 28 May 2021).
95. Bindea, G.; Mlecnik, B.; Hackl, H.; Charoentong, P.; Tosolini, M.; Kirilovsky, A.; Fridman, W.H.; Pages, F.; Trajanoski, Z.; Galon, J. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **2009**, *25*, 1091–1093. [CrossRef]
96. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef]
97. Bader, G.D.; Hogue, C.W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **2003**, *4*, 2. [CrossRef]
98. Lusk, R.W. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS ONE* **2014**, *9*, e110808. [CrossRef] [PubMed]
99. Strong, M.J.; Xu, G.; Morici, L.; Splinter Bon-Durant, S.; Baddoo, M.; Lin, Z.; Fewell, C.; Taylor, C.M.; Flemington, E.K. Microbial contamination in next generation sequencing: Implications for sequence-based analysis of clinical samples. *PLoS Pathog.* **2014**, *10*, e1004437. [CrossRef] [PubMed]
100. Simion, P.; Philippe, H.; Baurain, D.; Jager, M.; Richter, D.J.; Di Franco, A.; Roure, B.; Satoh, N.; Queinnec, E.; Ereskovsky, A.; et al. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr. Biol.* **2017**, *27*, 958–967. [CrossRef] [PubMed]
101. Tae, H.; Karunasena, E.; Bavarva, J.H.; McIver, L.J.; Garner, H.R. Large scale comparison of non-human sequences in human sequencing data. *Genomics* **2014**, *104*, 453–458. [CrossRef]

102. Nakazawa, M.; Minami, T.; Teramura, K.; Kumamoto, S.; Hanato, S.; Takenaka, S.; Ueda, M.; Inui, H.; Nakano, Y.; Miyatake, K. Molecular characterization of a bifunctional glyoxylate cycle enzyme, malate synthase/isocitrate lyase, in Euglena gracilis. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **2005**, *141*, 445–452. [CrossRef] [PubMed]
103. Nakazawa, M.; Nishimura, M.; Inoue, K.; Ueda, M.; Inui, H.; Nakano, Y.; Miyatake, K. Characterization of a bifunctional glyoxylate cycle enzyme, malate synthase/isocitrate lyase, of Euglena gracilis. *J. Eukaryot. Microbiol.* **2011**, *58*, 128–133. [CrossRef]
104. Liu, F.; Thatcher, J.D.; Barral, J.M.; Epstein, H.F. Bifunctional glyoxylate cycle protein of Caenorhabditis elegans: A developmentally regulated protein of intestine and muscle. *Dev. Biol.* **1995**, *169*, 399–414. [CrossRef]
105. Gilbert, D. Evigene Versus Transdecoder for Proteins from Transcripts. Available online: https://sourceforge.net/p/evidentialgene/blog/2017/11/-evigene-versus-transdecoder-for-proteins-from-transcripts/ (accessed on 3 April 2021).
106. Kitaoka, M.; Sasaki, T.; Taniguchi, H. Purification and properties of laminaribiose phosphorylase (EC 2.4 1.31) from Euglena gracilis Z. *Arch. Biochem. Biophys.* **1993**, *304*, 508–514. [CrossRef]
107. Kuhaudomlarp, S.; Patron, N.J.; Henrissat, B.; Rejzek, M.; Saalbach, G.; Field, R.A. Identification of Euglena gracilis beta-1,3-glucan phosphorylase and establishment of a new glycoside hydrolase (GH) family GH149. *J. Biol. Chem.* **2018**, *293*, 2865–2876. [CrossRef] [PubMed]
108. Barsanti, L.; Evangelista, V.; Passarelli, V.; Frassanito, A.M.; Gualtieri, P. Fundamental questions and concepts about photoreception and the case of Euglena gracilis. *Integr. Biol.* **2012**, *4*, 22–36. [CrossRef] [PubMed]
109. Gallo, J.M.; Schrevel, J. Homologies between paraflagellar rod proteins from trypanosomes and euglenoids revealed by a monoclonal antibody. *Eur. J. Cell Biol.* **1985**, *36*, 163–168. [PubMed]
110. Cachon, J.; Cachon, M.; Cosson, M.-P.; Cosson, J. The paraflagellar rod: A structure in search of a function. *Biol. Cell* **1988**, *63*, 169–181. [CrossRef]
111. Maharana, B.R.; Tewari, A.K.; Singh, V. An overview on kinetoplastid paraflagellar rod. *J. Parasit. Dis.* **2015**, *39*, 589–595. [CrossRef] [PubMed]
112. Verni, F.; Rosati, G.; Lenzi, P.; Barsanti, L.; Passarelli, V.; Gualtieri, P. Morphological relationship between paraflagellar swelling and paraxial rod in Euglena gracilis. *Micron Microsc. Acta* **1992**, *23*, 37–44. [CrossRef]
113. Iseki, M.; Matsunaga, S.; Murakami, A.; Ohno, K.; Shiga, K.; Yoshida, K.; Sugai, M.; Takahashi, T.; Hori, T.; Watanabe, M. A blue-light-activated adenylyl cyclase mediates photoavoidance in Euglena gracilis. *Nature* **2002**, *415*, 1047–1051. [CrossRef] [PubMed]
114. Koumura, Y.; Suzuki, T.; Yoshikawa, S.; Watanabe, M.; Iseki, M. The origin of photoactivated adenylyl cyclase (PAC), the Euglena blue-light receptor: Phylogenetic analysis of orthologues of PAC subunits from several euglenoids and trypanosome-type adenylyl cyclases from Euglena gracilis. *Photochem. Photobiol. Sci.* **2004**, *3*, 580–586. [CrossRef]
115. Cherry, J.M.; Hong, E.L.; Amundsen, C.; Balakrishnan, R.; Binkley, G.; Chan, E.T.; Christie, K.R.; Costanzo, M.C.; Dwight, S.S.; Engel, S.R.; et al. Saccharomyces Genome Database: The genomics resource of budding yeast. *Nucleic Acids Res.* **2012**, *40*, D700–D705. [CrossRef]
116. Markunas, C.M.; Triemer, R.E. Evolutionary History of the Enzymes Involved in the Calvin-Benson Cycle in Euglenids. *J. Eukaryot. Microbiol.* **2016**, *63*, 326–339. [CrossRef]
117. Lakey, B.; Triemer, R.; Müller, K. The tetrapyrrole synthesis pathway as a model of horizontal gene transfer in euglenoids. *J. Phycol.* **2017**, *53*, 198–217. [CrossRef]
118. Ponce-Toledo, R.I.; Moreira, D.; Lopez-Garcia, P.; Deschamps, P. Secondary Plastids of Euglenids and Chlorarachniophytes Function with a Mix of Genes of Red and Green Algal Ancestry. *Mol. Biol. Evol.* **2018**. [CrossRef]
119. Novak Vanclova, A.M.G.; Zoltner, M.; Kelly, S.; Soukal, P.; Zahonova, K.; Fussy, Z.; Ebenezer, T.E.; Lacova Dobakova, E.; Elias, M.; Lukes, J.; et al. Metabolic quirks and the colourful history of the Euglena gracilis secondary plastid. *New Phytol.* **2020**, *225*, 1578–1592. [CrossRef]
120. Cavalier-Smith, T. Principles of protein and lipid targeting in secondary symbiogenesis: Euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J. Eukaryot. Microbiol.* **1999**, *46*, 347–366. [CrossRef]
121. Cavalier-Smith, T. Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2003**, *358*, 109–133. [CrossRef] [PubMed]
122. Cenci, U.; Bhattacharya, D.; Weber, A.P.M.; Colleoni, C.; Subtil, A.; Ball, S.G. Biotic Host-Pathogen Interactions as Major Drivers of Plastid Endosymbiosis. *Trends Plant Sci.* **2017**, *22*, 316–328. [CrossRef] [PubMed]
123. Curtis, B.A.; Tanifuji, G.; Burki, F.; Gruber, A.; Irimia, M.; Maruyama, S.; Arias, M.C.; Ball, S.G.; Gile, G.H.; Hirakawa, Y.; et al. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **2012**, *6*, 59–65. [CrossRef] [PubMed]
124. Read, B.A.; Kegel, J.; Klute, M.J.; Kuo, A.; Lefebvre, S.C.; Maumus, F.; Mayer, C.; Miller, J.; Monier, A.; Salamov, A.; et al. Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature* **2013**, *499*, 209–213. [CrossRef] [PubMed]
125. Dorrell, R.G.; Gile, G.; McCallum, G.; Meheust, R.; Bapteste, E.P.; Klinger, C.M.; Brillet-Gueguen, L.; Freeman, K.D.; Richter, D.J.; Bowler, C. Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *eLife* **2017**, *6*. [CrossRef] [PubMed]
126. Simion, P.; Belkhir, K.; Francois, C.; Veyssier, J.; Rink, J.C.; Manuel, M.; Philippe, H.; Telford, M.J. A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biol.* **2018**, *16*, 28. [CrossRef]
127. Kumar, S.; Jones, M.; Koutsovoulos, G.; Clarke, M.; Blaxter, M. Blobology: Exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front. Genet.* **2013**, *4*, 237. [CrossRef]

128. Lukes, J.; Leander, B.S.; Keeling, P.J. Cascades of convergent evolution: The corresponding evolutionary histories of euglenozoans and dinoflagellates. *Proc. Natl. Acad. Sci. USA* **2009**, *106* (Suppl. 1), 9963–9970. [CrossRef]
129. Goh, W.W.B.; Wang, W.; Wong, L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends Biotechnol.* **2017**, *35*, 498–507. [CrossRef] [PubMed]
130. Almasi, A.; Pescod, M.B. Wastewater treatment mechanisms in anoxic stabilization ponds. *Water Sci. Technol.* **1996**, *33*, 125–132. [CrossRef]
131. Hauslage, J.; Strauch, S.M.; Eßmann, O.; Haag, F.W.M.; Richter, P.; Krüger, J.; Stoltze, J.; Becker, I.; Nasir, A.; Bornemann, G.; et al. Eu:CROPIS—"Euglena gracilis: Combined Regenerative Organic-food Production in Space"—A Space Experiment Testing Biological Life Support Systems Under Lunar And Martian Gravity. *Microgravity Sci. Technol.* **2018**, *30*, 933–942. [CrossRef]
132. Perez, E. *Analyses Biochimiques, Protéomiques et Transcriptomiques du Métabolisme Énergétique chez L'algue Secondaire verte Euglena gracilis (Euglenozoa, Excavata)*; Université de Liège: Liège, Belgique, 2015.
133. Simpson, A.G.; Stevens, J.R.; Lukes, J. The evolution and diversity of kinetoplastid flagellates. *Trends Parasitol.* **2006**, *22*, 168–174. [CrossRef] [PubMed]
134. Schwartzbach, S.D. Photo and nutritional regulation of Euglena organelle development. *Euglena Biochem. Cell Mol. Biol.* **2017**, 159–182. [CrossRef]
135. Gain, G.; Vega de Luna, F.; Cordoba, J.; Perez, E.; Degand, H.; Morsomme, P.; Thiry, M.; Baurain, D.; Pierangelini, M.; Cardol, P. Trophic state alters the mechanism whereby energetic coupling between photosynthesis and respiration occurs in Euglena gracilis. *New Phytol.* **2021**. (in revision).
136. Koski, L.B.; Golding, G.B. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **2001**, *52*, 540–542. [CrossRef]
137. Triemer, R.E. Feeding in Peranema trichophorum revisited (Euglenophyta) 1. *J. Phycol.* **1997**, *33*, 649–654. [CrossRef]
138. Schnepf, E.; Deichgräber, G. «Myzocytosis», a kind of endocytosis with implications to compartmentation in endosymbiosis. *Naturwissenschaften* **1984**, *71*, 218–219. [CrossRef]
139. Hehenberger, E.; Gast, R.J.; Keeling, P.J. A kleptoplastidic dinoflagellate and the tipping point between transient and fully integrated plastid endosymbiosis. *Proc. Natl. Acad. Sci. USA* **2019**. [CrossRef] [PubMed]
140. Bodyl, A. Did some red alga-derived plastids evolve via kleptoplastidy? A hypothesis. *Biol. Rev. Camb. Philos. Soc.* **2018**, *93*, 201–222. [CrossRef] [PubMed]
141. Yamaguchi, A.; Yubuki, N.; Leander, B.S. Morphostasis in a novel eukaryote illuminates the evolutionary transition from phagotrophy to phototrophy: Description of Rapaza viridis n. gen. et sp. (Euglenozoa, Euglenida). *BMC Evol. Biol.* **2012**, *12*, 29. [CrossRef] [PubMed]

4.3. CRitical Assessment of genomic COntamination detection at several Taxonomic ranks (CRACOT)

CRitical Assessment of genomic COntamination detection at several Taxonomic ranks (CRACOT)

Luc Cornet [abc*] Valérian Lupo[d] Stéphane Declerck[b] Denis Baurain[d]

a BCCM/IHEM, Mycology and Aerobiology, Sciensano, Brussels, Belgium
b BCCM/MUCL and Laboratory of mycology, Earth and Life Institute, Université catholique de Louvain, Belgium
c BCCM/ULC, InBioS–Molecular diversity and ecology of cyanobacteria, University of Liège, Liège, Belgium
d InBioS–PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, Liège, Belgium
* Corresponding author

Abstract

Genome contamination is a well known issue in genomics. Although it has already received a lot of attention, with an increasing number of detection tools made available over the years, no comparison between these tools exists in the literature. Here, we report the benchmarking of six of the most popular tools using a simulated framework. Our simulations were conducted on six different taxonomic ranks, from phylum to species. The analysis of the estimated contamination levels indicates that the precision of the tools is not good, often due to large overdetection but also underdetection, especially at the genus and species ranks. Furthermore, our results show that only redundant contamination is accurately estimated.

Keywords

Genomic contamination; Contaminant levels; Contamination simulations; Horizontal Gene Transfer simulations; Metagenomics

Background

Genomic contamination is a well-known, albeit recurrent, problem in genomics. It appears when a genome, often a Metagenome-Assembled Genome (MAG), contains DNA sequences that do not belong to the expected organism [1]. This umbrella concept actually masks different sources of DNA mis-affiliation, which can occur almost anytime between the selection of a sample and its bioinformatic analysis [1]. Nowadays, genomes are the basis of numerous studies, and it is no longer necessary to demonstrate that genomic contamination is a cause for artifacts, notably in phylogenomic inference [2-4]. Consequently, the detection of contaminants is a topic that has attracted the attention of scientists, with the development of numerous detection tools and an increasing rate of publications over the recent years. Although all these tools ultimately report a quantified level of contamination, they are based on various algorithms and do not measure the same information [1,5]. Indeed, among the most popular tools, two major categories can be distinguished: those relying on the presence of multiple marker genes (e.g., CheckM [6] and BUSCO [7]) and those based on whole-genome surveys (e.g., GUNC [8], Physeter [5] and Kraken2 [9]). Because of these differences in algorithms, Cornet et al 2018 [10] and Lupo et al 2021 [5] have reported the difficulty to meaningfully compare these tools, let alone computing correlations between their estimates. Simulations of genome contamination, in which the exact amount of contaminant sequences is known, can nevertheless be used to overcome such a limitation. In the present study, we compare the detection performance of six of the most used tools (CheckM [6], BUSCO [7], GUNC [8], Physeter [5], Kraken2 [9], and CheckM2 [11]) in order to assess their efficiency. To do so, we used simulations at multiple taxonomic ranks, while varying the contamination scenarios.

Results and Discussion

Regardless of the contaminant source, it is now established that it can be summarized into three main types at the genomic sequence level (**Figure 1**) [1,8]. The first type is **redundant contamination** that occurs when the contaminant sequence is redundant with an homologous genomic sequence of the expected organism [1]. The second type is **replaced contamination** that is similar to the first one, but with the genuine sequence of the expected organism lacking from its genome [1]. The third type is **single contamination** that occurs when the contaminant sequence has naturally no homologous sequence within the genome of the expected organism [1]. To mimic these three situations, we selected 705 high-quality reference genomes belonging to class *Clostridia* and genus *Lactobacillus* and simulated contamination events of the three types (**Figure 1**). Our simulations were performed out at six different taxonomic ranks, from intra-phylum to intra-species.

Surprisingly, our results reveal that, with the exception of Kraken2, none of the tested tools was able to accurately estimate the contamination level (CL) of our combined scenarios, when the three different contamination types were mixed (**Figure 2**). Separated simulations are available in Supplementary materials for redundant (**Figure S1**), replaced (**Figure S2**), and single (**Figure S3**) events. CheckM, based on the duplication of gene markers [6], overestimated the redundant CL (**Figure S1**), but quite logically, does not detect replaced (**Figure S2**) or single (**Figure S3**) contamination events. Similar to its main metric, CheckM's complementary metric used for genetically close contaminants (strain heterogeneity) also overestimated CL, but at the genus and species ranks (**Figure 2**). BUSCO, which is also based on marker duplication [7], largely overestimated the redundant CL (**Figure S1**) at all ranks and, as for CheckM, underdetected replaced (**Figure S2**) and single (**Figure S3**) contamination events. GUNC, which searches for sequence chimerism [8], presents a pattern of both over- and underestimation at four ranks (phylum, class, order and family) (**Figure 2**), with a minimum of 59% of underestimation (see **Table S1** for the percentage of underestimation of each tool at each taxonomic rank). At the genus and species ranks, GUNC only underestimated CL (**Figure 2**), notably for replaced events where it detected nothing (**Figure S2**). Physeter, which is based on Lowest Common Inference (LCA) of DIAMOND blastx [12] hits [5], overestimates CL at all ranks for all types of contaminants (**Figure 2**). In contrast, Kraken2, which takes advantage of exact long kmer matching [9], showed the best estimation of CL, fitting well to the simulations, with the exception of the species rank, which was largely underestimated (see **Table S1**). It is noteworthy that the genomes used in our simulations were of high quality (see Online Methods) and included in the Kraken2 database. Owing to its exact kmer matching algorithm [13], one cannot exclude that Kraken2 would perform less well on rare genomes, compared to our simulations. CheckM2, which uses machine learning based on genomic contamination simulation (gradient boost model) without relying on taxonomic information, [11], largely overestimated the redundant CL (**Figure S1**), especially at the genus and species ranks. Replaced (**Figure S2**) and single (**Figure S3**) CL were underestimated at all ranks, with the exception of the single type at the genus and species ranks. Percentages of underestimation(**Table S1)** showed that CheckM2 underestimates CL in more than 97% of the cases for both replacement and single events, while it never underdetected the redundant type (to the exception of the species rank in 1.3% of the cases). To overcome the impossibility to directly correlate the performance of the different tools (due to their algorithmic differences), we computed the correlation of each tool to the expected CL of our simulations. The tools, not including Kraken2, correlated badly, often negatively, with the expected CL level of the simulations, the correlation coefficient (R2) never going beyond 0.37 (**Figure 2**, **Figures S1-S3**).

Beside genomic contamination, another kind of genomic exchange naturally affects genomes: horizontal gene transfer (HGT). One of the major differences between HGT and contamination is that the first one accumulates mutations in the receiver (and donor) organisms [14] after transfer whereas contamination occurs shortly before or after genome sequencing, hence contaminant sequences are exact matches between donor and receiver genomes [1]. To investigate the effect of HGT on detection performance, a non-null mutation rate was optionally enabled during the simulations (see Online Methods), either at 10% (**Figure S4**) or 25% (**Figure S5**). None of the tools (to the exception of Physeter) conflated contamination and HGT, which suggested that HGT events should not increase CL on real data. While reassuring, a possible drawback is that if the contaminant sequence is an HGT, it has few chances to be detected. This can be damaging since HGT frequently occurs in bacteria [15-19]. Somewhat ironically, the inability of Physeter to differentiate between HGT and genomic contamination indicates that LCA algorithms are likely to prove useful in this case, even if too conservative due to their inclination for overdetection.

Conclusion

We conducted this study because no tools comparison, despite the availability of no less than 18 programs, had been published to date, raising the question in the community: “which tool should we use?” Our results have demonstrated that CL is frequently overestimated, resulting in unwarranted removal of sometimes precious genomes. Nevertheless, especially at the genus and species ranks, the odds of underestimation are always significant. This is a matter of concern because the risk of contamination by closely related taxa is higher when dealing with MAGs. We have also demonstrated that the replaced and single contamination types suffer less from underestimation compared to the redundant events. The results of this study are all the more surprising asour simulations were rather simple. Furthermore, simulations were conducted with only one contaminant genome, at low CL, while contamination by more than one taxon, at high CL, regularly occurs in public repositories [10,5]. Our conclusion is that, given the current algorithmic state of the field, which requires more innovation, users should use a combination of tools to estimate CL, and one of these tools should be Kraken2. Our contamination simulation framework, CRACOT, is freely available as a Nextflow workflow [20], sustained by a Singularity container [21], at https://github.com/Lcornet/GENERA/wiki/20.-CRACOT. It might be useful in future projects, for example to estimate the accuracy of new tools underdevelopment.

Online Methods

Contamination simulations (overview of CRACOT)

Our genome contamination simulations were carried out with the Nextflow workflow CRACOT, freely available at https://github.com/Lcornet/GENERA/wiki/20.-CRACOT.

705 high quality genomes belonging to class *Clostridia* and genus *Lactobacillus* were selected as input for CRACOT. These genomes were selected based on the GUNC [8] clade separation score (CSS), which measures the chimerism of genome contigs. Furthermore, we imposed on these genomes to have no more than five contigs with no 'N' within each contig. The contamination values of these genomes for the six tools are available in **Table S2**. The median contamination was 0.45% for CheckM V1.2.1 [6], 0.02% for GUNC V1.0.5 [8], 0.87% for BUSCO V5.4.3 [7], 22,3% for Physeter V0.213470 [5], 2,45% for Kraken2 V2.1.2 [9], 8% for CheckM2 V0.1.3 [11].

The first step of CRACOT, **Figure 1**, was to create random genome pairs, one genome being considered hereafter as the main "expected" organism and the second as the slave "contaminant" organism. The pairing, based on the NCBI Taxonomy [22,23] provided by Bio-Must-Core V0212670 (https://metacpan.org/dist/Bio-MUST-Core), associated with the genomes was made for one specific taxonomic rank, ranging from phylum to species. When a rank is selected, the two genomes should belong to the same taxon at this rank, but have a different taxonomy starting with the next rank. For instance, the phylum rank (e.g., Firmicutes) means that the two genomes belong to the same phylum but not to the same class (e.g., if one genome belongs to Bacilli, the other genome belongs to Clostridia).

The plasmids of the selected genomes were removed after the pairing step, so as to not interfere with the detection of contamination. Removal was performed with PlasmidPicker (https://github.com/haradama/PlasmidPicker) with default settings. Proteins werethen predicted with Prodigal V2.6.3 [24], used with default settings. Finally, OrthoFinder V2.5.4 [25], with default settings, wasused for orthologous inference.

The three types of contamination were simulated based on the common and single protein orthogroups (OGs). Common proteins weredefined as proteins present in only one copy for both the main and slave genome in the OG while single proteins were singletons of the slave genome. Duplicated contamination events were fished from the pool of common OGs, and the corresponding gene sequences of the slave genome were added to the end of the last contig of the master genome (with a serie of five 'N' added to either side of the gene). Replaced contamination events were also fished from the pool of common OGs but slave genes replaced the genuine genes within the main genome. Single contamination events were fished from the pool of singletons of the slave organism, and the corresponding gene sequences were added to the end of the last contig of the master genome, as above. The number of events of each type is a user-specified option. At each simulation, 150 chimeric genomes were asked as CRACOT output, but the real output number depends on the number of available common and single protein OGs. The number of simulated genomes used in this study are given in **Table S3**. Chimeric levels of the simulations are indicated in **Table S4**.

HGT can be simulated for each of the three sequence types. The mutation rate was computed with HgtSIM [26], with the rate option set at 1-0-1-1 so that a mutation rate in DNA sequences corresponds to the same simulation rate in the proteins [26].

CRACOT was used to simulate contamination events, not only the redundant, replaced or single type separately, but also as a combination of the three types. Two HGT simulations for a combination of the three contamination types, with a mutation rate of 10% and 25%, were also generated.

Genomic contamination estimation

Genomic contaminants were estimated using the Nextflow workflow GENcontams (https://github.com/Lcornet/GENERA/wiki/09.-Genome-quality-assessment) from the GENERA project [27]. CheckM V1.2.1 [6] was used with the lineage_wf option and the provided database. GUNC V1.0.5 [8] was used with default settings and the Progenomes 2.1 database [28]. BUSCO V5.4.3 [7] was used in auto-lineage mode and the provided database. BUSCO's number of duplicated markers was used as a proxy for the contamination level. Physeter V0.213470 was used with the auto-detect option and the database provided in Lupo et al. (2021) [5]. Kraken 2 V2.1.2 [9] was used with default settings and the database 'PlusFP' downloaded from https://benlangmead.github.io/aws-indexes/k2. Kraken2 levels of contamination were computed with the Physeter parser with the auto-detect option set to 'count_first'. The list of taxa used by the Physeter parser was automatically produced by the create-labeler.pl script using the list of genera found in the nodes.dmp file from the local mirror of NCBI Taxonomy. CheckM2 V0.1.3 [11] was used with default settings and the provided database.

Correlation and violin plot creation

Spearman correlations between the CL estimates of the tools and the simulated levels of contaminants, as created by CRACOT, were computed with R [29]. Violin plots were created with ggplot [30]. The R code for the creation of these plots is available at https://github.com/Lcornet/GENERA/wiki/21.-Supplemental-Scripts#4-make-cracot-tablepy-and-cracot-rscript.

Abbreviations

Metagenome-Assembled Genome (MAG)
contamination level (CL)
horizontal gene transfer (HGT)
clade separation score (CSS)
orthogroups (OGs)

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

CRACOT is freely available at https://github.com/Lcornet/GENERA/wiki/20.-CRACOT.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by a research grant (no. B2/191/P2/BCCM GEN-ERA) financed by the Belgian State – Federal Public Planning Science Policy Office (BELSPO). Computational resources were provided by the Consortium des Équipements de Calcul Intensif (CÉCI) funded by the F.R.S.-FNRS (2.5020.11), and through two research grants to DB: B2/191/P2/BCCM GEN-ERA (Belgian Science Policy Office - BELSPO) and CDR J.0008.20 (F.R.S.-FNRS).

Authors' contributions

LC and DB conceived the study. LC developed CRACOT and performed all analyses. VL developed the parser of Kraken2, used for CL estimation. LC and DB wrote the manuscript with the help of VL and SD.

Acknowledgements

Not applicable.

References

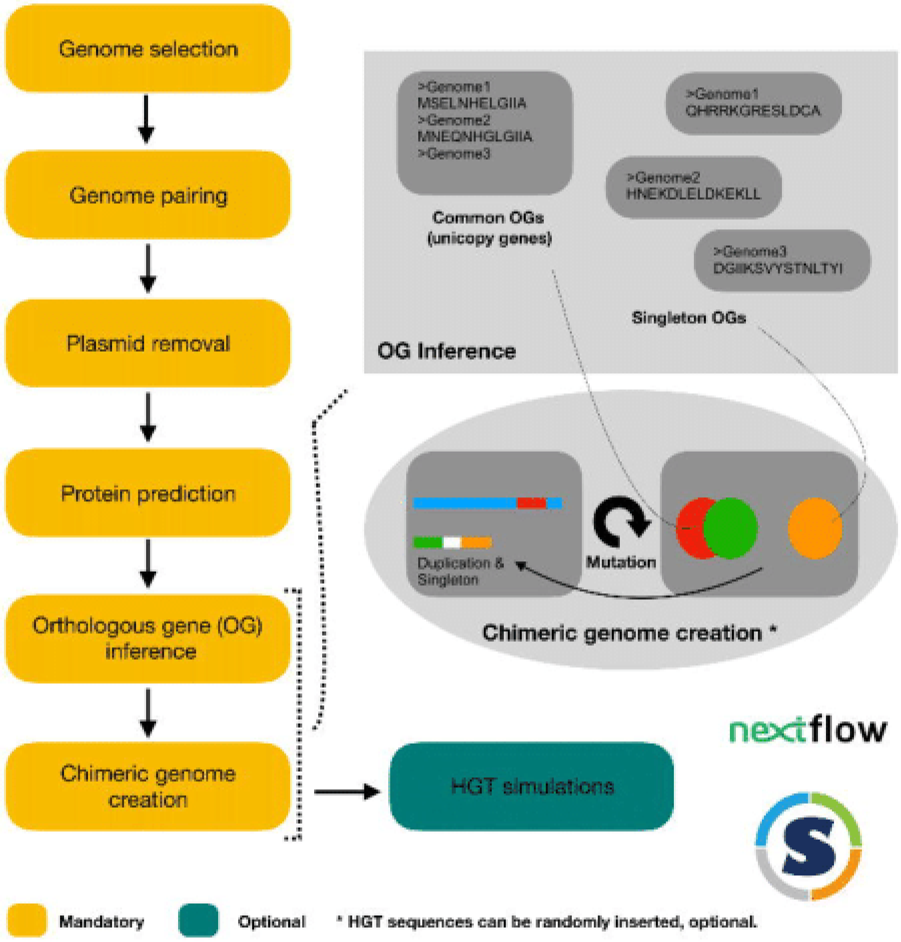
1. Cornet L, Baurain D. Contamination detection in genomic data: more is not enough. Genome Biology. 2022;23:60.
2. Schierwater B, Kolokotronis S-O, Eitel M, Desalle R. The Diploblast-Bilateria sister hypothesis: parallel evolution of a nervous systems in animals. Communicative & integrative biology. 2009;2:403–5.
3. Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, et al. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLOS Biology. 2011;9:e1000602.
4. Laurin-Lemay S, Brinkmann H, Philippe H. Origin of land plants revisited in the light of sequence contamination and missing data. Current Biology. 2012;22:R593–4.
5. Lupo V, Van Vlierberghe M, Vanderschuren H, Kerff F, Baurain D, Cornet L. Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics. Frontiers in Microbiology. 2021;12:3233.
6. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
7. Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. arXiv:210611799 [q-bio] [Internet]. 2021 [cited 2021 Oct 4]; Available from: http://arxiv.org/abs/2106.11799
8. Orakov A, Fullam A, Coelho LP, Khedkar S, Szklarczyk D, Mende DR, et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biology. 2021;22:178.
9. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. bioRxiv. 2019;762302.
10. Cornet L, Meunier L, Vlierberghe MV, Léonard RR, Durieu B, Lara Y, et al. Consensus assessment of the contamination level of publicly available cyanobacterial genomes. PLOS ONE. 2018;13:e0200323.
11. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning [Internet]. bioRxiv; 2022 [cited 2022 Aug 28]. p. 2022.07.11.499243. Available from: https://www.biorxiv.org/content/10.1101/2022.07.11.499243v1
12. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature Methods. 2015;12:59–60.
13. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology. 2014;15:R46.
14. Arnold BJ, Huang I-T, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol. 2021;1–13.
15. Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events.

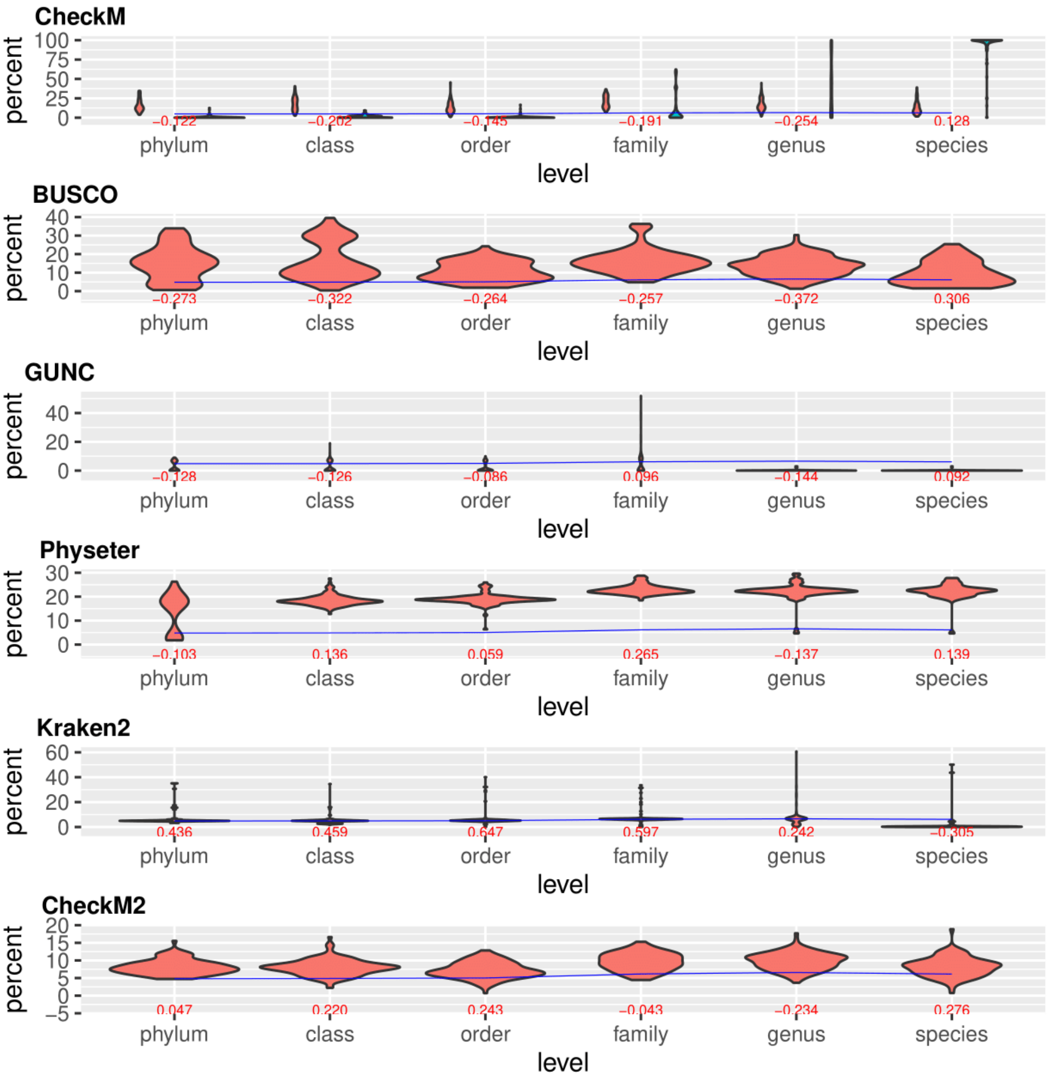
Genome Res. 2006;16:1099–108.
16. Dagan T, Martin W. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. PNAS. National Academy of Sciences; 2007;104:870–5.
17. Dagan T, Artzy-Randrup Y, Martin W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. PNAS. National Academy of Sciences; 2008;105:10039–44.
18. Bohr LL, Mortimer TD, Pepperell CS. Lateral Gene Transfer Shapes Diversity of Gardnerella spp. Front Cell Infect Microbiol [Internet]. Frontiers; 2020 [cited 2020 Dec 30];10. Available from: https://www.frontiersin.org/articles/10.3389/fcimb.2020.00293/full?report=reader#h3
19. Frazão N, Sousa A, Lässig M, Gordo I. Horizontal gene transfer overrides mutation in Escherichia coli colonizing the mammalian gut. PNAS. 2019;201906958.
20. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature Biotechnology. Nature Publishing Group; 2017;35:316–9.
21. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLOS ONE. 2017;12:e0177459.
22. Federhen S. The NCBI Taxonomy database. Nucleic Acids Research. 2012;40:D136–43.
23. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database. 2020;2020:baaa062.
24. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
25. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology. 2019;20:238.
26. Song W, Steensen K, Thomas T. HgtSIM: a simulator for horizontal gene transfer (HGT) in microbial communities. PeerJ. 2017;5:e4015.
27. Cornet L, Durieu B, Baert F, D'hooge E, Colignon D, Meunier L, et al. The GEN-ERA toolbox: unified and reproducible workflows for research in microbial genomics [Internet]. bioRxiv; 2022 [cited 2022 Nov 14]. p. 2022.10.20.513017. Available from: https://www.biorxiv.org/content/10.1101/2022.10.20.513017v1
28. Mende DR, Letunic I, Maistrenko OM, Schmidt TSB, Milanese A, Paoli L, et al. proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. Nucleic Acids Research. 2020;48:D621–5.
29. R Core Team. R: a language and environment for statistical computing. [Internet]. 2014. Available from: https://www.R-project.org/
30. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. [cited 2019 Aug 24]. Available from: https://ggplot2-book.org/

Figures

Figure 1: Flowchart of CRACOT.
CRACOT is a Nextflow workflow, supported by a Singularity container. It is a six step program. The first step is the genome selection according to a user-specified list. The second step is the association of genomes, by pairs of the same taxonomic group. Step 3 to 5 corresponds to the removal of plasmid, protein prediction and orthology inference. Finally, genome contamination simulations are based on the information produced during the orthology inference, the common (to both the expected and contaminant organisms) and single (of only the contaminant organism). Common genes are used for redundant and replaced contamination events while singleton are used for single contamination events. Optionally, a mutation rate can be enabled for each of these three basic types to simulate horizontal gene transfer.

Figure 2: Contamination estimation, at six taxonomic ranks, of the combined types of contamination.
Simulations were performed with a combination of the three contamination types (redundant, replaced, single). The median values of the contamination level (%CL) of these simulations are indicated by the blue line, while these CL estimated by the six tools are summarized by the violin plots. Spearman correlation values between the estimate of each tool and the simulated level of contamination are indicated in red.


Genome selection
Genome pairing
Plasmid removal
Protein prediction
Orthologous gene (OG) inference
Chimeric genome creation
HGT simulations
>Genome1
MSELNHELGIIA
>Genome2
MNEQNHGLGIIA
>Genome3
Common OGs
(unicopy genes)
>Genome1
QHRRKGRESLDCA
>Genome2
HNEKDLELDKEKLL
>Genome3
DGIIKSVYSTNLTYI
Singleton OGs
OG Inference
Duplication &
Singleton
Mutation
Chimeric genome creation *
nextflow
Mandatory
Optional
* HGT sequences can be randomly inserted, optional.


CheckM
percent
100
75
50
25
0
−0.122
−0.202
−0.145
−0.191
−0.254
0.128
phylum
class
order
family
genus
species
level
BUSCO
percent
40
30
20
10
0
−0.273
−0.322
−0.264
−0.257
−0.372
0.306
phylum
class
order
family
genus
species
level
GUNC
percent
40
20
0
−0.128
−0.126
−0.086
0.096
−0.144
0.092
phylum
class
order
family
genus
species
level
Physeter
percent
30
20
10
0
−0.103
0.136
0.059
0.265
−0.137
0.139
phylum
class
order
family
genus
species
level
Kraken2
percent
60
40
20
0
0.436
0.459
0.647
0.597
0.242
−0.305
phylum
class
order
family
genus
species
level
CheckM2
percent
20
15
10
5
0
−5
0.047
0.220
0.243
−0.043
−0.234
0.276
phylum
class
order
family
genus
species
level