Multi-Proteins Similarity-based sampling to select representative genomes from large databases

Rémi-Vinh Coudert[1,2,*], Jean-Philippe Charrier[2], Frédéric Jauffrit[2], Jean-Pierre Flandrois[1,*], Céline Brochier-Armanet[1,*]

[1] Université de Lyon, Université Lyon 1, CNRS, VetAgro Sup, Laboratoire de Biométrie et Biologie Evolutive, UMR5558, Villeurbanne, France

[2] Microbiology Research & Development, bioMérieux SA, 376 Chemin de l'Orme, 69280 Marcy L'Etoile, France

*To whom correspondence should be addressed.

Abstract

Motivation: Genomic sequence databases are growing exponentially, but with high redundancy and uneven data quality. For these reasons, most studies rely on samples of available genomic data. Current sampling approaches rely either on taxonomic, whole-genome similarity, or phylogenetic criteria. However, none of these approaches can select reliable representative genomes from huge data sets.

Results: Here we present MPS-Sampling (Multiple-Protein Similarity Sampling), a fast, scalable, and efficient method for selecting reliable representative samples of genomes from very large data sets. Using sets of single-copy protein families as input (e.g., core proteins, ribosomal proteins), MPS-Sampling delineates homogeneous groups of genomes through three successive clustering steps. Representative genomes are then selected within these groups according to quality and centrality criteria.

MPS-Sampling was applied to a set of 48 ribosomal protein families present in 178,203 bacterial genomes. MPS-Sampling generated a series of representative genome samples, reducing the initial dataset by 68% to 98%. Reduction using MPS-Sampling took 1.50 CPU hours, 150 times faster than Mash and 100,000 times faster than FastANI applied to the same data. Sample analysis showed that the selected genomes were both taxonomically and phylogenetically representative of the original dataset, demonstrating the relevance of the approach implemented in MPS-Sampling.

Availability and Implementation: MPS-Sampling is an open-source software available under CeCILL v2.1 at https://github.com/rvcoudert/MPS_Sampling.

Contact: remi.coudert@univ-lyon1.fr, celine.brochier-armanet@univ lyon1.fr

Supplementary information: Supplementary Material are available at Bioinformatics online.

Introduction

The burst of genome sequencing provides a wealth of data and an ever-increasing access to genetic information, including from uncultured organisms (Stephens *et al.*, 2015). However, available genomic data are growing in an unbalanced way, both in terms of quality, as most released genomes are in fact rough draft assemblies, and diversity, with the over-representation of a few species and taxonomic groups, reflecting socio-economic considerations (Land *et al.*, 2015) (Supplementary Material S1-S2). This huge amount of data makes exhaustive analyzes complicated, costly regarding computation time, or even technically impossible. For these reasons, most analyses rely on samples of representative genomes (Land *et al.*, 2015). Current sampling strategies are mainly based on either taxonomy, whole-genome similarity or phylogeny, but they all have the same objective: finding the best balance between sample size and representativeness, and proceed in the same way: the grouping of genomes, then the selection of representatives at random (Garcia *et al.*, 2021) or according to *a priori* defined criteria (e.g., type strains, model organisms, size or completeness of the genomes, external references) (Parks *et al.*, 2017; Chun *et al.*, 2018).

Taxonomy-based approaches group genomes according to their taxonomic affiliation (Chun *et al.*, 2018; Garcia *et al.*, 2021). Although taxonomy-based approaches are relatively easy to apply, they have strong limitations. For instance, many genomes have incomplete (or even no) taxonomic assignment (Supplementary Material S2). Despite they represent a significant part of the biodiversity, e.g. up to 60% in the study by Parks *et al.* (Parks *et al.*, 2020), these genomes are usually ignored, which could lead to significant omissions. Furthermore, taxonomy-based approaches are sensitive to assignation errors (Chun *et al.*, 2018) and hampered by historic legacy (Lan and Reeves, 2002). This issue is particularly problematic because the proportion and amplitude of these taxonomic errors is unknown. Finally, taxonomy is only a rough proxy of the genetic/phylogenetic diversity, as the criteria used to define species and supra-species taxa may vary depending of the considered groups (Chun *et al.*, 2018). For instance, although prokaryotic species are defined using whole-genome similarity (Varghese *et al.*, 2015; Olm *et al.*, 2020), some species are diverse enough to be separated into several distinct species, while distinct species are closely related enough to be merge into a single one (Rosselló-Móra and Amann, 2015). Regarding supra-species taxa, their delineation is usually based on the phylogenetic analysis of one or several markers (Chun *et al.*, 2018). However, phylogenetic relationships can be difficult to determine especially when the number of genomes is huge (Philippe *et al.*, 2017), not to mention that there is no consensus regarding the criteria used to delineate supra-species taxa (Chun *et al.*, 2018).

Most whole-genome similarity-based methods rely on the computation of Overall Genome Relatedness Indexes (OGRI) (Chun and Rainey, 2014). For instance, the Average Nucleotide Identity (ANI) is a major alignment-based method measuring the similarity between two genomes (Goris *et al.*, 2007; Palmer *et al.*, 2020), while the alignment-free method Mash uses hashing values to approximate the shared ratio of k-mers (Ondov *et al.*, 2016; Zhu *et al.*, 2019). Whole-genome similarity-based approaches are taxonomically and phylogenetically agnostic. However, they have also some limits. First, they are based on genome pairwise comparisons and are therefore constrained by the quadratic complexity $O(n^2)$ (Chun and Rainey, 2014), meaning that the computation time increases quadratically according to the number of genomes. To overcome this issue, they are guided either by taxonomy-defined representative genomes, used as centroids (Nayfach *et al.*, 2021), or used divide-and-conquer strategies using random or taxonomy-defined chunks (Bursteinas *et al.*, 2016; Léonard *et al.*, 2021; Almeida *et al.*, 2021). Second, the accuracy of OGRIs strongly decreases as the distance between genomes increases (i.e. above the species or genus level) (Qin *et al.*, 2014).

In phylogeny-based approaches, genomes are grouped according to their relatedness, which requires the inference of phylogenetic trees using either genome gene content or the phylogenetic signal carried by sets of conserved core gene/protein families. Main criteria used to delineate clusters of related genomes are tree topology, branch supports, and patristic distances (i.e. branch lengths). Several tools have been developed to delineate clusters in phylogenetic trees: (Cluster Picker (Ragonnet-Cronin *et al.*, 2013), RPANDA (Morlon *et al.*, 2016), PhyCLIP (Han *et al.*, 2019), AncestralClust (Pipes and Nielsen, 2022), Tree Pruner (Krishnamoorthy *et al.*, 2011), TreeTrimmer (Maruyama *et al.*, 2013)). However, most of them have limits (e.g., manual curation steps, taxonomy dependency) and use single-linkage clustering, which provides unbalanced clustering sensitive to chaining effect (see Discussion) (Menardo *et al.*, 2018). Furthermore, most of them cannot handle very large phylogenetic trees (>150,000 leaves) due to computational time and/or memory usage. To our knowledge, TreeCluster is likely the most appropriated approach, as it uses a reliable hierarchical clustering with either single-linkage, complete-linkage, or sum-length, and is able to process huge trees. However, TreeCluster does not include tools to select representative genomes. Finally, it is important to remember that the inference of phylogenetic trees is time-consuming and that the quality of trees decreases as the number of sequences or their divergence increase (Philippe *et al.*, 2017), which may have an impact on the relevance on the clustering and thus on the selection of representative genomes.

Several international consortia are providing sets of representative genomes (Supplementary Material S3), regarding biomedical and biotechnology research, while providing broad coverage of the Tree of Life. The advantage is that users do not have to manage the data sampling step, but the disadvantage is that users have no control over the sampling density, data redundancy, and data update. At last but not least, most of these samples do not include genomes of undescribed organisms.

To sum-up, current approaches to select representative genomes are not completely satisfactory, because they have important limitations and are not well suitable to process large collection of genomes. In our opinion, an ideal approach should (i) handle very large datasets in an acceptable computation time, (ii) be taxonomy-agnostic, to consider genomes with incomplete (or no) taxonomic affiliation, (iii) be phylogeny-independent, (iv) allow variable sampling densities, (v) use user-defined priority criteria for the selection of representative genomes, and (vi) be reproducible.

To address these needs, we developed MPS-Sampling (Multi Protein Similarity-based Sampling), a fast, scalable and reliable method to select representative genomes. Through three steps of clustering and matrix computation, MPS-Sampling overcomes quadratic complexity, enabling the processing of very large datasets. MPS-Sampling uses families of homologous proteins as input and provides in return genome sets of variable density, representative of both taxonomic and phylogenetic diversity of the initial dataset. MPS-Sampling was applied on 178,203 bacterial genomes, generating seven samples in 1.50 CPU hour using 48 ribosomal protein families as input.

Materials and Methods

1. MPS-Sampling workflow

The five steps of the workflow are illustrated in Figure 1 and in Supplementary Materials S4-S10.

Input: Protein sequences grouped into homologous families

MPS-Sampling uses families of single-copy homologous protein sequences as input. MPS-Sampling can handle missing data, meaning that protein families are not expected to be present in all genomes. The input order of data (i.e. genomes or protein families) does not impact the sampling process.

Step 1: Construction of Lin-clusters

Within each protein family, sequences are clustered using Linclust (Steinegger and Söding, 2018) of the MMseqs2 suite (Steinegger and Söding, 2018). Linclust was chosen for its efficiency, very high specificity, high sensitivity, and near-linear complexity (Steinegger and Söding, 2018). Well adapted to process huge datasets, Linclust is used for dereplication by two reference databases: RefSeq (Li *et al.*, 2021) and UniRef (Bateman *et al.*, 2017). Linclust identifies putative pairs of sequences through a kmer-based heuristic. The relevance of each pair is then evaluated based on sequence alignment using three parameters: the alignment e-value (eValue), coverage (minCov), and sequence identity (minSeqID). Confirmed links are then used to build sequence clusters (referred hereafter to as **Lin-clusters**) using the greedy set cover algorithm. The delineation of the Lin-clusters is a key information of MPS-Sampling, that is used for the three rounds of genome clustering (steps 2-4).

Step 2: Construction of elementary groups of genomes

This first round of genome clustering aims at gathering very close genomes into elementary groups of genomes (EGG). Lin-clusters are numbered from largest to smallest; the Lin-cluster encompassing the largest number of sequences being designated as 1 (step 2-1). Then, genomes are labeled according to the Lin-clusters to which their sequences belong; these genome labels are stored in the **Lin-clustering matrix** (step 2-2). Last, redundant lines of the Lin-clustering matrix are merged, leading to a smaller and non-redundant matrix, called the **Lin-combination matrix** (step 2-4). Each Lin-combination corresponds to a group of genomes, called elementary groups of genomes (EGG). This delineation is very stringent, as genomes differing by only one Lin-cluster will be put into distinct EGG. Within EGG, genomes are considered indistinguishable.

Step 3: Construction of pre-connected components

This second round of genome clustering aims at gathering close EGG into rough groups, called **pre-connected components**. To be computed very fast, EGG are gathered through an aggregation process. More precisely, from a starting EGG, each pre-connected component is expanded to any EGG sharing at least an user-defined minimum number of Lin-clusters (minNbLinclusters) (Supplementary Material S4).

Step 4: Construction of MPS-clusters

This third and last clustering aims at gathering close EGG into fine groups of EGG, called **MPS-clusters**, within pre-connected components. The similarity between two EGG is measured by the **Dice index**, also referred as Dice similarity (Dice, 1945), which corresponds to the proportion of Lin-clusters shared by two EGG (Supplementary Material S5). Dice indexes computed between all pairs of EGG are stored in the **similarity submatrix**. By restricting the computation of dice indexes within pre-connected components, only the most informative parts of the whole similarity matrix are computed, avoiding quadratic complexity (step 4-1). At the beginning, each EGG is considered as a MPS-cluster. Then, an aggregative hierarchical method with complete-linkage is used to group

MPS-clusters (step 4-2, Supplementary Material S6). The closest MPS-clusters are aggregated iteratively, two by two, provided their similarity is greater than a user-defined threshold called **minimum similarity (Δ)**. Thus, Δ represents the minimal proportion of common Lin-clusters shared between genomes within a given MPS-Cluster. The similarity between two MPS-clusters correspond to the Dice index of their most dissimilar Lin-combinations. For instance, when Δ = 0.7, two genomes belonging to the same MPS-Cluster will share 70% of their Lin-clusters, i.e. 70% of their protein sequences used as input belong to similar Lin-clusters. When Δ = 1, steps 3 and 4 have no effect and MPS-clusters correspond to the EGG.

Step 5: Selection of MPS-representatives

The final step of MPS-Sampling consists in the selection of one representative genome, called **MPS-representatives**, within each MPS-cluster. The choice of the MPS-representatives relies on fame, protein family distribution, and centrality criteria (Supplementary Material S7). More precisely, genomes considered as the most representative by the community and with the best fame will be chosen preferentially (e.g., RefSeq genomes, type strains). Then, those with the largest distribution across protein families used as input are favored to avoid the selection of incomplete genomes. Finally, genomes which are the most central genomes, i.e. with the most frequent Lin-clusters within the MPS-cluster, are chosen (Supplementary Material S8). If several genomes meet these three criteria, they will be distinguished using a pseudo-random criterion. This ensures the reproducibility of MPS-Sampling.

Output: List of MPS-representatives

MPS-Sampling returns the list of MPS-representative genomes, as well as all intermediate results (e.g. the Lin-clustering matrix, the Lin-combination matrix, the similarity submatrix of each pre-connected component, the link between each input genome and its MPS-representative).

2. Technical details

MPS-Sampling is based on Snakemake, a scalable bioinformatics workflow engine based on Python (Köster and Rahmann, 2012). The environment is managed by Conda, which installs automatically all the needed dependencies. The main dependency is Linclust of the MMseqs2. All intermediate scripts are written in R. The structure of the MPS-Sampling Snakemake pipeline is detailed in a flowchart and an entity relationship diagram (ERD) (Supplementary Material S9-S10).

3. MPS-Sampling run

MPS-Sampling has been run on the bacterial sequences of RiboDB v15.0 (Feb 2023) (Jauffrit *et al.*, 2016), a dedicated database gathering ribosomal proteins (r-prots). To limit biases induced by large amounts of missing data, the eight r-prot families present in less than 80% of the bacterial genomes and the 19,189 bacterial genomes containing less than 80% of the r-prot families were not considered (Supplementary Material S11-S12). Multiple-copy sequences, representing ~1% of the 8,436,399 sequences, have been discarded (Supplementary Material S13). Altogether, the bacterial dataset encompasses 8,315,939 single-copy sequences spread over 48 r-prot families and 178,203 genomes. 20,798 (12%) of the genomes come from Genbank, 157,405 (88%) from RefSeq, among which 16,135 (9%) are labeled as RefSeq-representative genomes. Genomes were labelled according to NCBI assembly accessions and versions. According to the NCBI taxonomy. 135,315 genomes (76%) had complete taxonomic information, meaning that each relevant taxonomic level (i.e. phylum, class, order, family, genus, and species) is defined.

4. MPS-Sampling run

MPS-Sampling parameters were optimized according to a standardized process (Supplementary Material S14-S20) and set as follow: eValue = 10^{-5}, coverageMode = 0, minCov = 0.8, minSeqID = 0.6, minNbLinclusters = 25, and seven values of Δ (1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4).

5. Phylogeny

Phylogenetic inferences are detailed in the Supplementary Material S21. Phylogenetic tree figures were drown using iTOL (Letunic and Bork, 2021).

Results

MPS-Sampling runs

Using 48 r-prot families from 178,203 genomes, seven MPS-samples of decreasing density were computed (1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4). Using a single-thread process, a single MPS-sampling run required 38 CPU min, while the seven MPS-samplings required 86 CPU min (Supplementary Material S22). All intermediate results are detailed in Supplementary Material S23-S29 and all final results in Supplementary Material S30-S37. As expected, the higher the value of Δ, the lower the number of MPS-representatives: from 57,332 (Δ = 1) to 3,474 (Δ = 0.4). This corresponded to a sampling of 32.17% down to 1.95% of the bacterial dataset (Figure 2A). When Δ = 0.7, only one-third of the MPS-clusters (4,909 out of 12,775) contained and, due to priority rules, were represented by a RefSeq-representative genome. These 4,909 MPS-clusters are representative of 159,235 genomes (90%) (Supplementary Material S38). Conversely, two-thirds of the MPS-clusters (7,866 out of 12,775) representing 18,968 genomes (10%) did not contain any RefSeq-representative genome. This indicated that RefSeq-representative genomes only partially represented the sequenced bacterial diversity. Therefore, using RefSeq-representative genomes as proxy of the bacterial diversity might be misleading, as it would lead to an underestimation of the true bacterial diversity, ignoring potentially two-thirds of the genetic diversity. The reliability of MPS-sampling to capture diversity of the bacterial dataset was evaluated according to taxonomic and phylogenetic criteria. Although they are inadequate to guide the selection of representative genomes, they constitute interesting external references to evaluate the quality of a sample.

Taxonomic representativeness of the samples

As expected, decreasing the density of the sampling led to the progressive reduction of the number of species (Δ ≤ 0.9), genus (Δ ≤ 0.8), families (Δ ≤ 0.6), and eventually orders and classes (Δ = 0.4)(Figure 2B). At the same time, the average number of retained genomes at each taxonomic level decreased: from eight (complete dataset) genomes to one (Δ ≤ 0.9) for species, 41 to one (Δ ≤ 0.5) for genera, 249 to three (Δ = 0.4) for family, from 641 to nine (Δ = 0.4) for orders, and so on (Supplementary Material S38C). However, even when only 1.95% of genomes were selected (Δ = 0.4), almost all (92%) high-level taxa (orders, classes, phyla) were still represented in the sample (Figure 2B). These results illustrated the flexibility of MPS-Sampling, which, with appropriate parameters, can be used to dereplicate redundancy and sample diversity at the desired taxonomic level, while avoiding most biases inherent to taxonomy-based sampling methods.

Being taxonomy-agnostic, MPS-Sampling efficiently processed genomes with incomplete taxonomy. These genomes represented 24% of the bacterial dataset and were proportionally less reduced than genomes with complete taxonomy (Figure 2C). This showed that genomes with incomplete taxonomy were less redundant (and thus less dereplicated) than genomes with complete taxonomy. This is not surprising, given that the best described taxa are also the most studied and sequenced.

Phylogenetic representativeness of the samples

As the sample size decreased, genetic redundancy of MPS-representatives was expected to decrease, while phylogenetic diversity increased. This could be tested by monitoring the ratio between the length the tree inferred with r-prots of the sampled genomes divided by the number of leaves. This ratio is equal to 0.0159 for the complete bacterial dataset, then it increased linearly as Δ decreased, reaching 0.0489 (Δ = 1) to 0.3049 (Δ = 0.4), confirming that discarded genomes were indeed the most redundant according to phylogenetic criteria (Figure 2D). Interestingly, the ratio increased linearly whether we considered the complete dataset, or the subset of genomes with either complete or incomplete taxonomy (Supplementary Material S39C), although they had very different initial phylogenetic diversity (0.0159, 0.0072 and 0.0549, respectively). Altogether, these results showed that MPS-Sampling was successful in capturing the phylogenetic diversity of the complete bacterial dataset.

Phylogenetic distribution of the MPS-representatives

To take this a step further, we mapped MPS- samples with the lowest density (Δ = 0.7 to Δ = 0.4) onto a reference bacterial phylogeny. Because the tree inferred with 178,203 genomes was difficult to visualize, we reduced the number of considered genomes with the goal of best representing the diversity of the 178,203 genomes

(Supplementary Material S38). More precisely, we kept all the 16,135 RefSeq-representative genomes, as they are considered as a standard against which other data should be compared (O'Leary *et al.*, 2016). When Δ = 0.7, these genomes were representative and/or belonged to 4,909 MPS-clusters, that together represent 159,235 genomes (90%). We therefore assumed that these 16,135 RefSeq-representative genomes were a reliable reference to represent these 159,235 genomes. The remaining 18,968 genomes (10%) were distributed across the 7,866 MPS-clusters that did not contain any RefSeq-representative genomes. Because these genomes could not be linked to any reference external to our analysis, we considered all of them for the phylogenetic analysis. Thus, in total, 35,103 genomes were used to infer a reference phylogeny from the bacterial dataset (Figure 3). As expected, the sampling density across the tree decreased when Δ decreased, while maintaining good tree coverage, even when the sample contained only 1.95% of the genomes (Δ = 0.4). This showed that, although based on sequence comparisons only, MPS-Sampling was able to reliably capture the phylogenetic diversity of the complete bacterial dataset.

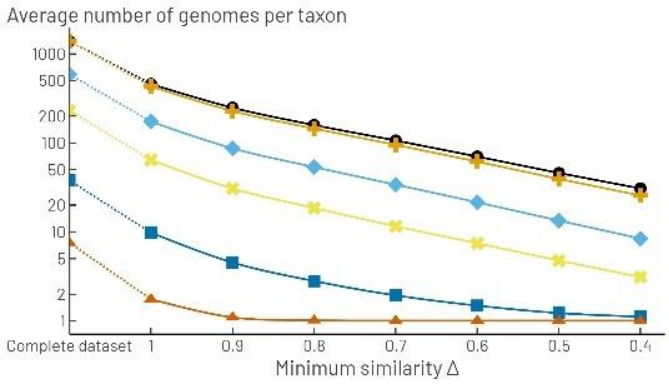
Next, the taxonomic and phylogenetic representativeness of MPS-samples was investigated at lower evolutionary scales, through the analysis of three large and well described families of Gram-positive and Gram-negative bacteria: the *Lactobacillaceae* (Zheng *et al.*, 2020), the *Bacillaceae* (Gupta *et al.*, 2020; Patel and Gupta, 2020), and the *Enterobacteriaceae* (Adeolu *et al.*, 2016), whose taxonomy and phylogeny were recently updated. This local inspection demonstrated that MPS-sampling succeed in adjusting the sampling density according to the internal redundancy of these taxa, homogenizing the taxonomic and phylogenetic diversity when needed (Supplementary Material S39 – Reduction and taxonomic affiliation

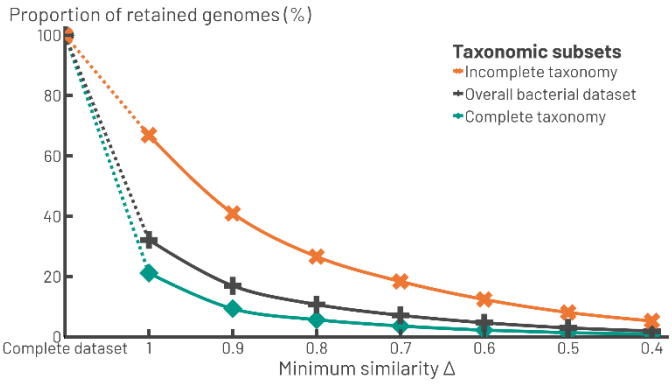
A: Average number of genomes representing each taxonomic level in the samples.

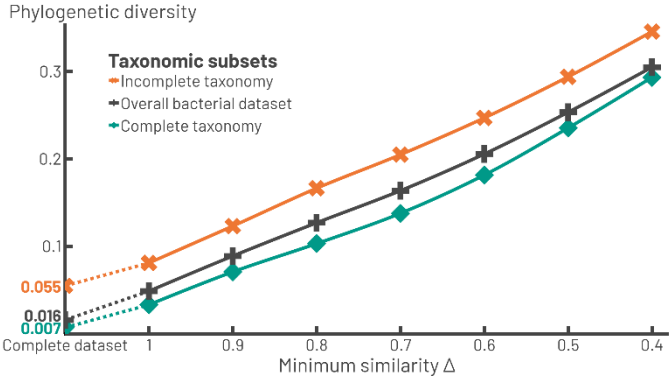
B: Genomic reduction of the 135,315 genomes and the 42,888 genomes with a complete and incomplete taxonomic affiliation, respectively, with a normalized scale.

C: Phylogenetic diversity of three subsets: the overall bacterial dataset, the subset of genomes with complete taxonomic affiliation and the subset of genomes with incomplete taxonomic affiliation. The phylogenetic diversity was computed by the length of all branches divided by the number of leaves.

All precise statistics are available in Supplementary Material S30.







Supplementary Material S40-S42).

Discussion

MPS-Sampling (Multi-Protein Similarity Sampling) is a fast, scalable and reliable method for selecting sets of representative genomes based on the similarity across a set of protein families. In this sense, MPS-Sampling is a Multilocus Sequence Analysis (MLSA) based on similarity (Chun and Rainey, 2014). Through several steps of sequence clustering and matrix calculation, MPS-Sampling overcomes the quadratic complexity, making it possible to process huge genomic datasets in acceptable computation time. The development of MPS-Sampling was inspired by the work of Sørensen (1948), which uses the Dice index and the hierarchical clustering with complete-linkage to study vegetal populations and suggested that: "it may be a good plan first to arrange the material roughly into preliminary groups" (Sørensen, 1948).

MPS-Sampling clustered genomes using two very fast clusterings before applying a third and finer clustering: first very close genomes are dereplicated into reliable EGG, which are then grouped in rough pre-connected components. Finally, these pre-connected components are divided into reliable clusters using complete-linkage. The Dice index is used to calculate the similarity between EGG. Indeed, this index, that reflects the common number of elements between two sets, has interesting properties and is fast to compute. It has been used in many comparative studies (Dalirsefat *et al.*, 2009; McDougal *et al.*, 2003; Chen *et al.*, 2011; Carass *et al.*, 2020). For instance, it was used to compute the percentage of conserved proteins (POCP) between genomes in order to define genus boundaries (Qin *et al.*, 2014). Using ANI could be more accurate at low evolutionary scales, but at larger scales, ANI values become uninformative (Qin *et al.*, 2014).

The use of the pre-connection to preliminary arrange the data allows to reduce considerably the computation time (see below). As a consequence, MPS-Sampling is faster than most clustering approaches based on whole-genome similarity or phylogeny. For example, FastANI (Jain *et al.*, 2018) would take ~153,000 CPU hours and Mash (Ondov *et al.*, 2016) would take ~239 CPU hours, which is respectively 100,000 and 150 times more time than the 1.50 CPU hours for selecting seven samples using MPS-Sampling. Phylogeny-based approaches require the inference of a tree which is extensively time-consuming. For instance, FastTree, a very fast method, took 51.50 CHU hours to infer a rough tree of the bacterial dataset. Then, Treemmer took ~721 CPU hours to select a sample of similar size than MPS-Sampling with $\Delta = 0.4$ (~3,500 representatives) (Menardo *et al.*, 2018). TreeCluster was able to cluster a large tree through complete-linkage in a very short time (less than 10 seconds for 178,203 leaves) (Balaban *et al.*, 2019). However, the quality of the sampling is highly dependent of the quality of the reconstructed tree, which become uncertain when the number and the divergence of sequences increase and/or when rough phylogenetic methods and models are used.

The pre-connection is used to roughly separate genomes that are sufficiently dissimilar, using the information contained in the Lin-combinations. Organizing genomes into pre-connected components reduces the quadratic complexity by computing only the informative subparts of the similarity matrix. MPS-Sampling does not have a quadratic complexity and is therefore much faster than methods requiring the computation of a global similarity matrix. For the bacterial dataset, the construction of EGG reduced the number of 178,203 genomes to 57,332 Lin-combinations, providing a gain of 90% for subsequent pairwise comparisons (Supplementary Material S44). Pre-connection provided another gain of 68%, reaching a total reduction of 97%. Computing the whole similarity matrix would have taken 14h, instead of 26 min for the submatrices, and 2 TB of memory usage instead of 40 GB. This confirmed the significant gain brought by the MPS-Sampling workflow, both in terms of computation time and required memory space. MPS-Sampling has a locally-restricted quadratic complexity and thus is much faster than methods requiring the computation of a global similarity matrix.

By using hierarchical method with complete-linkage, closest EGG or MPS-clusters under construction are aggregated into larger MPS-clusters as long as the similarity between their most dissimilar Lin-combinations is greater than Δ. As a consequence, the intrinsic diversity within each MPS-cluster is controlled, because the Dice index between any pair of Lin-combination is necessarily greater than Δ. For example, when Δ = 0.7, all genomes share at least 70% of their Lin-clusters with their MPS-representative (i.e. here this corresponds to 34 / 48 r-prots). Being highly constrained, complete-linkage cannot create false links but can miss some of them. Thus, MPS-Sampling has a perfect specificity and a very good sensitivity, meaning that there is a slight risk of over-sampling, mostly concerning large groups, which was preferred to a possible loss of representativeness.

As complete-linkage produces bounded clusters, MPS-Sampling is resistant to the chaining effect, which could create elongated clusters (Jain *et al.*, 1999) (Supplementary Material S46-S47). In contrast, Treemmer (Menardo *et al.*, 2018) and ToRQuEMaDA (Léonard *et al.*, 2021) are based on single-linkage and are thus sensitive to this effect. For example, running ToRQuEMaDA on 63,863 genomes, in which *Cyanobacteria* and *Chlamydiae* were the fifth and sixth most represented phyla, 0.67% and 0.56% genomes, respectively, led to the selection of 11 *Chlamydiae* (7.3%) but no *Cyanobacteria* among the 151 sampled genomes. Thus, the sampling led to an enrichment in *Chlamydiae*, despite corresponding strains are very close and nearly indistinguishable on phylogenies inferred with the 151 genomes (Léonard *et al.*, 2021). In contrast, thanks to complete-linkage, MPS-Sampling provided much more balanced samples. For instance, when Δ = 0.4, *Cyanobacteria* was reduced from 1,363 genomes to eight representatives and *Chlamydiae* from 568 genomes to 13 distantly representatives (Supplementary Material S48 and Supplementary Material S49).

A major issue with hierarchical clustering and complete-linkage is that it does not handle outliers (Aguinis *et al.*, 2013). Here, outliers are genomes that cannot be linked to any other genome and thus correspond to singleton MPS-clusters. In our analyses, 21% (Δ = 0.4) to 42% (Δ = 0.7) of the MPS-clusters are singletons, which represents 0.42% to 2.99% of the 178,203 genomes. This may seem high, but it must be kept in mind that it is impossible to determine whether these outliers are artefactual (e.g., genome sequencing or assembly error, contamination) or if they represent real but poorly studied part of the diversity (e.g., new lineages, HGTs). As with any type of unsupervised learning, when the correct answer is not known, it is up to the expert to determine which data is valid.

Here, MPS-Sampling has been applied with r-prot families. These proteins are well suited for evolutionary studies on large evolutionary scales (i.e. from species to phyla) (Yutin *et al.*, 2012; Ramulu *et al.*, 2014). However, at smaller scales of evolution (e.g., intra-species), using faster evolving protein families would be more appropriate.

MPS-Sampling does not explicitly handle duplicated sequences. This was initially considered (Supplementary Material S43), but not implemented in the current version of MPS-Sampling. Multiple copies are rather rare in prokaryotic core protein families, e.g. for the bacterial dataset, duplications concerned only 1.34% of the sequences, and the explored possibilities led more noise than information. Therefore, if the dataset contains paralogs, we recommend users to separate the paralogs into different protein families. This option avoids the loss of information, without bringing noise in the data.

As for all algorithms, the quality of the analyzed data is essential. However, the way MPS-Sampling was designed makes it relatively robust. In fact, the Dice index encapsulates information on a large number of protein families, reducing the impact of missing data (e.g., gene losses, incomplete genome sequencing) or errors, especially concerning protein family assembly (i.e. the inclusion of non-homologous sequences). For instance, although absences are considered when constructing Lin-clusters and Lin-combinations, which allows to correctly discriminate genomes from each other,

they are omitted when calculating the Dice index. As a consequence, missing data do not impact the calculation of the MPS-clusters, thus MPS-Sampling can used with non-core protein families.

Furthermore, if non-homologous sequences are included in protein families, they will form singleton Lin-clusters and will therefore be isolated from other sequences. The main consequence will be the creation of an excessive number of Lin-clusters and EGG. However, if these events are punctual, these EGG will not remain isolated, as they will be gathered with others during pre-connection and MPS-clusters delineation, and due to the centrality criterion, the corresponding genomes have little chance of being retained as MPS-representatives. MPS-Sampling is also relatively resistant to horizontal gene transfer (HGT), as the impact of the non-majority signal from HGTs is minimized by the calculation of the Dice index, allowing genomes to be correctly linked. However, high level of protein family assembly errors or HGT may impact the sampling procedure. For example, if more than 30% of the protein families of genome A used as input belonging to taxon T have been acquired by HGT, MPS-Sampling will be unable to group A with other members of T when $\Delta \geq 0.7$. Most likely, A will be present in the final sample in addition to the representative(s) from T. In that case, this is consistent because according to analyzed data, A represents a significative part of the genetic diversity contained in the dataset.

Conclusion

MPS-Sampling is a new method for selecting samples of representative genomes from a huge database, based on Multi-Proteins Similarity (MPS). Our study shows that MPS-Sampling was particularly performant to reduce a large dataset of bacterial genomes, holding most of the evolutive diversity of the original set, at various evolutionary scales. MPS-Sampling is still consistent when taxonomy is misleading and diverging from phylogeny.

Acknowledgments

The authors wish to thank Pierre S. Garcia and Guy Perrière for their assistance in providing feedback on MPS-Sampling.

Funding information

This work was supported by the National Association of Research and Technology (ANRT - Association Nationale de la Recherche et de la Technologie, France)(grant CIFRE N°2019/1231) and bioMérieux S.A., Marcy l'Étoile, France.

References

Adeolu,M. *et al.* (2016) Genome-based phylogeny and taxonomy of the 'Enterobacteriales': Proposal for enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morgane. *Int. J. Syst. Evol. Microbiol.*, **66**, 5575–5599.

Aguinis,H. *et al.* (2013) Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organ. Res. Methods*, **16**, 270–301.

Almeida,A. *et al.* (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.*, **39**, 105–114.

Balaban,M. *et al.* (2019) TreeCluster: Clustering biological sequences using phylogenetic trees. *PLoS One*, **14**, 1–20.

Bateman,A. *et al.* (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

Bursteinas,B. *et al.* (2016) Minimizing proteome redundancy in the UniProt Knowledgebase. *Database*, **2016**, 1–9.

Carass,A. *et al.* (2020) Evaluating White Matter Lesion Segmentations with Refined Sørensen-Dice Analysis. *Sci. Rep.*, **10**, 1–19.

Chen,C. *et al.* (2011) Representative Proteomes : A Stable, Scalable and Unbiased proteome set for sequence analysis and functional annotation. *PLoS One*, **6**.

Chun,J. *et al.* (2018) Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.*, **68**, 461–466.

Chun,J. and Rainey,F.A. (2014) Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *Int. J. Syst. Evol. Microbiol.*, **64**, 316–324.

Dalirsefat,S.B. *et al.* (2009) Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the silkworm, Bombyx mori. *J. Insect Sci.*, **9**, 1–8.

Dice,L.R. (1945) Measures of the Amount of Ecologic Association Between Species Author (s): Lee R . Dice Published by : Ecological Society of America Stable URL : http://www.jstor.org/stable/1932409. *Ecology*, **26**, 297–302.

Garcia,P.S. *et al.* (2021) A Comprehensive Evolutionary Scenario of Cell Division and Associated Processes in the Firmicutes. *Mol. Biol. Evol.*, **38**, 2396–2412.

Goris,J. *et al.* (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.*, **57**, 81–91.

Gupta,R.S. *et al.* (2020) Robust demarcation of 17 distinct bacillus species clades, proposed as novel bacillaceae genera, by phylogenomics and comparative genomic analyses: Description of robertmurraya kyonggiensis sp. nov. and proposal for an emended genus bacillus limiting it o. *Int. J. Syst. Evol. Microbiol.*, **70**, 5753–5798.

Han,A.X. *et al.* (2019) Phylogenetic clustering by linear integer programming (PhyCLiP). *Mol. Biol. Evol.*, **36**, 1580–1595.

Jain,A.K. *et al.* (1999) Data clustering: A review. *ACM Comput. Surv.*, **31**, 264–323.

Jain,C. *et al.* (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, **9**, 1–8.

Jauffrit,F. *et al.* (2016) RiboDB Database: A Comprehensive Resource for Prokaryotic Systematics. *Mol. Biol. Evol.*, **33**, 2170–2172.

Köster,J. and Rahmann,S. (2012) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

Krishnamoorthy,M. *et al.* (2011) Tree pruner: An efficient tool for selecting data from a biased genetic database. *BMC Bioinformatics*, **12**, 1–8.

Lan,R. and Reeves,P.R. (2002) Escherichia coli in disguise: Molecular origins of Shigella. *Microbes Infect.*, **4**, 1125–1132.

Land,M. *et al.* (2015) Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics*, **15**, 141–161.

Lassmann,T. (2020) Kalign 3: Multiple sequence alignment of large datasets. *Bioinformatics*, **36**, 1928–1929.

Léonard,R.R. *et al.* (2021) ToRQuEMaDA: Tool for retrieving queried Eubacteria, metadata and

dereplicating assemblies. *PeerJ*, **9**, 1–28.
Letunic,I. and Bork,P. (2021) Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.
Li,W. *et al.* (2021) RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.*, **49**, D1020–D1028.
Maayer,P. De *et al.* (2019) Reorganising the order Bacillales through phylogenomics. *Syst. Appl. Microbiol.*, **42**, 178–189.
Maruyama,S. *et al.* (2013) Treetrimmer: A method for phylogenetic dataset size reduction. *BMC Res. Notes*, **6**.
McDougal,L.K. *et al.* (2003) Pulsed-Field Gel Electrophoresis Typing of Oxacillin-Resistant Staphylococcus aureus Isolates from the United States: Establishing a National Database. *J. Clin. Microbiol.*, **41**, 5113–5120.
Menardo,F. *et al.* (2018) Treemmer: A tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics*, **19**, 1–8.
Morlon,H. *et al.* (2016) RPANDA: An R package for macroevolutionary analyses on phylogenetic trees. *Methods Ecol. Evol.*, **7**, 589–597.
Nayfach,S. *et al.* (2021) A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.*, **39**, 499–509.
O'Leary,N.A. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
Olm,M.R. *et al.* (2020) Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems*, **5**, e00731-19.
Ondov,B.D. *et al.* (2016) Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 1–14.
Palmer,M. *et al.* (2020) All anis are not created equal: Implications for prokaryotic species boundaries and integration of anis into polyphasic taxonomy. *Int. J. Syst. Evol. Microbiol.*, **70**, 2937–2948.
Parks,D.H. *et al.* (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.*, **38**, 1079–1086.
Parks,D.H. *et al.* (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.*, **2**, 1533–1542.
Patel,S. and Gupta,R.S. (2020) A phylogenomic and comparative genomic framework for resolving the polyphyly of the genus bacillus: Proposal for six new genera of bacillus species, peribacillus gen. nov., cytobacillus gen. nov., mesobacillus gen. nov., neobacillus gen. nov., metabacillu. *Int. J. Syst. Evol. Microbiol.*, **70**, 406–438.
Philippe,H. *et al.* (2017) Pitfalls in supermatrix phylogenomics. *Eur. J. Taxon.*, **2017**, 1–25.
Pipes,L. and Nielsen,R. (2022) AncestralClust: clustering of divergent nucleotide sequences by ancestral sequence reconstruction using phylogenetic trees. *Bioinformatics*, **38**, 663–670.
Price,M.N. *et al.* (2010) FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**.
Qin,Q.L. *et al.* (2014) A proposed genus boundary for the prokaryotes based on genomic insights. *J. Bacteriol.*, **196**, 2210–2215.
Ragonnet-Cronin,M. *et al.* (2013) Automated analysis of phylogenetic clusters. *BMC Bioinformatics*, **14**.
Ramulu,H.G. *et al.* (2014) Ribosomal proteins: Toward a next generation standard for prokaryotic systematics? *Mol. Phylogenet. Evol.*, **75**, 103–117.
Rosselló-Móra,R. and Amann,R. (2015) Past and future species definitions for Bacteria and Archaea. *Syst. Appl. Microbiol.*, **38**, 209–216.
Sørensen,T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.
Steinegger,M. and Söding,J. (2018) Clustering huge protein sequence sets in linear time. *Nat. Commun.*, **9**.
Stephens,Z.D. *et al.* (2015) Big data: Astronomical or genomical? *PLoS Biol.*, **13**, 1–11.
Varghese,N.J. *et al.* (2015) Microbial species delineation using whole genome sequences. *Nucleic Acids Res.*, **43**, 6761–6771.
Yates,A.D. *et al.* (2022) Ensembl Genomes 2022: An expanding genome resource for non-vertebrates. *Nucleic Acids Res.*, **50**, D996–D1003.
Yutin,N. *et al.* (2012) Phylogenomics of Prokaryotic Ribosomal Proteins. *PLoS One*, **7**, e36972.
Zheng,J. *et al.* (2020) A taxonomic note on the genus Lactobacillus: Description of 23 novel genera,

emended description of the genus Lactobacillus beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. *Int. J. Syst. Evol. Microbiol.*, **70**, 2782–2858.
Zhu,Q. *et al.* (2019) Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.*, **10**.

Figure 1 – Overview of the MPS-Sampling workflow

Input. MPS-Sampling uses families of homologous single-copy protein sequences as input. In this example, ten genomes (g_A, g_B, …, g_J) and four protein families (**uL1**, **uL2**, **uL3**, and **uL4**) are considered. Dashes indicate the absence of a sequence in a genome. Here, **uL2** is missing in g_A, g_B, g_C, g_I and g_J, and **uL4** is missing in g_I.

Step 1: Construction of Lin-clusters. For each protein family, sequence clusters, called **Lin-clusters**, are built using Linclust (Steinegger and Söding, 2018)of the MMseqs2 suite. Because Linclust is applied to each protein family, the clustering, and thus Lin-clusters, may differ from one protein family to another. Here, the **uL1** sequence of g_A is clustered with **uL1** sequences of g_E, g_F and g_G, while **uL4** sequences of g_A and g_D are grouped together.

Step 2-1: Labeling of protein sequences. For each protein family, protein sequences are labeled according to the Lin-cluster to which they belong. Sequences from the largest Lin-clusters are labeled first. For instance, **uL1** sequences from the largest Lin-cluster (g_A, g_E, g_F, and g_G) are labeled as **1**, sequences from the second largest Lin-cluster (g_B, g_C, and g_D) as **2**, while the third largest Lin-cluster (g_H, g_I, and g_J) is labeled as 3.

Step 2-2: Construction of the Lin-clustering matrix. Lin-cluster labels are stored in a matrix, called **Lin-clustering matrix**, whose rows correspond to genomes and columns to protein families. Here, genome g_A (first row) is labeled as **(1, -, 1, 2)**.

Step 2-3: Re-ordering of the Lin-clustering matrix. To insure reproducibility of the sampling, protein families (columns) and genomes (rows) are re-ordered. Protein families are ordered according to their number of Lin-clusters and to the hashing value of their name. Genomes are ordered according to the lexicographic order and to the hashing value of their name. Here, all the four protein families have the same number of Lin-clusters (3), thus are ordered according to their hashing value, giving the new order : **uL3, uL1, uL2, uL4**. Then, the genomes are ordered according first to the lexicographic order, then to the hashing value of their name, given the new order : g_F, g_F, g_G, g_A, g_B, g_C, g_D, g_H, g_I, g_J. Here the ordered of the triplet (g_E, g_F, g_G) and the duet (g_B, g_C) is determined given the hashing value of the genome names.

Step 2-4: Construction of the Lin-combination matrix. Redundant lines of the Lin-matrix are fused, leading to a smaller matrix, called **Lin-combination matrix**, whose rows correspond to unique Lin-combinations of Lin-clusters. In this example, genomes g_B and g_C harbor the same Lin-combination **(1, 2, -, 1)** and are thus bound together in the same Lin-combination (**Comb3**). At this step, the absence of sequences is not taken into consideration; for example, the absence of **uL2** in g_B and g_C is not considered as a difference to separate these two genomes. The genomes grouped together into the same Lin-combination are called an elementary group of genomes (EGG), because they constitute a set of genomes indistinguishable according to the parameters used to the run.

Step 3: Construction of pre-connected components. Lin-combinations are gathered into rough groups called **pre-connected components**. This step, called pre-connection, provides a rough and fast delineation while conserving all pairwise links above a given threshold. In the example, using minNbLinclusters = 2, two pre-connected components are built. A first pre-connected component gathers **Comb1** (g_F, g_E, g_G), **Comb2** (g_A), **Comb3** (g_B, g_C), and **Comb4** (g_D), corresponding to the seven genomes (g_F, g_E, g_G, g_A, g_B, g_C, and g_D), while the second pre-connected component gathers **Comb5** (g_H), **Comb6** (g_I),and **Comb7** (g_J), encompassing the three genomes (g_H, g_I, and g_J) (Supplementary Material S4).

Step 4-1: Computation of similarity submatrices. Within pre-connected components, the similarity between each pair of Lin-combinations is computed and stored in square submatrices, called **similarity submatrices**. The similarity is expressed by the Dice index which corresponds to the proportion of common Lin-clusters between two Lin-combinations (missing values are omitted) (Supplementary Material S5). Here, **Comb6** and **Comb7** share two Lin-clusters out of five, so their Dice index in the matrix is **2/5**.

Step 4-2: Construction of MPS-clusters. The Lin-combinations (and the corresponding genomes) are clustered into **MPS-clusters** according to a hierarchical method with complete-linkage up to a minimum similarity Δ (Supplementary Material S6). In the example, five MPS-clusters are built using minimum similarity Δ = 0.5. A first MPS-cluster gathers two Lin-combinations: **Comb1** (g_F, g_E, g_G) and **Comb2** (g_A), corresponding to four genomes (g_F, g_E, g_G, and g_A). A second MPS-cluster gathers two Lin-combinations: **Comb5** (g_H) and **Comb6** (g_I), corresponding to the two genomes (g_H and g_I). A third MPS-cluster encompasses only one Lin-combination **Comb3** (g_B and g_C), but it corresponds to two genomes (g_B and g_C). The two last genomes corresponding to **Comb4** (g_D) and **Comb7** (g_J) are isolated and correspond to singleton MPS-clusters.

Step 5: Selection of MPS-representatives. One **MPS-representative** genome is selected per MPS-cluster. These MPS-representatives are selected according to centrality, quality, and fame criteria. Here, g_G, g_B, g_D, g_H, and g_J are selected, each representing one MPS-cluster.

Output: MPS-Sampling returns the list of the MPS-representative genomes, as well as the links between each input genome and its MPS-representative genome.

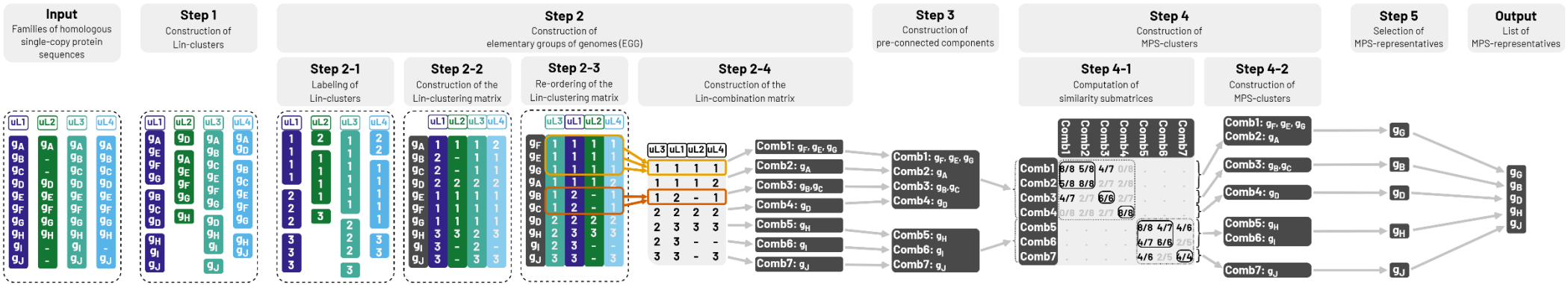


Figure 2 – Sampling of the bacterial dataset (178,203 genomes) by MPS-Sampling.
A: Size of the samples built by MPS-Sampling.
dB: Taxonomic diversity of the samples. The proportion of phyla, classes, orders, families, genera, and species represented in each sample is indicated.
C: Phylogenetic diversity of the samples, computed by the length of all branches of the tree inferred with the sample genomes divided by the number of leaves.
D: Genomic reduction of the 135,315 genomes and the 42,888 genomes with a complete and incomplete taxonomic affiliation, respectively.
All precise statistics and additional figures are available in Supplementary Materials S30 and S38.

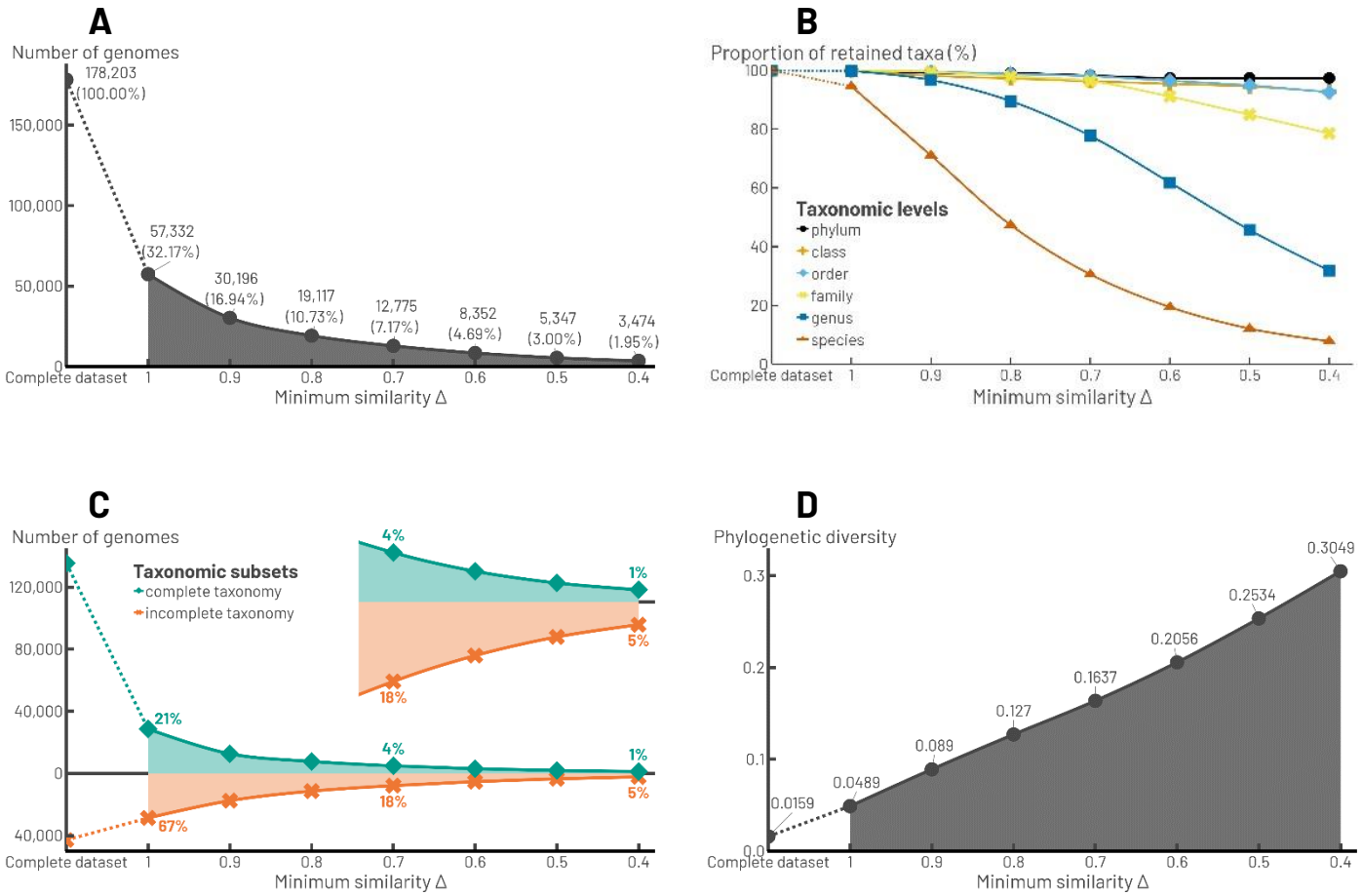
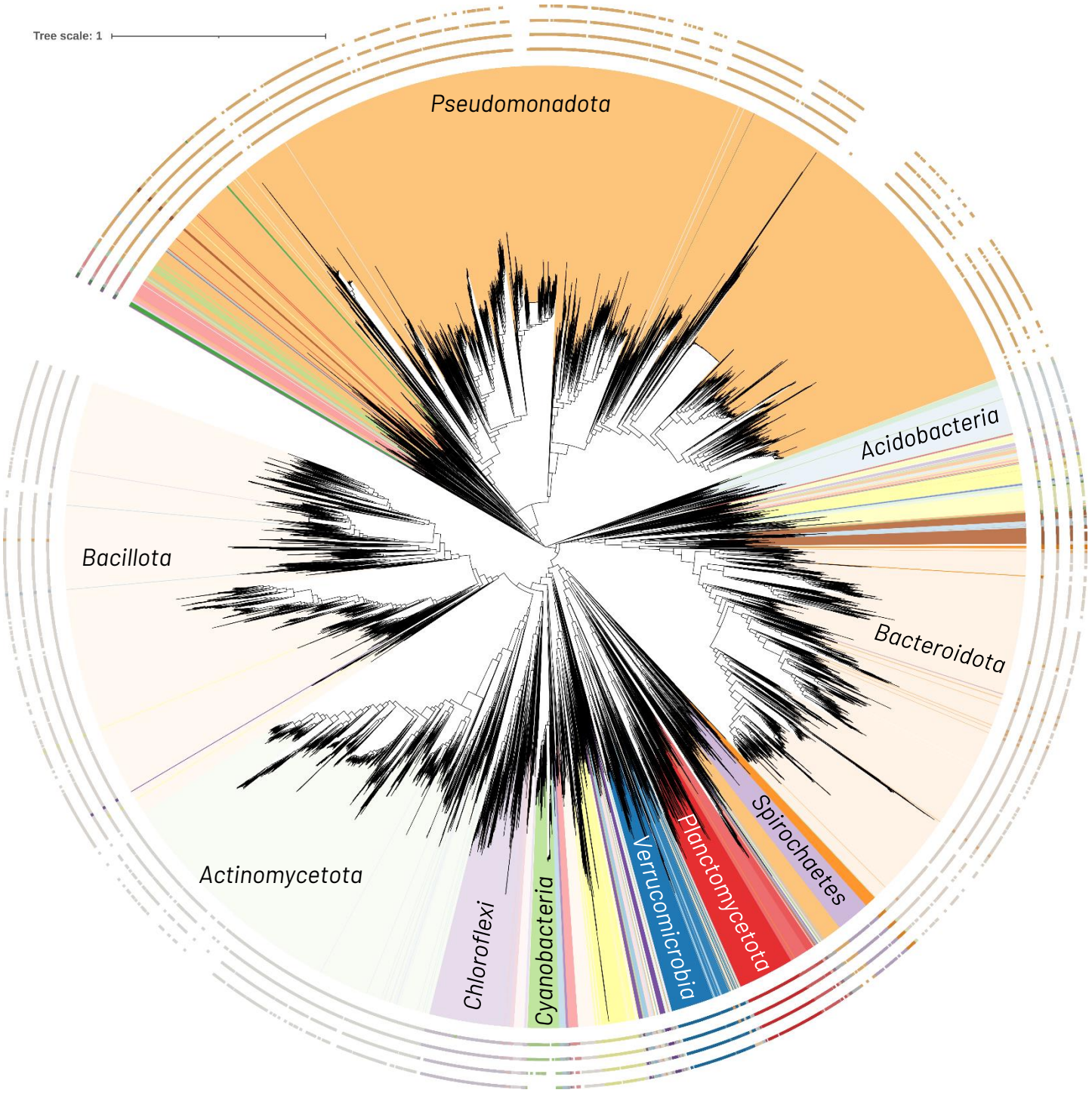


Figure 3 – Phylogenetic distribution of the MPS-representatives on a reference bacterial phylogeny

MPS-representatives corresponding to four MPS-runs are mapped on a reference ML phylogeny of *Bacteria*. From the innermost to the outermost circle, samplings obtained with $\Delta = 0.7$ (12,775 MPS-representatives), $\Delta = 0.6$ (8,352 MPS-representatives), $\Delta = 0.5$ (5,347 MPS-representatives), and $\Delta = 0.4$ (3,474 MPS-representatives). The tree has been inferred using the r-prot sequences from 35,103 genomes. The 10 most represented phyla in the bacterial backbone are: *Pseudomonadota* (ancient *Proteobacteria*; 12,774 leaves in orange at the top), *Bacillota* (ancient *Firmicutes*; 5,768 leaves in light beige at the left), *Bacteroidota* (ancient *Bacteroidetes*; 4,715 leaves in light orange at the right), *Actinomycetota* (ancient *Actinobacteria*; 4,245 leaves in light green at the bottom-left), *Chloroflexi* (1,001 leaves in light purple at the bottom), *Planctomycetota* (ancient *Planctomycetes*; 682 in red at the bottom-right), *Acidobacteria* (615 leaves in light blue at the right slightly at the top), *Verrucomicrobia* (556 leaves in blue at the bottom), *Spirochaetes* (415 leaves in purple at the bottom-right), and *Cyanobacteria* (390 leaves in green at the bottom). The scale bar represents the average number of amino acid substitutions per site in protein sequences used to infer the tree.



Supplementary material

Multi-Proteins Similarity based sampling to select evolutionary significant representative genomes from huge databases

Rémi-Vinh Coudert[1,2,*], Jean-Philippe Charrier[2], Frédéric Jauffrit[2], Jean-Pierre Flandrois[1], Céline Brochier-Armanet[1]

[1] Université de Lyon, Université Lyon 1, CNRS, VetAgro Sup, Laboratoire de Biométrie et Biologie Evolutive, UMR5558, Villeurbanne, France

[2] Microbiology Research & Development, bioMérieux SA, 376 Chemin de l'Orme, 69280 Marcy L'Etoile, France

*To whom correspondence should be addressed.

Table of contents

List of Figures

List of Supplementary Materials

References

Adeolu,M. *et al.* (2016) Genome-based phylogeny and taxonomy of the 'Enterobacteriales': Proposal for enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morgane. *Int. J. Syst. Evol. Microbiol.*, **66**, 5575–5599.

Aguinis,H. *et al.* (2013) Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organ. Res. Methods*, **16**, 270–301.

Almeida,A. *et al.* (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.*, **39**, 105–114.

Balaban,M. *et al.* (2019) TreeCluster: Clustering biological sequences using phylogenetic trees. *PLoS One*, **14**, 1–20.

Bateman,A. *et al.* (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

Bursteinas,B. *et al.* (2016) Minimizing proteome redundancy in the UniProt Knowledgebase. *Database*, **2016**, 1–9.

Carass,A. *et al.* (2020) Evaluating White Matter Lesion Segmentations with Refined Sørensen-Dice Analysis. *Sci. Rep.*, **10**, 1–19.

Chen,C. *et al.* (2011) Representative Proteomes : A Stable, Scalable and Unbiased proteome set for sequence analysis and functional annotation. *PLoS One*, **6**.

Chun,J. *et al.* (2018) Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.*, **68**, 461–466.

Chun,J. and Rainey,F.A. (2014) Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *Int. J. Syst. Evol. Microbiol.*, **64**, 316–324.

Dalirsefat,S.B. *et al.* (2009) Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the silkworm, Bombyx mori. *J. Insect Sci.*, **9**, 1–8.

Dice,L.R. (1945) Measures of the Amount of Ecologic Association Between Species Author (s): Lee R . Dice Published by : Ecological Society of America Stable URL : http://www.jstor.org/stable/1932409. *Ecology*, **26**, 297–302.

Garcia,P.S. *et al.* (2021) A Comprehensive Evolutionary Scenario of Cell Division and Associated Processes in the Firmicutes. *Mol. Biol. Evol.*, **38**, 2396–2412.

Goris,J. *et al.* (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.*, **57**, 81–91.

Gupta,R.S. *et al.* (2020) Robust demarcation of 17 distinct bacillus species clades, proposed as novel bacillaceae genera, by phylogenomics and comparative genomic analyses: Description of robertmurraya kyonggiensis sp. nov. and proposal for an emended genus bacillus limiting it o. *Int. J. Syst. Evol. Microbiol.*, **70**, 5753–5798.

Han,A.X. *et al.* (2019) Phylogenetic clustering by linear integer programming (PhyCLiP). *Mol. Biol. Evol.*, **36**, 1580–1595.

Jain,A.K. *et al.* (1999) Data clustering: A review. *ACM Comput. Surv.*, **31**, 264–323.

Jain,C. *et al.* (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, **9**, 1–8.

Jauffrit,F. *et al.* (2016) RiboDB Database: A Comprehensive Resource for Prokaryotic Systematics. *Mol. Biol. Evol.*, **33**, 2170–2172.

Köster,J. and Rahmann,S. (2012) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

Krishnamoorthy,M. *et al.* (2011) Tree pruner: An efficient tool for selecting data from a biased genetic database. *BMC Bioinformatics*, **12**, 1–8.

Lan,R. and Reeves,P.R. (2002) Escherichia coli in disguise: Molecular origins of Shigella. *Microbes Infect.*, **4**, 1125–1132.

Land,M. *et al.* (2015) Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics*, **15**, 141–161.

Lassmann,T. (2020) Kalign 3: Multiple sequence alignment of large datasets. *Bioinformatics*, **36**, 1928–1929.

Léonard,R.R. *et al.* (2021) ToRQuEMaDA: Tool for retrieving queried Eubacteria, metadata and

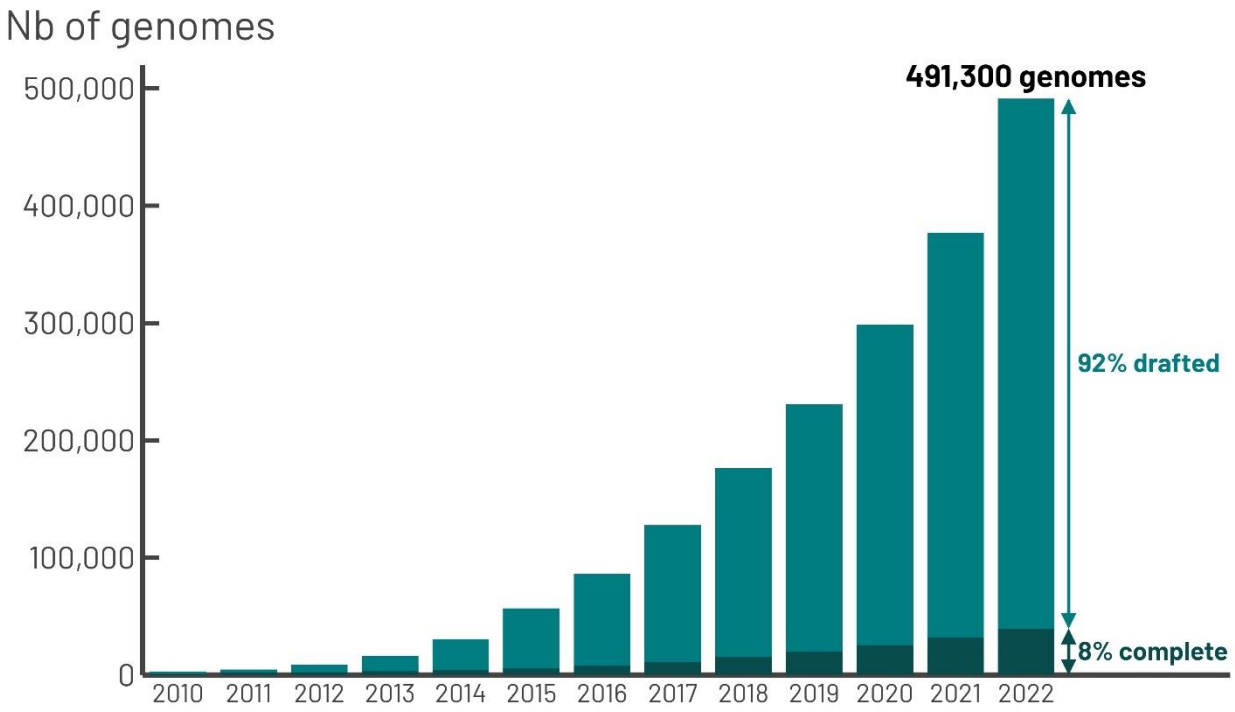
dereplicating assemblies. *PeerJ*, **9**, 1–28.
Letunic,I. and Bork,P. (2021) Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.
Li,W. *et al.* (2021) RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.*, **49**, D1020–D1028.
Maayer,P. De *et al.* (2019) Reorganising the order Bacillales through phylogenomics. *Syst. Appl. Microbiol.*, **42**, 178–189.
Maruyama,S. *et al.* (2013) Treetrimmer: A method for phylogenetic dataset size reduction. *BMC Res. Notes*, **6**.
McDougal,L.K. *et al.* (2003) Pulsed-Field Gel Electrophoresis Typing of Oxacillin-Resistant Staphylococcus aureus Isolates from the United States: Establishing a National Database. *J. Clin. Microbiol.*, **41**, 5113–5120.
Menardo,F. *et al.* (2018) Treemmer: A tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics*, **19**, 1–8.
Morlon,H. *et al.* (2016) RPANDA: An R package for macroevolutionary analyses on phylogenetic trees. *Methods Ecol. Evol.*, **7**, 589–597.
Nayfach,S. *et al.* (2021) A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.*, **39**, 499–509.
O'Leary,N.A. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
Olm,M.R. *et al.* (2020) Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems*, **5**, e00731-19.
Ondov,B.D. *et al.* (2016) Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 1–14.
Palmer,M. *et al.* (2020) All anis are not created equal: Implications for prokaryotic species boundaries and integration of anis into polyphasic taxonomy. *Int. J. Syst. Evol. Microbiol.*, **70**, 2937–2948.
Parks,D.H. *et al.* (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.*, **38**, 1079–1086.
Parks,D.H. *et al.* (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.*, **2**, 1533–1542.
Patel,S. and Gupta,R.S. (2020) A phylogenomic and comparative genomic framework for resolving the polyphyly of the genus bacillus: Proposal for six new genera of bacillus species, peribacillus gen. nov., cytobacillus gen. nov., mesobacillus gen. nov., neobacillus gen. nov., metabacillu. *Int. J. Syst. Evol. Microbiol.*, **70**, 406–438.
Philippe,H. *et al.* (2017) Pitfalls in supermatrix phylogenomics. *Eur. J. Taxon.*, **2017**, 1–25.
Pipes,L. and Nielsen,R. (2022) AncestralClust: clustering of divergent nucleotide sequences by ancestral sequence reconstruction using phylogenetic trees. *Bioinformatics*, **38**, 663–670.
Price,M.N. *et al.* (2010) FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**.
Qin,Q.L. *et al.* (2014) A proposed genus boundary for the prokaryotes based on genomic insights. *J. Bacteriol.*, **196**, 2210–2215.
Ragonnet-Cronin,M. *et al.* (2013) Automated analysis of phylogenetic clusters. *BMC Bioinformatics*, **14**.
Ramulu,H.G. *et al.* (2014) Ribosomal proteins: Toward a next generation standard for prokaryotic systematics? *Mol. Phylogenet. Evol.*, **75**, 103–117.
Rosselló-Móra,R. and Amann,R. (2015) Past and future species definitions for Bacteria and Archaea. *Syst. Appl. Microbiol.*, **38**, 209–216.
Sørensen,T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.
Steinegger,M. and Söding,J. (2018) Clustering huge protein sequence sets in linear time. *Nat. Commun.*, **9**.
Stephens,Z.D. *et al.* (2015) Big data: Astronomical or genomical? *PLoS Biol.*, **13**, 1–11.
Varghese,N.J. *et al.* (2015) Microbial species delineation using whole genome sequences. *Nucleic Acids Res.*, **43**, 6761–6771.
Yates,A.D. *et al.* (2022) Ensembl Genomes 2022: An expanding genome resource for non-vertebrates. *Nucleic Acids Res.*, **50**, D996–D1003.
Yutin,N. *et al.* (2012) Phylogenomics of Prokaryotic Ribosomal Proteins. *PLoS One*, **7**, e36972.
Zheng,J. *et al.* (2020) A taxonomic note on the genus Lactobacillus: Description of 23 novel genera,

emended description of the genus Lactobacillus beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. *Int. J. Syst. Evol. Microbiol.*, **70**, 2782–2858.
Zhu,Q. *et al.* (2019) Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.*, **10**.

Supplementary Material S1 – Growth of available genomic data over time

From the NCBI website, the file ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt was downloaded. It represents the available genomic data of GenBank. All rows until and including 2022 were considered. A cumulated count by year was computed and showed in the bar plot below. Among them, the drafted genomes (i.e. scaffolds and contigs) were colored in light blue at the top of each bar and the complete genomes in dark blue.

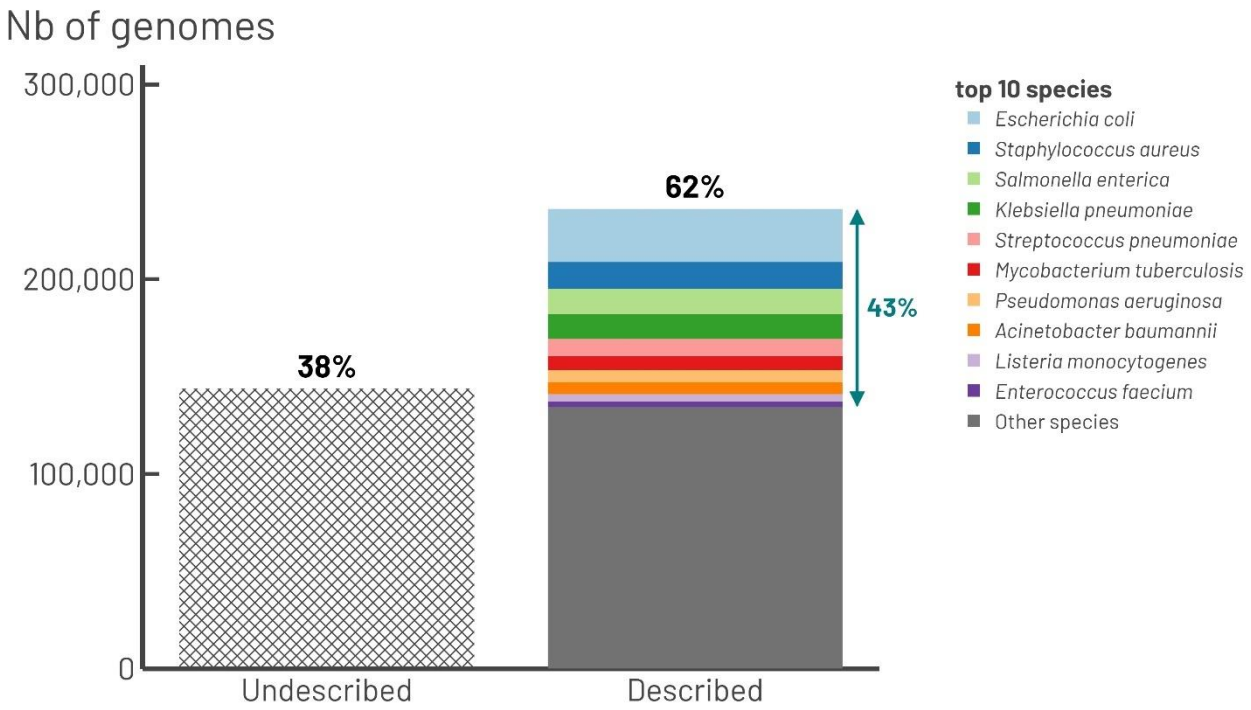
The cumulated amount doubled every 2 or 3 years and reached 491,300 genomes at the end of 2022. At this time, only 8% of the data were complete genomes while the remaining 92% were only drafted genomes.



Supplementary Material S2 – Unbalanced representativity in available genomic data

From the NCBI website, the file ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt was downloaded. It represents the available genomic data of GenBank. All rows until and including were considered. Genomes with an incorrect species name, e.g. ending with sp. or bacterium, were classified as "undescribed". Remaining genomes were classified as "described". Among them, the genomes of the 10 most represented species were highlighted, as mentioned in the legend (*Escherichia coli*, *Staphylococcus aureus*…).

Among the genomes available in 2022, 38% of them are undescribed while 62% of them are correctly described. Moreover, 43% of these correctly described genomes represent only 10 species, the most sequenced ones.



Supplementary Material S3 – Recognized libraries of representative genomes

Several consortia have made libraries of representative genomes available to users. NCBI (O'Leary *et al.*, 2016), Uniprot (Bateman *et al.*, 2017) and Ensembl Genomes (Yates *et al.*, 2022) propose representative genomic data, which contain, as of 2023/01/27, 17,145 RefSeq-representative genomes (including 16,557 for *Bacteria*), 22,121 UniProt reference proteomes (including 8,821 for *Bacteria*), and 33,316 Ensembl-representative genomes (including 31,332 for *Bacteria*) respectively. RefSeq-representative genomes (O'Leary *et al.*, 2016) are " computationally or manually selected as a representative from among the best genomes available for a species or clade". They are chosen on "chosen among eligible assemblies based on [some] criteria", the first criterium being the "manual selection". The UniProt reference proteomes (Bateman *et al.*, 2017) are "selected among all proteomes (manually and algorithmically, according to a number of criteria) to provide broad coverage of the tree of life". Ensembl-representative genomes (Yates *et al.*, 2022) are chosen among UniProt according to automatic dereplication rules (Bursteinas *et al.*, 2016). These three references genomic sets include proteomes of interest for biomedical and biotechnological research, and are thus highly impacted by socio-induced biases. These libraries bring together a qualitative and supposedly representative sampling of the taxonomic diversity of genomic data. The advantage for the users is that they do not have to manage the data sampling step. However, the use of these libraries also has limitations. First, the users have no control over the selection of genomes, so some taxonomic groups of interest may not be represented. Second, the sampling density and the redundancy of the data is not controlled, which may require a second sampling step. Finally, these libraries do not include, most of the time, the genomes of undescribed organisms.

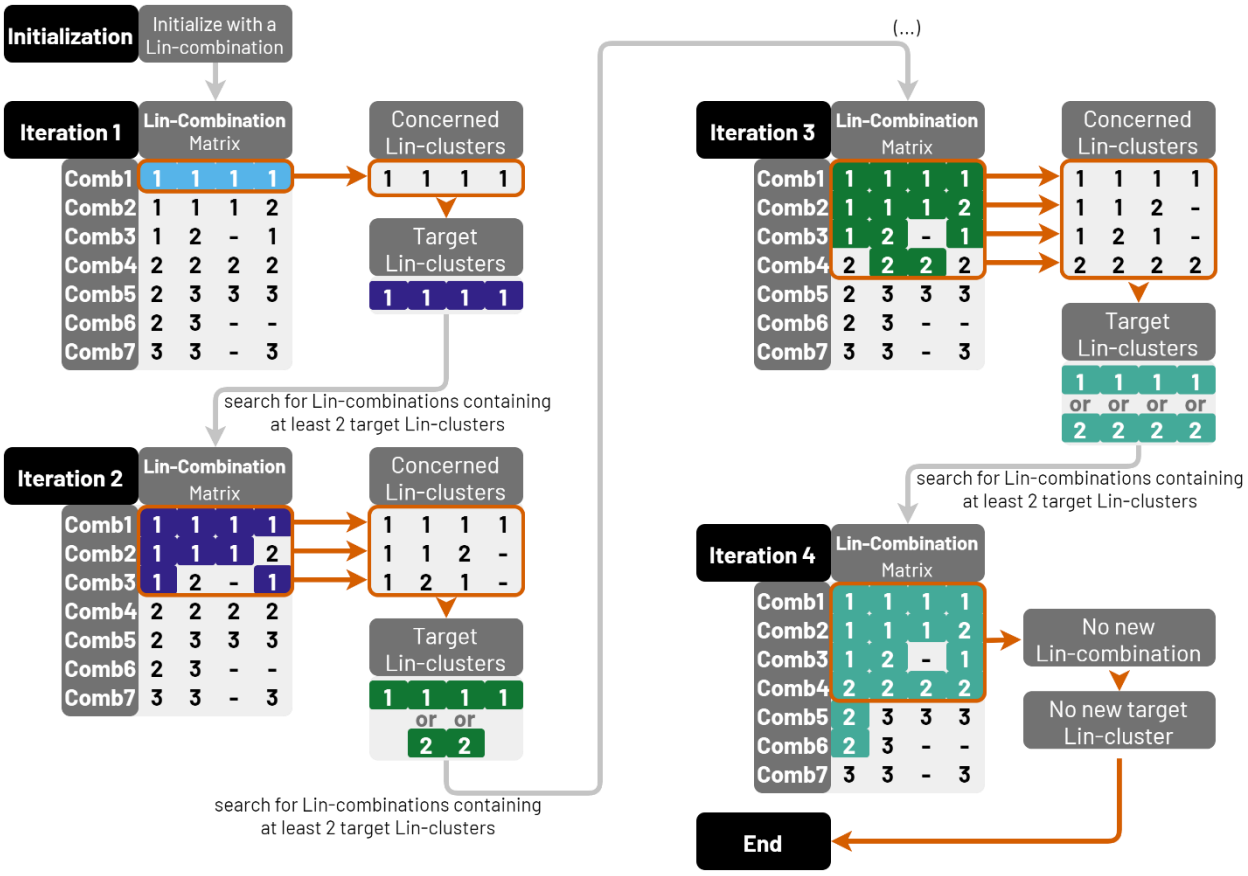
These encompass both RefSeq representative sensu stricto and RefSeq reference according to the NCBI. RefSeq reference genomes are "manually selected high quality genome assembly that NCBI and the community have identified as being important as a standard against which other data are compared", while RefSeq representative genomes are "computationally or manually selected as a representative from among the best genomes available for a species or clade that does not have a designated reference genome" (Li *et al.*, 2021). To simplify, there are both referred to as RefSeq representatives.

Supplementary Material S4 – From Lin-combinations to pre-connection

Here an example of the construction of a pre-connected component during the pre-connection, with the parameter minNbLinclusters = 2. From a starting Lin-combination, close Lin-combinations are iteratively absorbed according to a single common reference of target Lin-clusters.

Initialization: The pre-connected component is initialized with a given Lin-combination: **Comb1**.
Iteration 1: Only the Lin-clusters of the starting Lin-combination are targeted: **(1111)**.
Iteration 2: Lin-combinations sharing at least minNbLinclusters = 2 targeted Lin-clusters are searched. It results that two Lin-combinations have at least 2 target Lin-clusters: **Comb2** and **Comb3**. These two Lin-combinations are added to the current pre-connected component. This is the aggregative aspect of pre-connection. For the next iteration, the Lin-clusters of the 3 involved Lin-combinations (**Comb1**, **Comb2**, and **Comb3**) are gathered into a single common reference of target references.
Iteration 3: Searching Lin-combination with at least minNbLinclusters = 2 target Lin-clusters, a new Lin-combination matches: **Comb4**. This Lin-combination is added and the current pre-connected component contains now 4 Lin-combinations: **Comb1**, **Comb2**, **Comb3**, and **Comb4**.
Iteration 4: No new Lin-combination is found. Thus the pre-connected component is now stable and its delineation ends.

To resume, from the Lin-combination **Comb1**, a pre-connected component has been delineated using minNbLinclusters = 2 through 4 iterations and contains four Lin-combinations: **Comb1**, **Comb2**, **Comb3**, and **Comb4**. The process could be used to build a second pre-connected component. Starting from the Lin-combination **Comb5**, it ends containing three Lin-combinations: **Comb5**, **Comb6**, and **Comb7**.

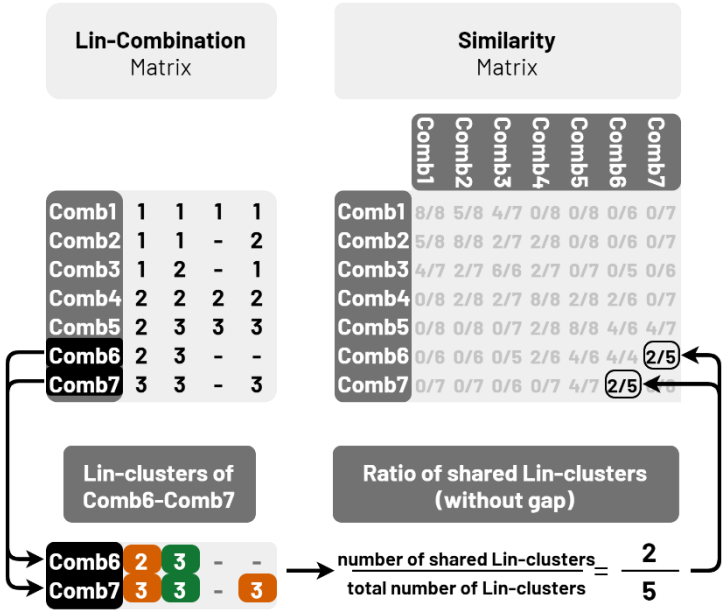


Supplementary Material S5 – Dice index and similarity submatrices

The computation of the Dice index between the pair of the two Lin-combinations (Comb6 – Comb7) will be detailed.
To begin with, consider the Lin-clusters of the two Lin-combinations: **(2, 3, –, -)** and **(3, 3, –, 3)**.
Then count the common Lin-clusters and the total number of Lin-clusters (without missing values), respectively **2** and **5**.
At the end, compute the proportion of common Lin-clusters, which is **2 / 5**.

Noting X and Y the two compared Lin-combinations, the Dice index is given by the formula: $\frac{2|\mathbf{X}\cap\mathbf{Y}|}{|\mathbf{X}|+|\mathbf{Y}|}$.

The Dice index handles missing values. It completely ignores the missing values when the protein family is missing in both Lin-combinations (protein family 3, third column, – vs –) and balances when there is only one missing value against a Lin-cluster (protein family 4, fourth column, - vs 3).



Supplementary Material S6 – From pre-connection to MPS-clustering

1: The pre-connected components indicate which parts to calculate in the sparse similarity matrix. In this example, two pre-connected components have been previously computed (Supplementary Material S4). The first one with **Comb1** to **Comb4** and the second one with **Comb5** to **Comb7**. As a result, only two similarity submatrices are computed (orange dotted lines).

2: Among computed similarities, only high similarities are used. In this example, the minimum similarity Δ was set to Δ=0.5 ; thus all similarities greater or equal to 0.5 are used. All similarities greater than Δ=0.5 are highlighted in black in the similarity matrix and are marked as links in the graph theory representation. In this example, four links have been found: **Comb1**-**Comb2**, **Comb2**-**Comb3**, **Comb5**-**Comb6** and **Comb5**-**Comb7**.

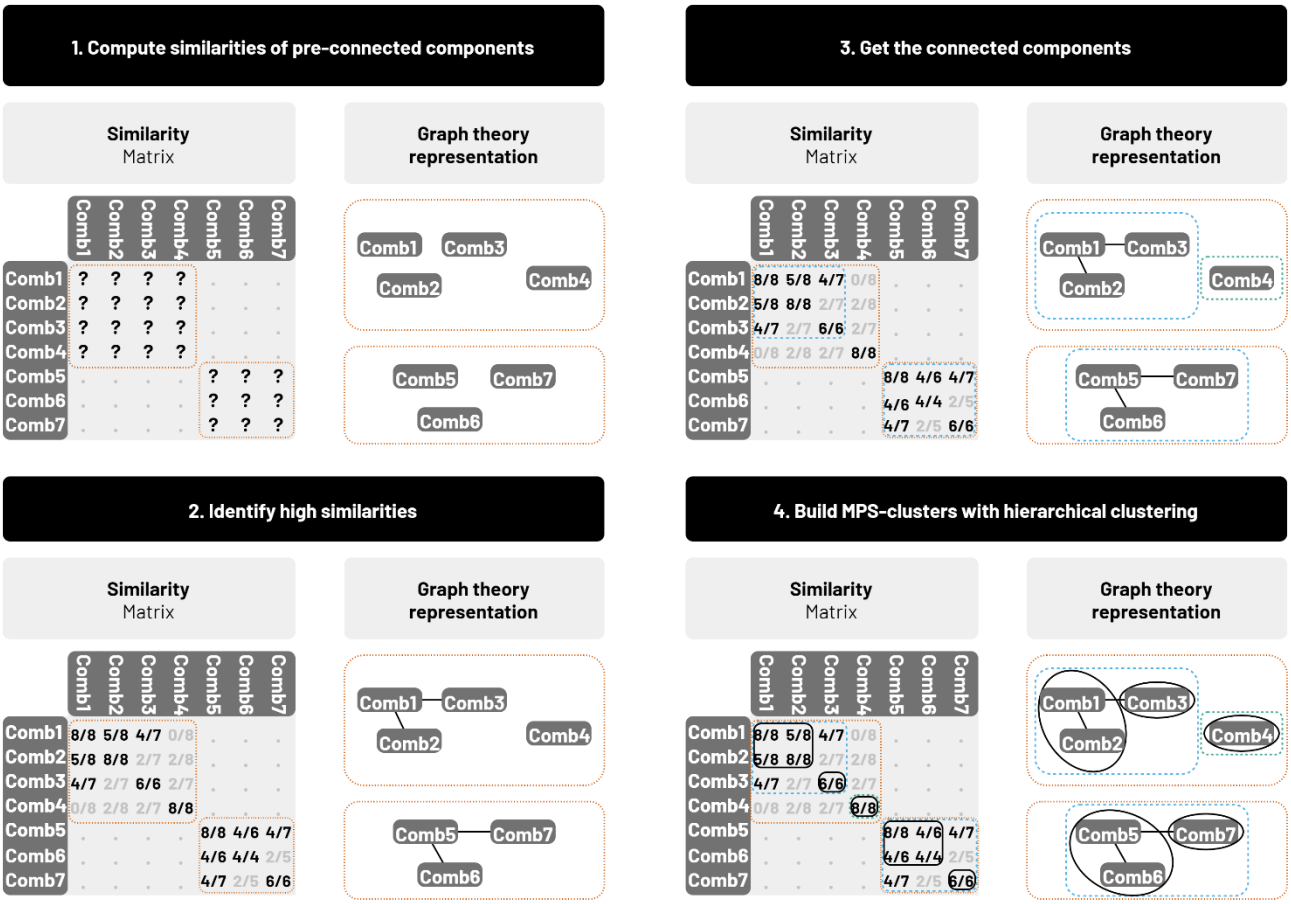
3: Connected components are built using all the found links above the minimum similarity Δ=0.5. In this example, 3 connected components are built (blue dashed lines): {**Comb1**, **Comb2**, **Comb3**}, {**Comb4**} and {**Comb5**, **Comb6**, **Comb7**}.

4: MPS-clusters are built using hierarchical method with complete-linkage. It leads to build 5 MPS-clusters (black lines). 2 MPS-clusters gather 2 Lin-combinations, respectively {**Comb1**, **Comb2**} and {**Comb5**, **Comb6**}. 3 Lin-combinations remain isolated into singleton MPS-clusters, respectively {**Comb3**}, {**Comb4**}, and {**Comb7**}.

Note that the pre-connection avoids to compute empty parts of the similarity matrix, saving a substantial amount of time. In this example, only 4x4 + 3x3 = 16 + 9 = 25 similarities have been computed among the total of 7x7 = 49 similarities. As a result, the similarity matrix computation was reduced to 50%.

The pre-connection, the connection and the complete-linkage are three increasingly fine partitions. In this example, it is visible that each MPS-cluster is completely included in a connected component. Then each connected component is itself completely included in a pre-connected component.

This degressive partitioning is a suitable divide-and-conquer: it splits the complete dataset into preliminary subparts, breaking the quadratic complexity, without discarding any important link. To find interesting links, it has a perfect sensibility, but a medium specificity. (Considering the founded links, there are many false positives, but no false negative.)



Supplementary Material S7 – Priority rules for the selection of the representative genomes

For each MPS-cluster, the MPS-representative genome is selected according to a rigorous hierarchical process, which favors:

(1) Genomes with the best fame degree. For example, a RefSeq-representative* genome is always preferred **(R)**, if available, then a complete assembly genome from a type strain **(TC)**, etc.
(2) Genomes with the largest protein coverage across families used as input.
(3) Genomes with the best centrality within the MPS-cluster, i.e. belonging to the largest Lin-clusters within the MPS-cluster.
(4) Genome whose name has the smallest hash value. The hash values are computed with the function hash() of the R package rlang.

The frequency of the priority rules is showed in the bar plot of the Supplementary Material S29.

	Step	Description	12,775 MPS-clusters (Δ=0.7)	12,775 MPS-clusters (Δ=0.7)	178,203 genomes (complete dataset)
1	Fame	Genome with the best fame degree	7,446 (58%)		
	(R)	RefSeq-representative*		4,909 (38%)	16,135 (9%)
	(TC)	Complete assembly from type strain		6 (0%)	1,144 (1%)
	(TS)	Scaffold assembly from type strain		9 (0%)	2,010 (1%)
	(EC)	Complete assembly from Ensembl! Bacteria (Yates *et al.*, 2022)		64 (1%)	2,730 (2%)
	(ES)	Scaffold assembly from Ensembl! Bacteria(Yates *et al.*, 2022)		490 (4%)	5,368 (3%)
	(C)	Complete assembly		300 (2%)	25,230 (14%)
	(S)	Scaffold assembly		2,567 (20%)	46,784 (26%)
	(U)	Unassembled		214 (2%)	65,319 (37%)
	(Others)	-		4,216 (33%)	13,483 (8%)
2	Protein coverage	Genome with the most sequences as possible	4,833 (38%)		
3	Centrality	Genome with the most frequent Lin-clusters within the MPS-cluster	2,603 (20%)		
4	Pseudo-Random	Genome whose name has the smallest Hash value.	817 (6%)		

* These encompass both RefSeq-representative *sensu stricto* and RefSeq-reference according to the NCBI. RefSeq-reference genomes are "manually selected high-quality genome assembly that NCBI and the community have identified as being important as a standard against which other data are compared", while RefSeq-representative genomes are "computationally or manually selected as a representative from among the best genomes available for a species or clade that does not have a designated reference genome" (O'Leary *et al.*, 2016). To simplify, there are both referred to as RefSeq-representatives.

Supplementary Material S8 – Centrality criterium

Proof of the **Corollary 1:** The most central genomes can be chosen according to the frequency of the Lin-clusters. The genomes whose the frequency of their Lin-clusters will be the most central, according to the Dice index.

Introduction

Let's consider M genomes and N protein families. For each genome i and each protein family j, a sequence s_j^i belongs to a Lin-cluster x_j^i. Let's see an example with $M = 4$ genomes (g_A, g_B, g_C, g_D) and $N = 3$ protein families (f_1, f_2, f_3). It gives a matrix (x_j^i) of 4 rows and 3 columns.

	f_1	f_2	f_3
g_A	1	1	1
g_B	1	1	1
g_C	1	1	2
g_D	2	1	3

Lin-combination matrix

For example, the sequence s_3^D comes from the genome g_D, is attached to the protein family f_3 and was put in the Lin-cluster x_3^D which is labeled 3, i.e. $x_3^D = 3$.

Frequency of Lin-clusters

Now, for each Lin-cluster x_j^I can be calculated its frequency $freq(x_j^I)$. Noting the Dirac function δ, this frequency can be given by :

$$freq(x_j^I) = \frac{1}{M} \times \sum_{i=1}^{M} \delta(x_j^I, x_j^i)$$

Then an average frequency can be associated to each genome g_I by calculated the average of the frequency of its Lin-clusters. This average frequency can be given by :

$$freq(g_I) = \frac{1}{N} \times \sum_{j=1}^{N} freq(x_j^I) = \frac{1}{NM} \times \sum_{j=1}^{N} \sum_{i=1}^{M} \delta(x_j^I, x_j^i)$$

Let's see some applications from the previous example.

	f_1	f_2	f_3	$freq$
g_A	3/4	4/4	2/4	9/12
g_B	3/4	4/4	2/4	9/12
g_C	3/4	4/4	1/4	8/12
g_D	1/4	4/4	1/4	6/12

Average frequency of Lin-clusters

For example, the Lin-cluster $x_3^D = 3$ appears once out of four genomes in the family f_3, so its frequency is equal to $freq(x_3^D) = 1/4$. Another example is the Lin-cluster $x_1^A = 1$. It appears in three genomes out of four or in the family f_1, so its frequency is $freq(x_1^A) = 3/4$. Last, the genome g_A has three Lin-clusters with the respective frequencies $3/4$, $4/4$ and $2/4$, so the average frequency of this genome g_A is:

$$freq(g_A) = \frac{1}{3} \times (\frac{3}{4} + \frac{4}{4} + \frac{2}{4}) = \frac{9}{12}$$

Dice index and centrality

Now, let's introduce the Dice index between the Lin-clusters of two genomes. In this context, the Dice index counts the proportion of common Lin-clusters between two genomes. The Dice index is given by the formula:

$$Dice(g_A, g_B) = \frac{1}{2\mathrm{N}} \times \sum_{j=1}^{N} 2\delta(x_j^A, x_j^B) = \frac{1}{\mathrm{N}} \times \sum_{j=1}^{N} \delta(x_j^A, x_j^B)$$

Let's see what happens for the previous example.

	g_A	g_B	g_C	g_D	$centrality$
g_A	6/6	6/6	4/6	2/6	18/24
g_B	6/6	6/6	4/6	2/6	18/24
g_C	4/6	4/6	6/6	2/6	16/24
g_D	2/6	2/6	2/6	6/6	12/24

Dice index matrix

For example, the two genomes g_A and g_B share identical Lin-clusters for the three protein families : $(1, 1, 1)$ so their Dice index is $Dice(g_A, g_B) = 6/6$. On the contrary, the two genomes g_C and g_D share only the same Lin-cluster for the second protein family f_2, so their Dice index is: $Dice(g_A, g_B) = 2/6$.

Next, the centrality of a genome can be defined by its average Dice index to all the genomes (including itself). It can be given by the formula:

$$centrality(g_A) = \frac{1}{\mathrm{M}} \times \sum_{i=1}^{M} Dice(g_A, g_i) = \frac{1}{\mathrm{MN}} \times \sum_{i=1}^{M} \sum_{j=1}^{N} \delta(x_j^A, x_j^i)$$

The following lemma will link the centrality of a genome and the frequency of its Lin-clusters.

Lemma 1: The centrality of a genome is equals to the average frequency of its Lin-clusters, i.e.:

$$centrality(g_A) = freq(g_A)$$

Prof:

$$centrality(g_A) = \frac{1}{\mathrm{M}} \times \sum_{i=1}^{M} Dice(g_A, g_i) = \frac{1}{\mathrm{MN}} \times \sum_{i=1}^{M} \sum_{j=1}^{N} \delta(x_j^A, x_j^i)$$

The two sums can be switched because the sum is finite.

$$centrality(g_A) = \frac{1}{\mathrm{NM}} \times \sum_{j=1}^{N} \sum_{i=1}^{M} \delta(x_j^A, x_j^i) = \frac{1}{N} \times \sum_{j=1}^{N} freq(x_j^A) = freq(g_A)$$

□

Then the following theorem will be easily proved.

Theorem 1: The most central genomes are the genomes whose the average frequency of their Lin-clusters are the higher.

Prof: It is simply because the centrality of a genome is equal to the average frequency of its Lin-clusters. So the genome whose Lin-clusters have the highest average frequency will also be the genome with the higher centrality.

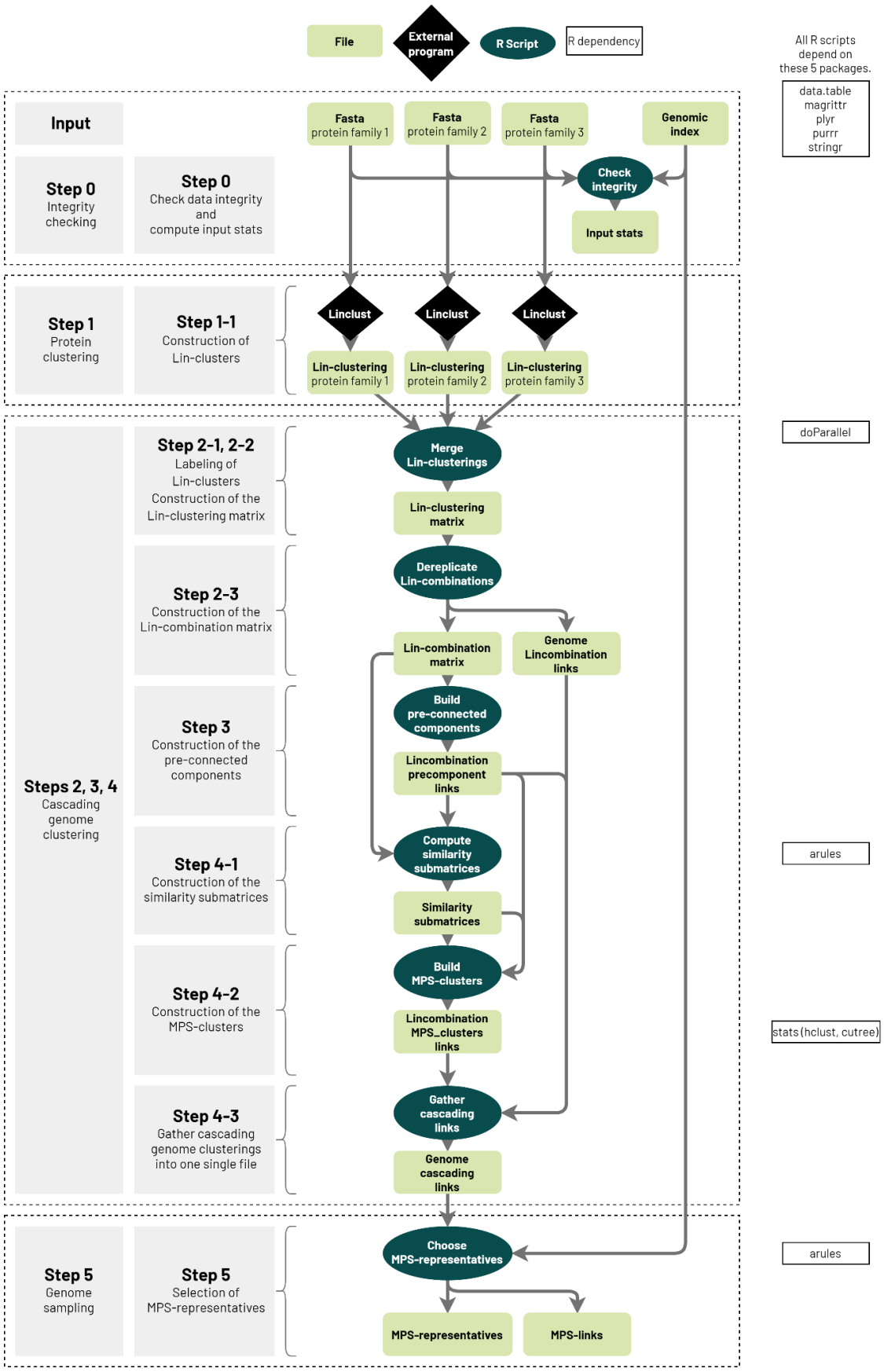
□

Corollary 1: The most central genomes can be chosen according to the frequency of the Lin-clusters. The genomes whose the frequency of their Lin-clusters will be the most central, according to the Dice index.

In the previous example, the most central genomes will be g_A and g_B because the average frequency of their Lin-clusters and their centrality will be the higher, i.e. $\frac{9}{12} = \frac{18}{24}$.

Supplementary Material S9 – Flowchart of MPS-Sampling

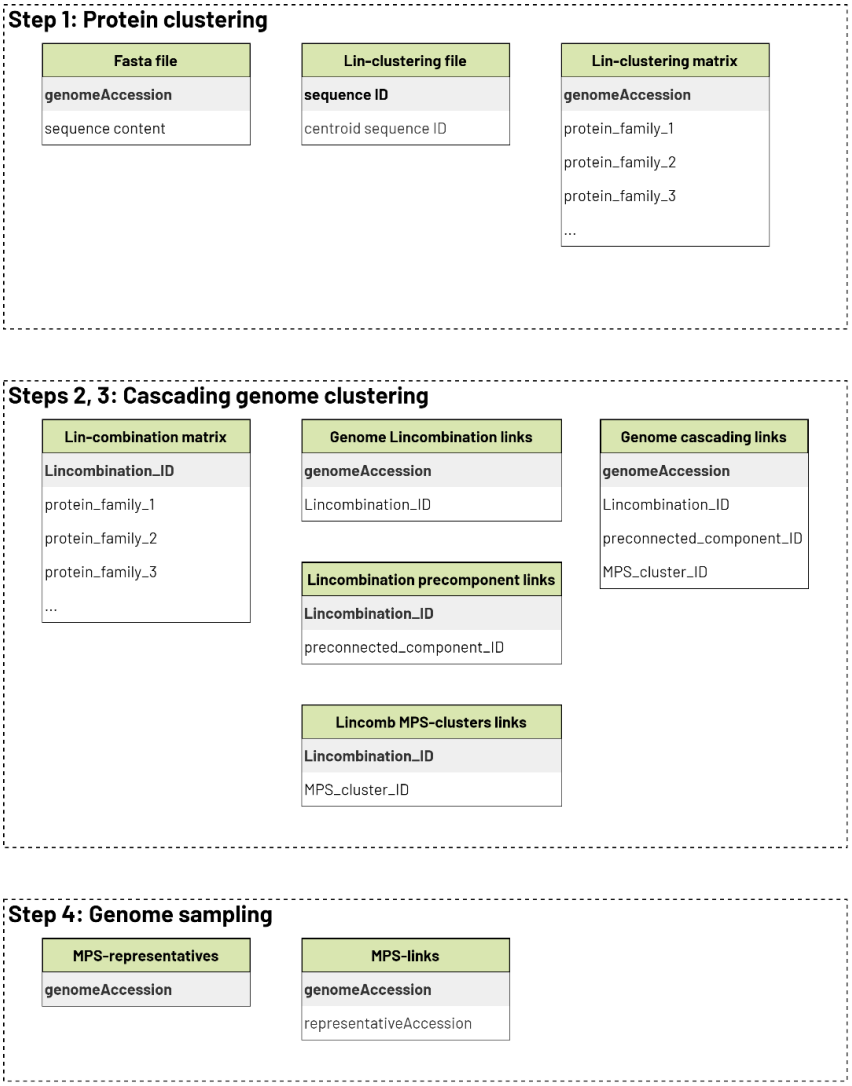
This flowchart shows the data analysis processed by MPS-Sampling, centralized thanks to a Snakemake pipeline. Each file is colored in light green. The only external program is Linclust and colored in black. All other tasks are carried out by home-made R scripts, colored in strong blue. The R dependencies is mentioned at the right, in white boxes.



Supplementary Material S10 – Entity relationship diagram (ERD) of MPS-Sampling

The entity relationship diagram (ERD) of MPS-Sampling describes the different data tables used during the analysis. The primary key of each table is highlighted in grey.

For example, the data table "Genome cascading links" contains a row per genome. For each row, the primary key **genomeAccession** indicates the genome, then **Lincombination_ID**, **preconnected_component_ID** and **MPS_cluster_ID** indicate to which Lin-combination, pre-connected component and MPS-cluster the genome is attached, respectively.



Supplementary Material S11 – Trimming of the bacterial dataset

The initial data encompassed 200,565 genomes along 55 protein families, with 9,693,984 single-copy sequences representing 87.88% of the cells (see Supplementary Material S12). First, only protein families present in more than 50% of the genomes were kept, leading to the elimination of 55 – 53 = 2 protein families. Then, bacterial genomes containing more than 80% of the 53 remaining protein families were kept, leading to the elimination of 200,565 - 178,693 = 21,872 genomes. Next, protein families were selected again, keeping only those present in more than 80% of the genomes. And again with the genomes present in more than 80% of the 50 remaining protein families. Now, protein sequences with abnormally short size (i.e. < 75% of the median length within protein families) were excluded from the analysis. This leads to the exclusion of 8,546,835 - 8,334,499 = 212,336 sequences. After two last trimming on the protein families and the genomes, the final trimming of the bacterial dataset encompassed 178,203 genomes along 48 protein families, with 8,315,939 single-copy sequences representing 97,22% of the cells.

Supplementary Material S12 – Content dimensions during the trimming of the bacterial dataset

Each line of this table represents the dimensions of the genomic data at a given stage. For example, the initial data encompassed 200,565 genomes (rows) along 55 protein families (columns), with 9,693,984 single-copy sequences (filled cells) representing 87.88% of the cells.

Step	Number of rows	Number of columns	Number of filled cells	Ratio of filled cells
Initial data	200,565	55	9,693,984	87.88%
Column ≥ 50%	200,565	53	9,685,194	91.11%
Row ≥ 80%	178,693	53	9,029,460	95.34%
Column ≥ 80%	178,693	50	8,633,315	96.63%
Row ≥ 80%	178,693	50	8,633,315	96.63%
Column: Sequence Length ≥ 75% of Median	178,693	50	8,546,835	95.66%
Column ≥ 80%	178,693	48	8,334,499	97.17%
Row ≥ 80%	178,203	48	8,315,939	97.22%

Supplementary Material S13 - Duplication cases in the bacterial dataset

The bacterial dataset encompassed 178,203 x 48 = 8,553,744 cases (see Supplementary Material S12). 179,103 of these cases (2.09%) were empty, i.e. there were 8,553,744 – 179,103 = 8,374,641 filled cases (97.91%) encompassing 8,436,399 sequences. 51,026 of the 8,374,641 filled cases (0.61%) are duplication cases. 112,784 of the 8,436,399 sequences (1.34%) were involved in duplication cases. Only single-copy sequences were considered, meaning that 112,784 sequences (1.34% of the sequences) were excluded of the analysis. At the end, this represented 8,436,399 - 112,784 = 8,323,615 protein sequences.

Supplementary Material S14 - Optimization of the parameters of MPS-Sampling

For step 1, Linclust parameters were set as follow: eValue = 10^{-5}, coverageMode = 0, minCov = 0.8, and minSeqID = 0.6. This corresponded to the code for Linclust: "-e 0.00001 -cov-mode-0 -c 0.8 –min-seq-id 0.6". The coverage mode 0 corresponded to a bidirectional coverage between query and target. The step involving Linclust lasted 1 minute without parallelization (Supplementary Material S22) and could be automatically parallelized with the Snakemake pipeline if needed. Thus, it was possible to test various sets of parameters. Here, the Linclust parameters were optimized to generate a number of Lin-clusters between 2% and 10% of the bacterial dataset, which corresponds to 3,000 and 17,000 genomes, respectively. Regarding our tests, only the minimum sequence identity seems to impact the number of Lin-clusters (Supplementary Material S15-S18). The minimum sequence identity of Linclust was set to minSeqID = 0.6 because it brought a sample of 12,775 and 3,474 MPS-representatives for $\Delta = 0.7$ and $\Delta = 0.4$, respectively (Figure 2A). This value of minSeqID = 0.6 brought a median of 773 Lin-clusters along the 48 proteins (Supplementary Material S24), which represent roughly the number of taxonomic families of 661. Based on these parameters, within each Lin-cluster, protein sequences were expected to share at least half of their amino acid positions with the centroid sequences (minSeqID x minCov = 0.60 x 0.80 = 0.48).

For step 2, there is no parameter to adjust the construction of the EGG. Intrinsically, this step corresponds to a MPS-clustering of the step 4 with $\Delta = 1$.

For step 3, the pre-Connection was very fast (2 minutes) (Supplementary Material S22). The optimization of minNbLinclusters parameter was based on two indicators: the size of the largest pre-connected component and the number of pre-connected components. In fact, when the size of the largest pre-connected component becomes too large, the quadratic complexity makes the calculation of its similarity matrix too tedious. The first requirement concerns the largest pre-connected component and its size (Supplementary Material S19). Because of quadratic complexity, when its size exceeds the critical threshold of ~25k, computing its similarity submatrix lasts too long regarding the machine we used. The second requirement concerns the minimum similarity Δ (Supplementary Material S20). As there is at least a MPS-cluster per pre-connected component, there are at least as many MPS-representative genomes as there are pre-connected components. It leads to choosing the smallest possible minNbLinclusters with the size of the largest pre-Connected component below the critical area.

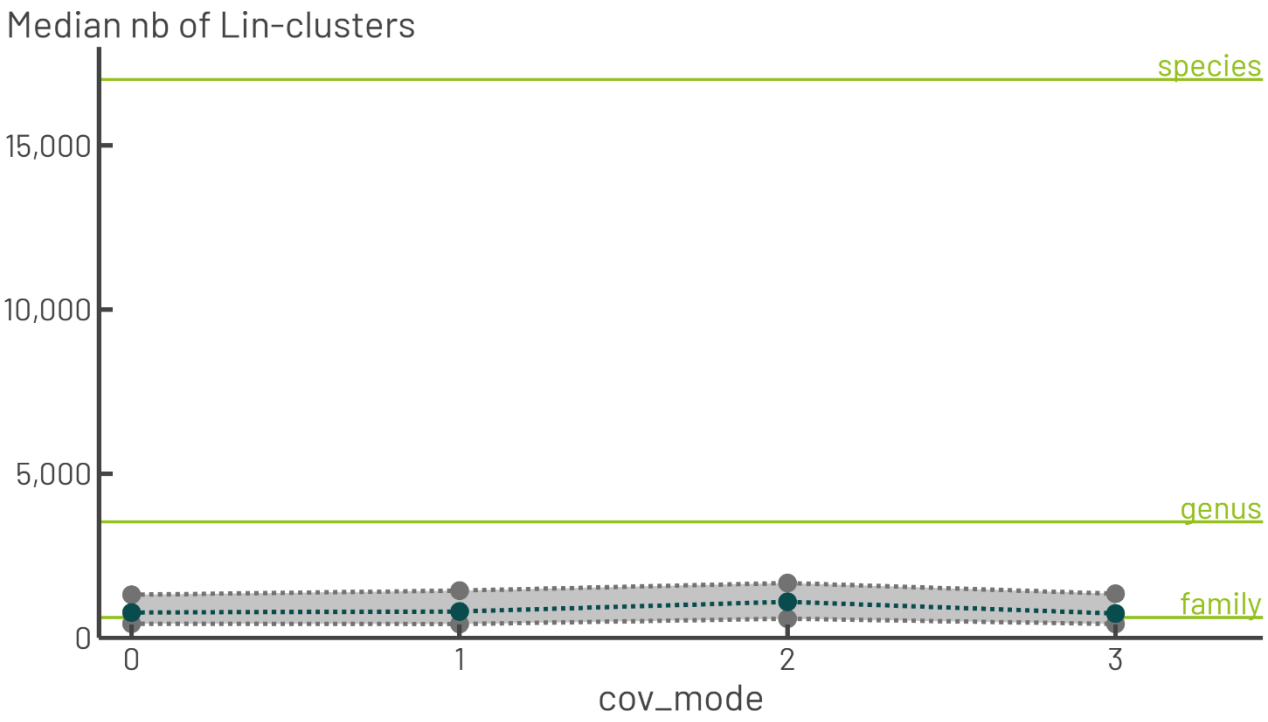
The two figures are coherent because it confirms that, the greater the minNbLinclusters was, the more the overall genomic dataset was divided. Regarding the size of the largest pre-connected component, a first platoon can be seen until 24 (Supplementary Material S19); with the largest pre-connected component comprising most of the dataset, it means that the pre-connection was not able to separate it into several parts. Then the curve is decreasing, with two other observable platoons. Parallelly, the number of pre-connected components was increasing as minNbLinclusters increased, mostly at the end (Supplementary Material S20). For the bacterial dataset, the pre-connection was adjust with the optimal value minNbLinclusters = 25. This leaded to 488 pre-connected components. This means there will be at least 488 MPS-representatives. The size of the pre-connected components whose size varied comprised from 1 to 79,145 genomes. 25 shared Lin-clusters over 48 (≈ 0.5) are required to be pre-connected, which corresponds to a minimal similarity of $\Delta = 0.5$. It means that all similarities from 1 and 0.50 were captured by pre-connection and that the minimum similarity Δ of the complete-linkage was constrained between $\Delta = 1$ and $\Delta = 0.5$. However, Δ could decreases slightly below Δ, down to $\Delta = 0.4$, to accentuate the dereplication of areas containing many relatively close genomes.

For step 4, the minimum similarity Δ was set to seven different values $\Delta \in \{1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4\}$, leading to seven different runs and thus to seven samplings of different sizes. Thanks to complete-linkage, within each MPS-cluster when $\Delta = 0.4$, any pair of genomes share more than 40% of similar sequences, i.e. share r-prots sequences that have been clustered in the same Lin-clusters.

Supplementary Material S15 – Median number of Lin-clusters according to coverage mode

Univariate study of the median number of Lin-clusters according to the coverage mode (**covMode**)
with the fixed parameters: **eValue = 10e-5**, **MinCov = 0.8** and **minSeqID = 0.6**.

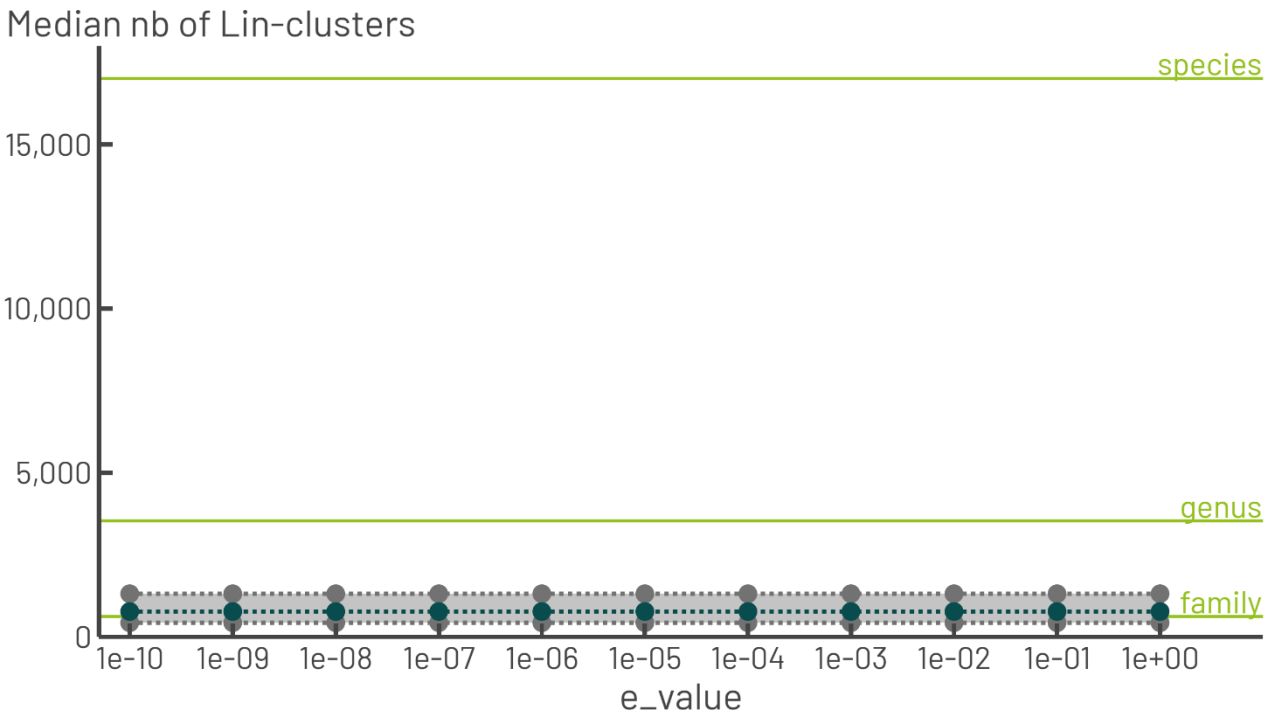
Three lines represent the median number (blue) as well as the first and third quartiles (grey) of Lin-clusters.
The line is almost flat, so the coverage mode did not have a large impact on the sequence cluster finesse.



Supplementary Material S16 – Median number of Lin-clusters according to eValue

Univariate study of the median number of Lin-clusters according to the **eValue**
with the fixed parameters: **covMode = 0**, **minCov = 0.8** and **minSeqID = 0.6**.

Three lines represent the median number (blue) as well as the first and third quartiles (grey) of Lin-clusters.
The line is completely flat, so the eValue had almost no impact on the sequence cluster finesse.

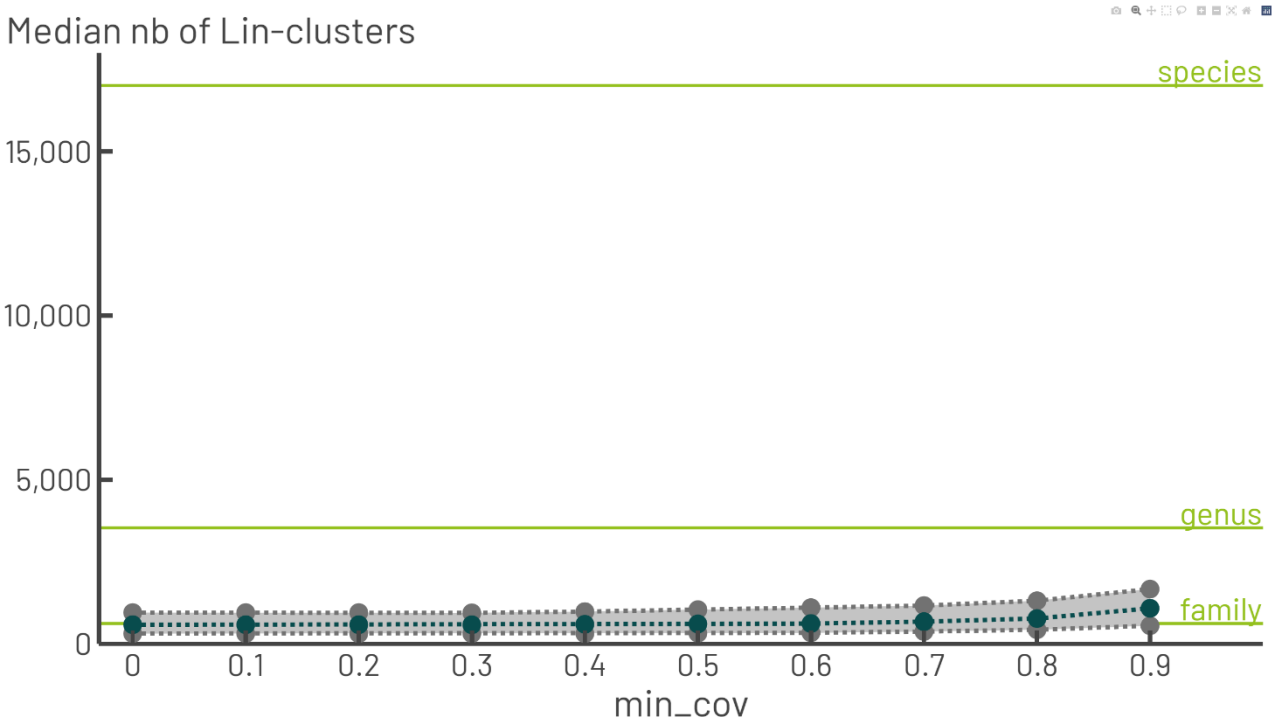


Supplementary Material S17 – Median number of Lin-clusters according to minimum coverage

Univariate study of the median number of Lin-clusters according to the minimal coverage (**minCov**) with the fixed parameters: **covMode = 0**, **eValue = 10e-5** and **minSeqID = 0.6**.

Three lines represent the median number (blue) as well as the first and third quartiles (grey) of Lin-clusters.

The minimum coverage had almost no impact from 0.3 to 0.7. From 0.8 to 0.9, it had a little impact.

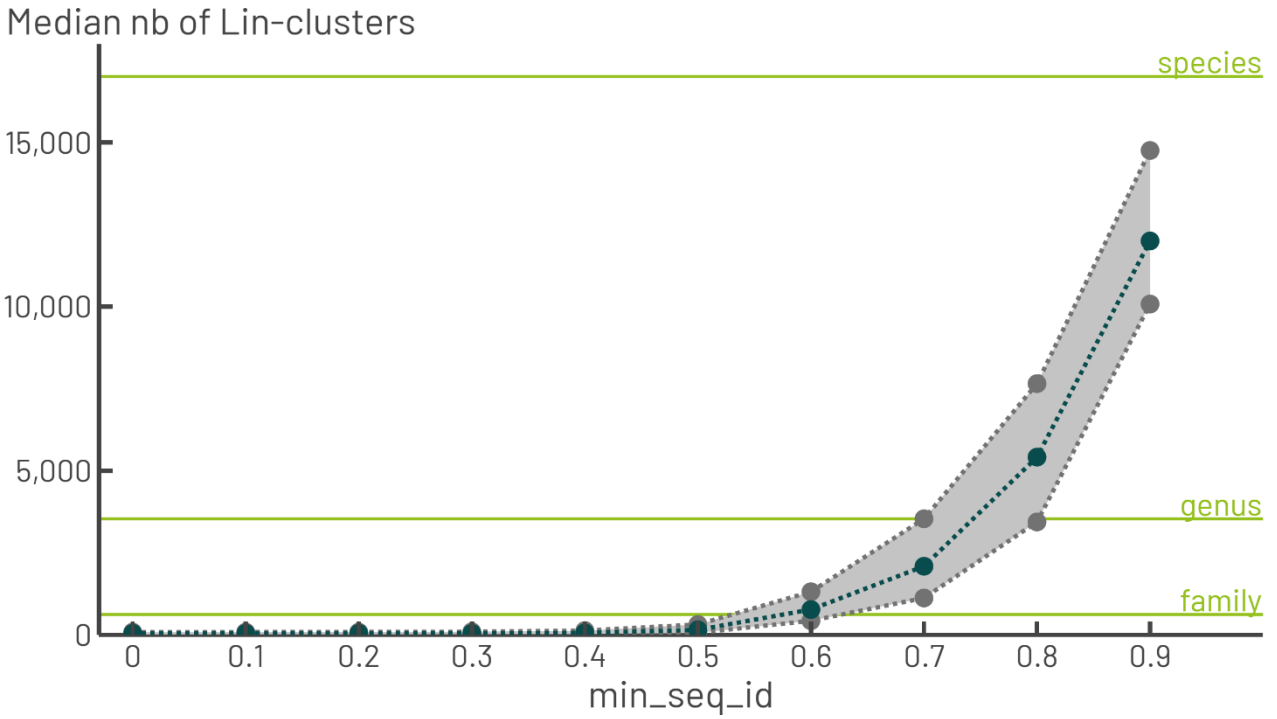


Supplementary Material S18 – Median number of Lin-clusters according to minimum sequence identity

Univariate study of the median number of Lin-clusters according to the minimal sequence identity (**minSeqID)** with the fixed parameters: **covMode = 0**, **eValue = 10e-5** and **covMode = 0.8**.

Three lines represent the median number (blue) as well as the first and third quartiles (grey) of Lin-clusters.

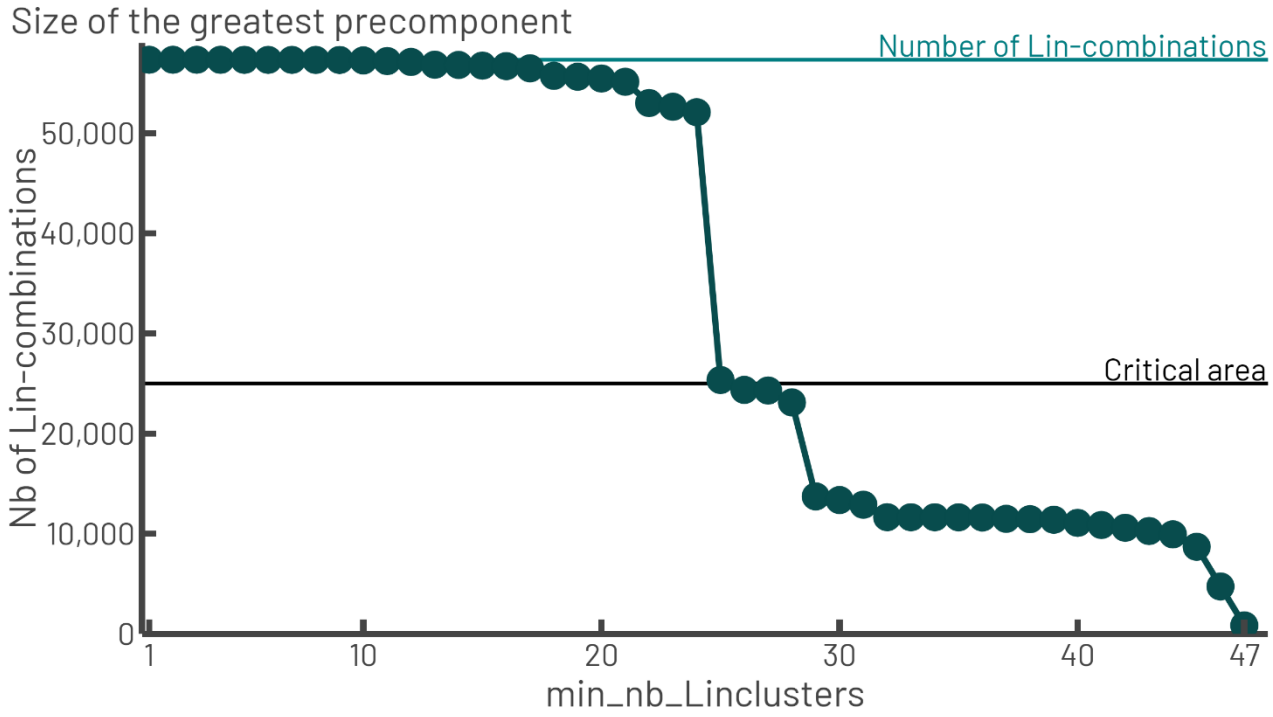
The minimum sequence identity had clearly an important impact on the cluster finesse and constitutes the major parameter to fine-tune for Linclust. By fixing it to **minSeqID = 0.6**, the median number of sequence clusters (773) approached the number of taxonomic families (661).



Supplementary Material S19 – Size of the largest pre-connected component depending on MinNbLinclusters

The decreasing of the curve is coherent because the greater the MinNbLinclusters, the more the overall genomic dataset was divided and the smaller the largest pre-connected component.

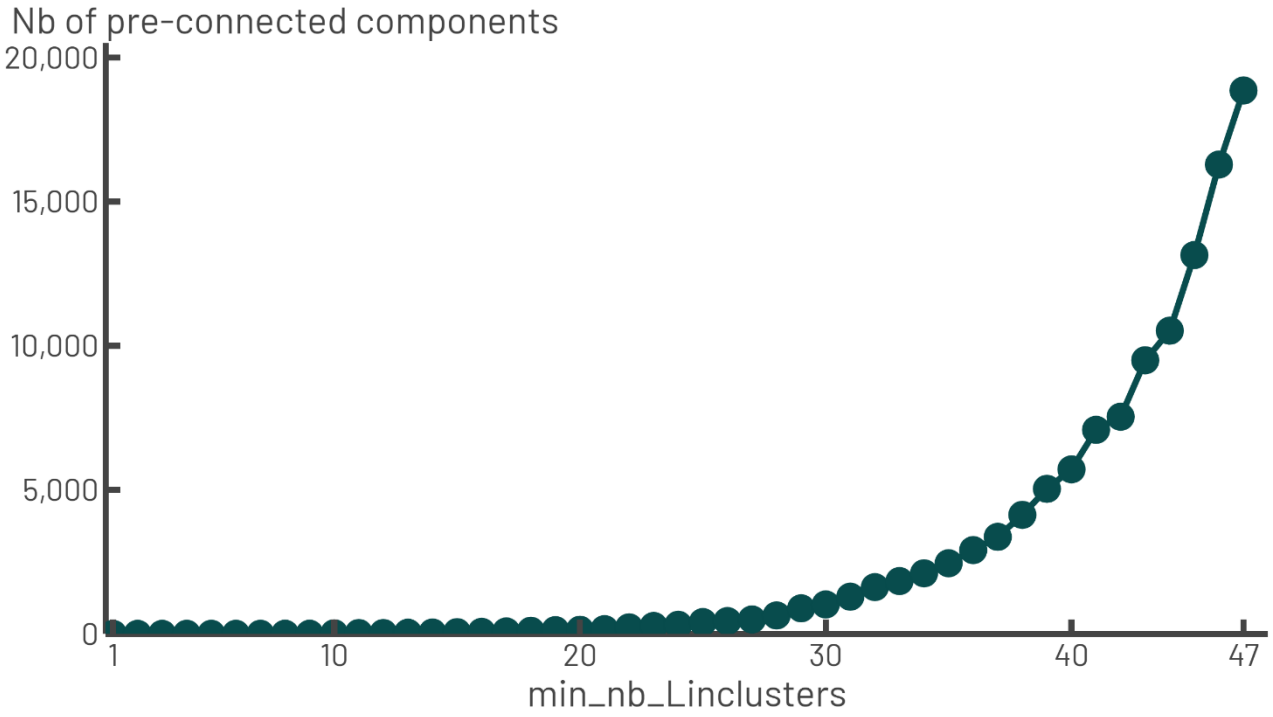
The parameter was set to **MinNbLinclusters = 25**, the smallest value according to the critical area. The largest pre-connected component encompassed 25,351 Lin-combinations, a little above the critical area of 25,000.



Supplementary Material S20 – Number of pre-connected components depending on MinNbLinclusters

The decreasing of the curve is coherent because the greater the MinNbLinclusters, the more the overall genomic dataset was divided and the more pre-connected components there were.

The parameter was set to **MinNbLinclusters = 25**, leading to 488 pre-connected components.



Supplementary Material S21 - Phylogenetic reconstruction

Briefly about phylogenetic inference, for each of the 48 studied r-prot families, protein sequences have been aligned with kalign v3.2.2 (March 22) (Lassmann, 2020). The resulting alignments have been trimmed by removing positions containing more than 30% of gaps and sequences with more than 30% of gaps and combined to build a large supermatrix. A second trimming round has been performed on the supermatrix to remove the blocks of positions containing more than 20% of gaps and sequences containing more than 20% of gaps. These filters are used to limit the amount of missing data in the supermatrices that may bias phylogenetic reconstructions (Philippe *et al.*, 2017). Phylogenies have been reconstructed using FastTree v2.1.11 (Feb 20) (Price *et al.*, 2010). The phylogeny inference of the 178,203 genomes of the bacterial dataset lasted 51.50 hours with a parallelization using 3 cores.

Supplementary Material S22 – Computational time MPS-Sampling concerning the 178,027 genomes of the bacterial dataset

MPS-Sampling was executed through a single- threaded process on a dedicated server with 20 cores / 40 threads (2 Intel Xeon E5- 2660v2 CPUs @2.20Ghz) and 128 GB of DDR3 memory under Debian 10 (Buster).

The computational time of one MPS-sample was 1 + 1 + 2 + 26 + 1 + 7 = 38 min. Because the steps 1 to 4-1 are common to several MPS-samples, the computational time for seven MPS-samples was 1 + 1 + 2 + 26 + (1 + 7) * 7 = 86 min = 1h26.

The most time-consuming step was the computing of similarity submatrices from pre-connected components (step 4-1) due to quadratic complexity. This limitation was partially relaxed thanks to pre-connection, saving a substantial amount of time (see Discussion).

Algorithm	Step	Task	Computational time
-	-	Retrieval and download of ribosomal sequence families from RiboDB website	3 min
-	-	Filter genomes and ribosomal protein families	5 min
-	-	Format FASTA files	4 min
MPS-Sampling	Step 1	Construction of Lin-clusters	1 min
MPS-Sampling	Step 2-1	Labelling of Lin-clusters	1 min
MPS-Sampling	Step 2-2	Construction of the Lin-clustering matrix	
MPS-Sampling	Step 2-3	Re-ordering of the Lin-clustering matrix	
MPS-Sampling	Step 2-4	Construction of the Lin-combination matrix	
MPS-Sampling	Step 3	Construction of pre-connected components	2 min
MPS-Sampling	Step 4-1	Computation of similarity submatrices	26 min
MPS-Sampling	Step 4-2	Construction of MPS-clusters	1 min
MPS-Sampling	Step 5	Selection of MPS-representatives	7 min

Supplementary Material S23 – Intermediate results of MPS-Sampling concerning the bacterial dataset

Starting from the 178,203 bacterial genomes, 59 to 2,789 Lin-clusters per r-prot family were built (median = 773) (step 1, Supplementary Material S24). This reflects sequence variation (e.g., selective pressure, sequencing or assembly errors) among r-prot families: the higher the identity among sequences within a protein family, the smaller the number of Lin-clusters. The number of Lin-clusters was not correlated with the length of the protein sequences (Supplementary Material S25). These Lin-clusters constituted 57,332 Lin-combinations and thus as many EGG, from which 48,296 (84,8%) were singleton (i.e. contain a single genome) (step 2, Supplementary Material S26). In contrast, the three largest encompassed 4.21% of initial dataset (i.e. 2,891, 2,575, and 2,037 genomes, respectively).

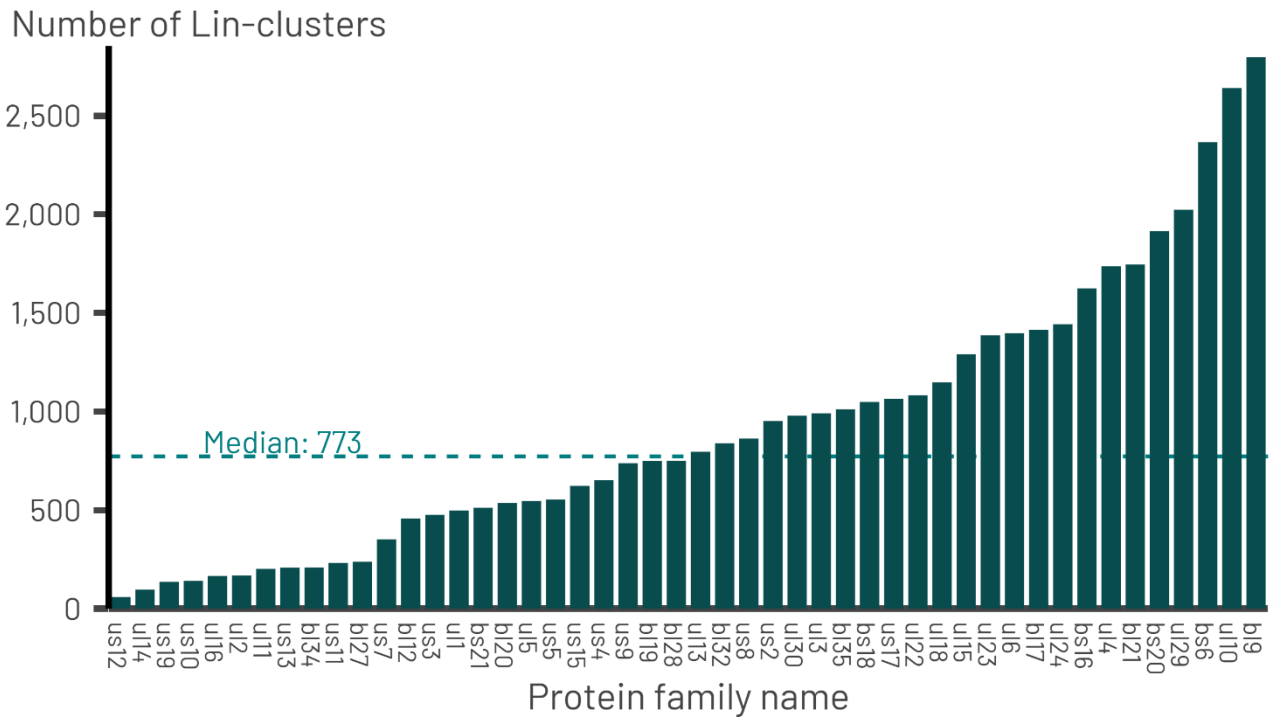
From these Lin-combinations, 488 pre-connected components were built (step 3, Supplementary Material S27) whose size varied greatly: 163 pre-connected components (46%) contained a single genome, while the two largest encompassed together 150,858 genomes and 57,332 Lin-combinations, respectively 84.66% and 76.48% of the bacterial dataset.

Then, MPS-clusters were computed (step 4). At this stage, the higher the value of Δ, the lower the number of MPS-clusters: from 57,332 ($\Delta = 1$) to 3,474 ($\Delta = 0.4$). This corresponded to a sample from 32.17% to 1.95% of the complete bacterial dataset (Figure 2A). As an example, when $\Delta = 0.7$, 12,775 MPS-clusters were built (Supplementary Material S28), among which 5,329 contained a single genome, 2,404 two genomes, while the two largest gather 13,820 and 3,385 genomes, respectively.

The process ended with the selection of the MPS-representatives according to priority rules (step 4). The frequency of each criteria use is shown in the Supplementary Material S29. The most decisive criteria were fame, protein distribution, centrality and pseudo-randomness, respectively.

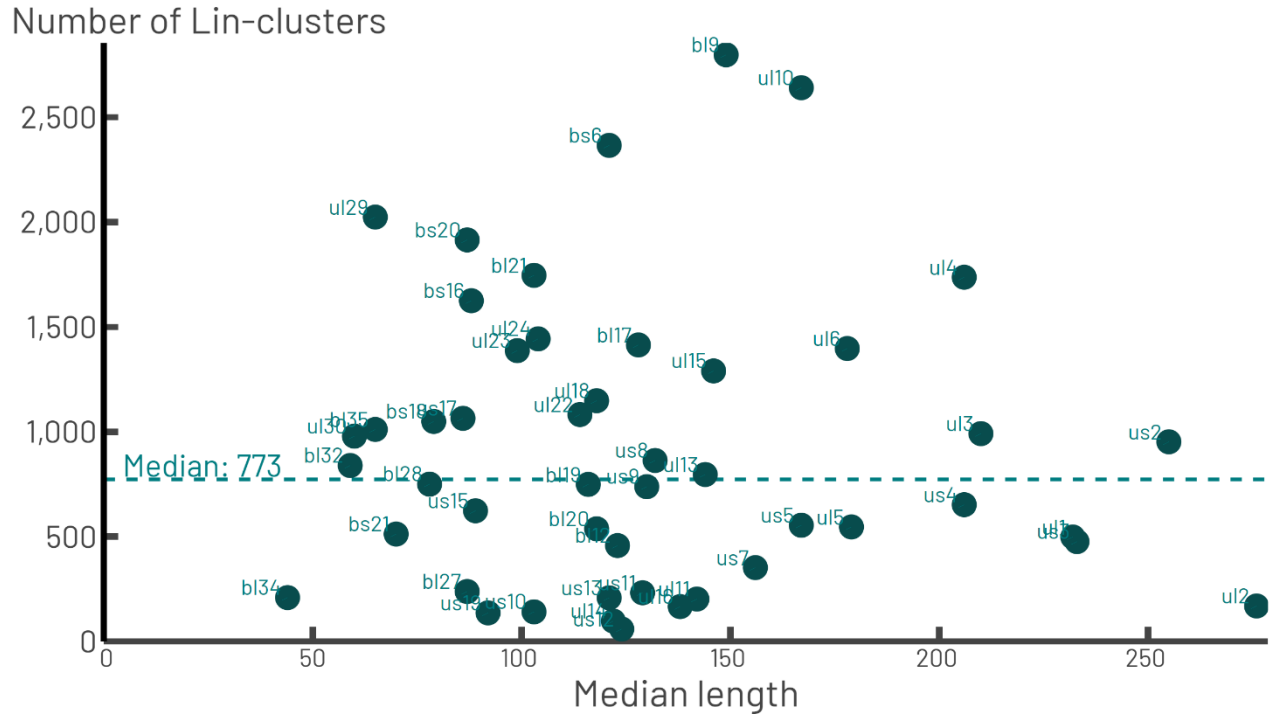
Supplementary Material S24 – Number of Lin-clusters per protein family

During the analysis of the bacterial dataset, the number of Lin-clusters varied from 59 to 2,798 with a median of 773.



Supplementary Material S25 – Median length of protein sequences VS Number of Lin-clusters

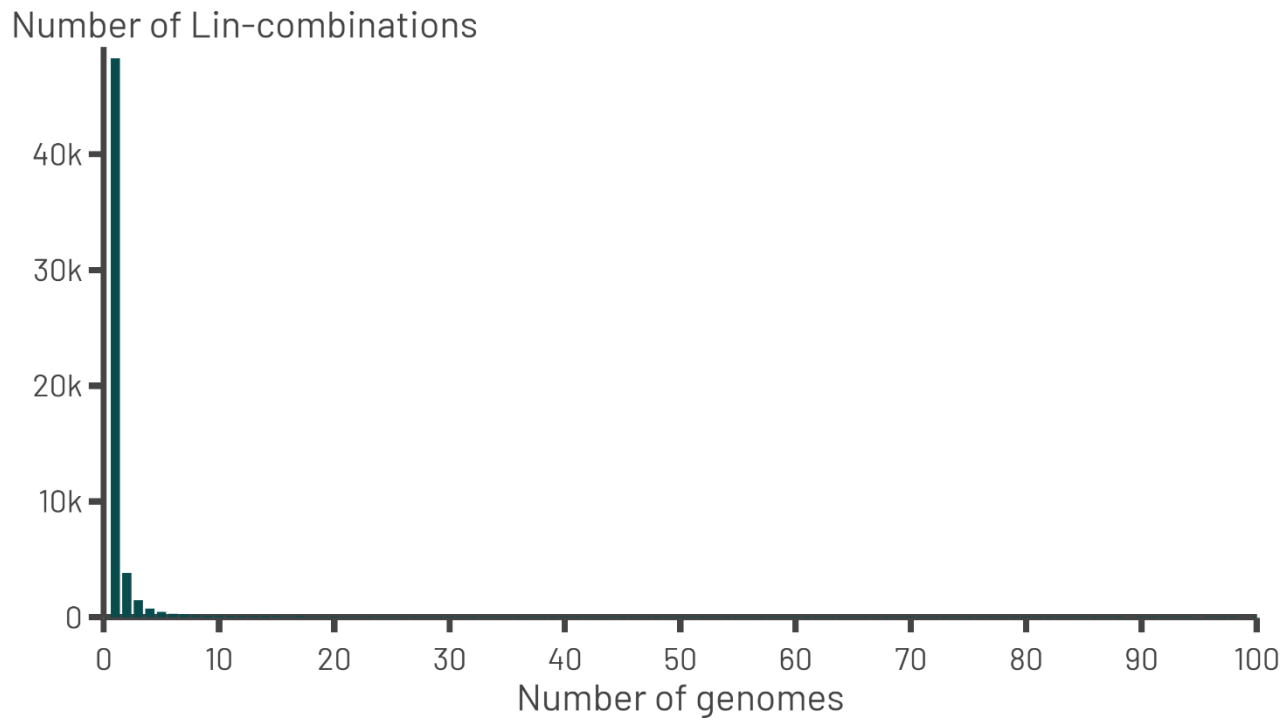
During the analysis of the bacterial dataset, along the 48 considered r-prot families, the number of Lin-clusters was not correlated with the median length of protein sequences.

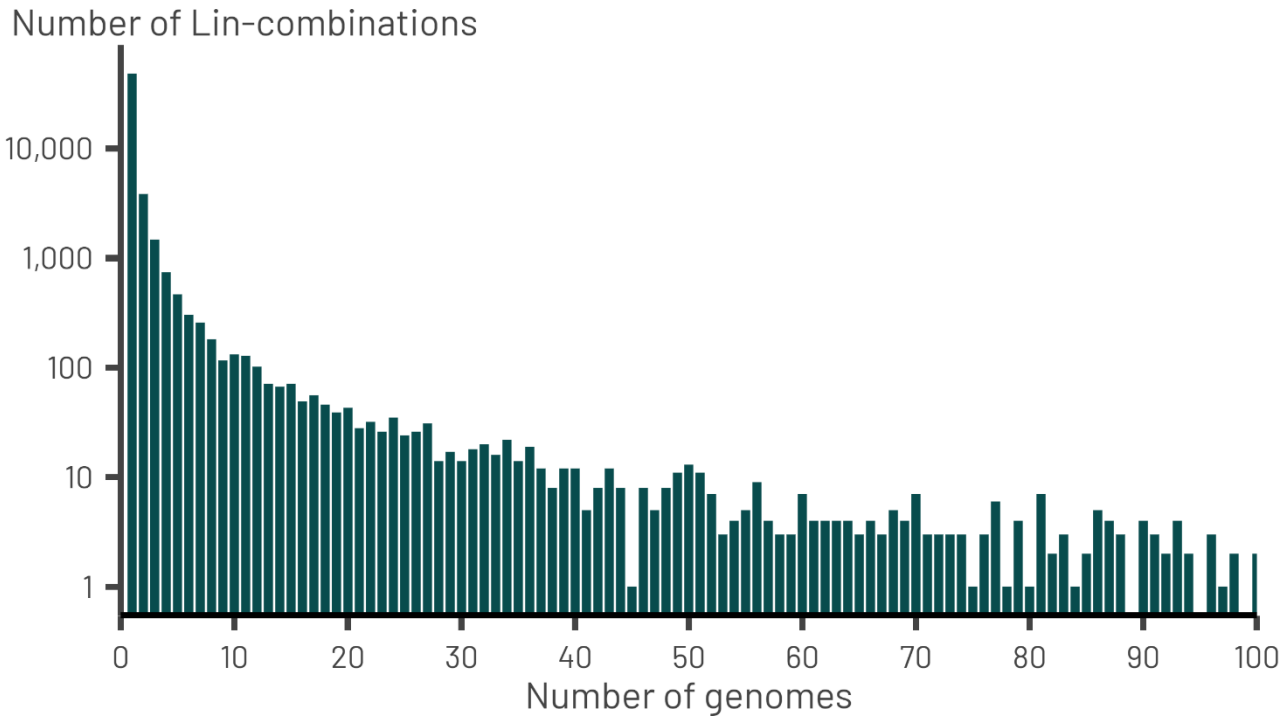


Supplementary Material S26 – Number of genomes within the 57,332 Lin-combinations

During the analysis of the bacterial dataset, 57,332 Lin-combination were built. 48,296 Lin-combinations admitted only one genome. On the contrary, the three largest Lin-combinations contained respectively 2,891, 2,575 and 2,037 genomes. 198 Lin-combinations have more than 100 genomes and are "out of the plot" on the right.

The second histogram has a logarithmic scale on the y-axis.

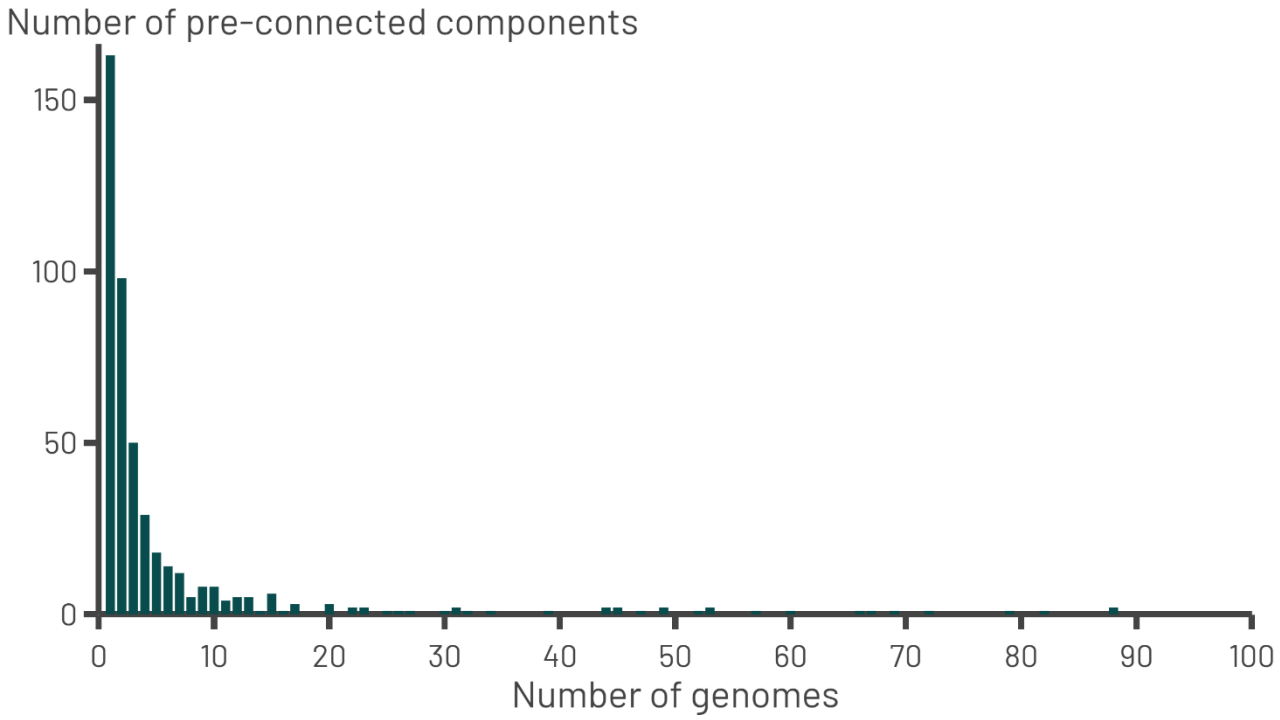


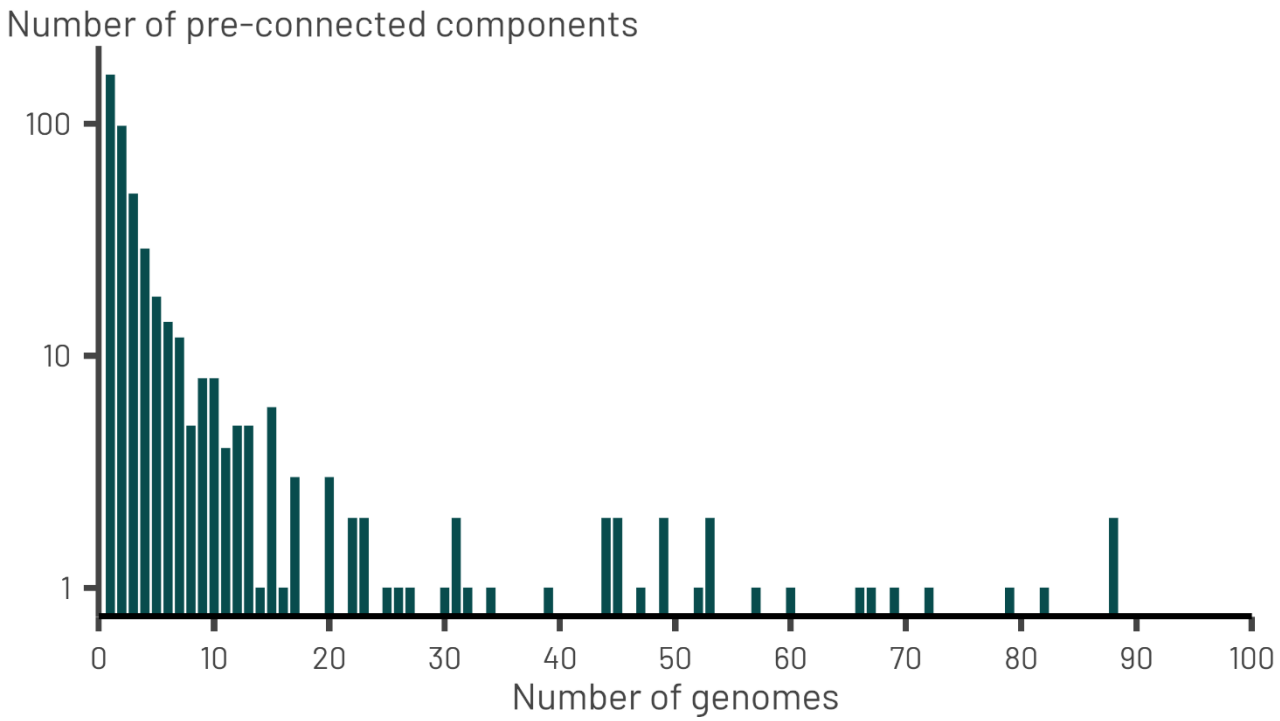


Supplementary Material S27 – Number of genomes within the 488 pre-connected components

During the analysis of the bacterial dataset, 488 pre-connected components were built. 163 pre-connected components admitted only one genome. On the contrary, the two largest pre-connected components contained 79,145 and 71,713 genomes respectively, encompassing 150,858 out of the 178,203 genomes of the bacterial dataset (84.66%). These two largest pre-connected components contained 25,351 and 18,499 Lin-combinations respectively, encompassing 43,850 out of the 57,332 Lin-combinations of the bacterial dataset (76.48%). 22 pre-connected components has more than 100 genomes and are "out of the plot" on the right.

The second histogram has a logarithmic scale on the y-axis.

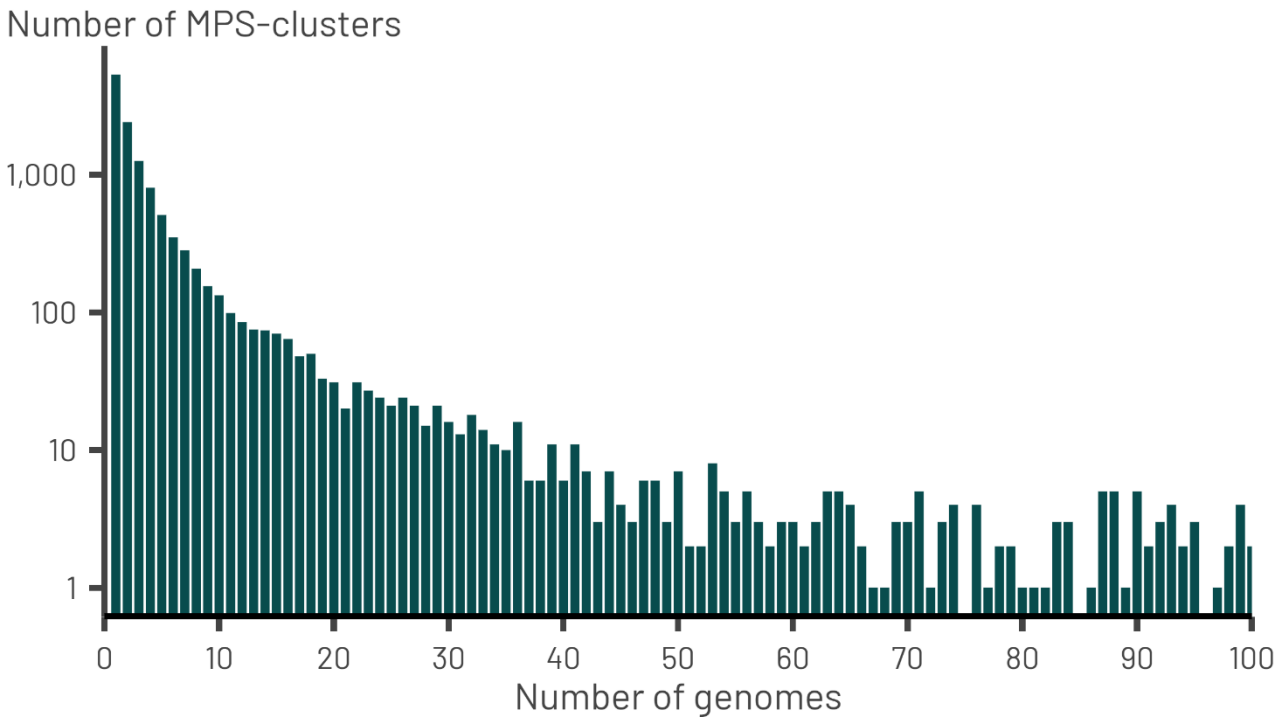


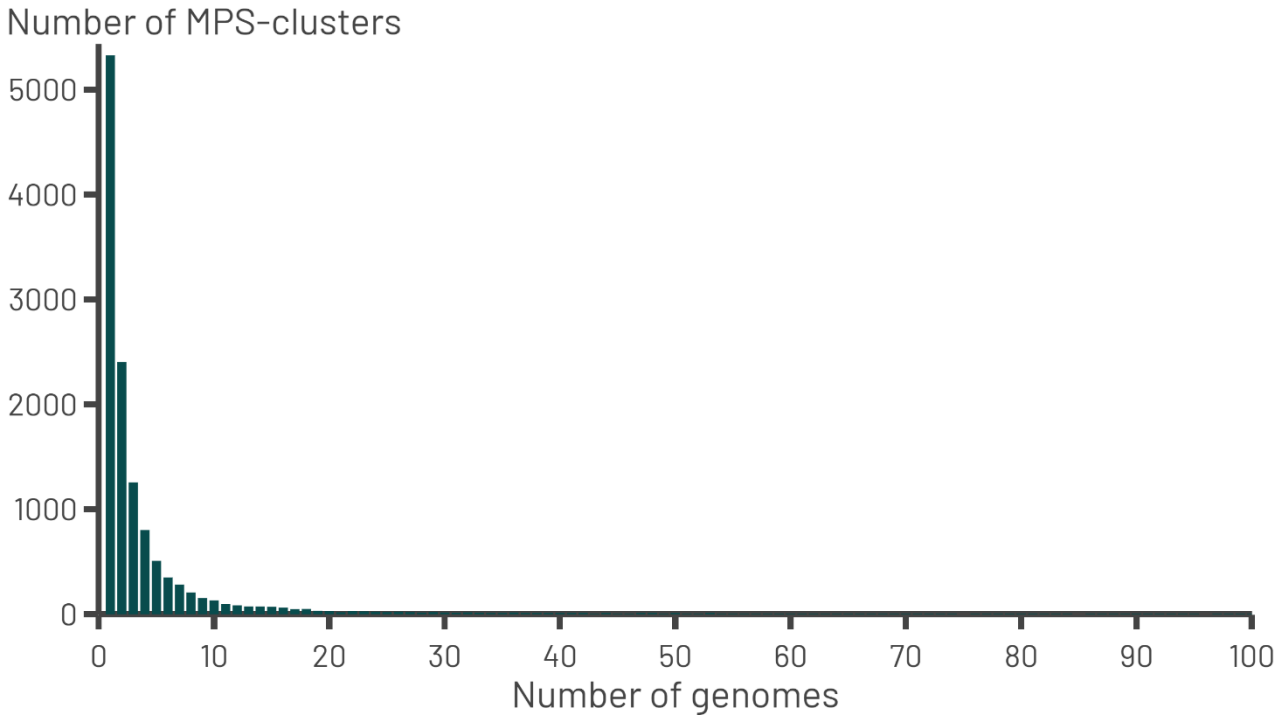


Supplementary Material S28 – Number of genomes within the 12,775 MPS-clusters (Δ=0.7)

During the analysis of the bacterial dataset, 12,775 MPS-clusters were built using Δ=0.7. 5,329 MPS-clusters admitted only one genome and 2,404 have only two genomes. On the contrary, the two largest MPS-clusters contained respectively 13,820 and 3,385 genomes and respectively 563 and 367 Lin-combinations. 194 MPS-clusters have more than 100 genomes and are "out of the plot" on the right.

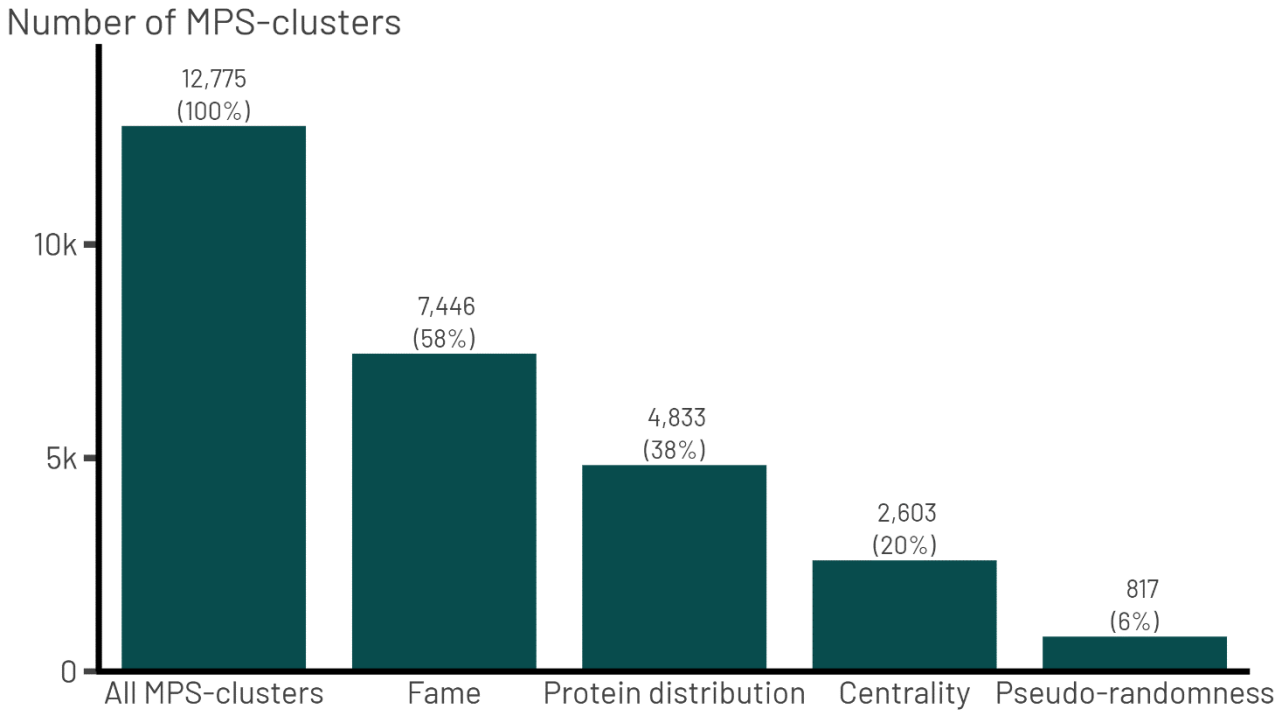
The second histogram has a logarithmic scale on the y-axis.





Supplementary Material S29 – Number of MPS-clusters where each selection rule was applied

With Δ=0.7, 12,775 MPS-clusters were generated. Among them, 12,775 - 7,446 = 5,329 MPS-clusters contain only one genome, so no selection rule needed to be applied. It meant that 7,446 MPS-clusters had more than one genome. The first rule (fame) was applied to all these 7,446 MPS-clusters. After that, 4,833 MPS-clusters still had more than one genome, so the second rule (protein distribution) was applied to these 4,833 MPS-clusters. Then, the third rule (centrality) was applied to 2,603 MPS-clusters. Finally, the last rule (pseudo-randomness) was applied to 817 MPS-clusters.



Supplementary Material S30 – Reduction of the whole *Bacteria* dataset using MPS-Sampling

	Complete dataset	MPS-Sampling runs						
		Δ=1	Δ=0.9	Δ=0.8	Δ=0.7	Δ=0.6	Δ=0.5	Δ=0.4
Genomic size								
Number of genomes	178,203	57,332	30,196	19,117	12,775	8,352	5,347	3,474
	100%	32.17%	16.94%	10.73%	7.17%	4.69%	3.00%	1.95%
Number of genomes with complete taxonomic affiliation	135,315	28,641	12,634	7,698	4,888	3,033	1,872	1,205
	100%	21.17%	9.34%	5.69%	3.61%	2.24%	1.38%	0.89%
Number of genomes with partial taxonomic affiliation	42,888	28,691	17,562	11,419	7,887	5,319	3,475	2,269
	100%	66.90%	40.95%	26.63%	18.39%	12.40%	8.10%	5.29%
Number of genomes with no taxonomic affiliation	44	41	35	30	23	21	14	9
	100%	93.18%	79.55%	68.18%	52.27%	47.73%	31.82%	20.45%
Number of genomes with no phylum affiliation	86	82	70	56	42	36	26	17
	100%	95.35%	81.40%	65.12%	48.84%	41.86%	30.23%	19.77%
Number of genomes with no class affiliation	5,619	4,897	3,620	2,651	2,001	1,488	1,084	817
	100%	87.15%	64.42%	47.18%	35.61%	26.48%	19.29%	14.54%
Number of genomes with no order affiliation	9,507	8,726	6,542	4,835	3,697	2,720	1,922	1,393
	100%	91.79%	68.81%	50.86%	38.89%	28.61%	20.22%	14.65%
Number of genomes with no family affiliation	13,616	12,554	9,437	6,932	5,255	3,817	2,668	1,844
	100%	92.20%	69.31%	50.91%	38.59%	28.03%	19.59%	13.54%
Number of genomes with no genus affiliation	18,663	17,344	12,995	9,387	6,905	4,815	3,175	2,079
	100%	92.93%	69.63%	50.30%	37%	25.80%	17.01%	11.14%
Number of genomes with no species affiliation	41,752	28,077	17,161	11,120	7,663	5,157	3,347	2,172
	100%	67.25%	41.10%	26.63%	18.35%	12.35%	8.02%	5.20%
Number of genomes with at least one Candidatus entry	2,991	2,621	1,723	1,256	946	698	526	431
	100%	87.63%	57.61%	41.99%	31.63%	23.34%	17.59%	14.41%
Number of genomes that are RefSeq-representative	16,135	14,730	11,146	7,518	4,909	3,100	1,947	1,276
	100%	91.29%	69.08%	46.59%	30.42%	19.21%	12.07%	7.91%

	Complete dataset	MPS-Sampling runs						
		Δ=1	Δ=0.9	Δ=0.8	Δ=0.7	Δ=0.6	Δ=0.5	Δ=0.4
Genomic size								
Number of genomes	178,203 100%	178,203 100%	178,203 100%	178,203 100%	178,203 100%	178,203 100%	178,203 100%	178,203 100%
Number of genomes linked to a representative with complete taxonomic affiliation	135,315 100%	142,726 105.48%	150,483 111.21%	155,64 115.02%	158,617 117.22%	161,017 118.99%	162,785 120.30%	164,754 121.76%
Number of genomes linked to a representative with partial taxonomic affiliation	42,888 100%	35,477 82.72%	27,720 64.63%	22,563 52.61%	19,586 45.67%	17,186 40.07%	15,418 35.95%	13,449 31.36%
Number of genomes linked to a representative with no taxonomic affiliation	44 100%	43 97.73%	43 97.73%	38 86.36%	33 75%	36 81.82%	31 70.45%	33 75%
Number of genomes linked to a representative with no phylum affiliation	86 100%	85 98.84%	84 97.67%	83 96.51%	71 82.56%	71 82.56%	78 90.70%	69 80.23%
Number of genomes linked to a representative with no class affiliation	5,619 100%	5,609 99.82%	5,608 99.80%	5,535 98.51%	5,432 96.67%	5,272 93.82%	5,023 89.39%	4,722 84.04%
Number of genomes linked to a representative with no order affiliation	9,507 100%	9,492 99.84%	9,455 99.45%	9,277 97.58%	8,98 94.46%	8,659 91.08%	8,145 85.67%	7,601 79.95%
Number of genomes linked to a representative with no family affiliation	13,616 100%	13,590 99.81%	13,421 98.57%	13,110 96.28%	12,617 92.66%	12,033 88.37%	11,360 83.43%	10,406 76.42%
Number of genomes linked to a representative with no genus affiliation	18,663 100%	18,497 99.11%	17,866 95.73%	16,970 90.93%	15,682 84.03%	14,307 76.66%	12,88 69.01%	11,265 60.36%
Number of genomes linked to a representative with no species affiliation	41,752 100%	34,284 82.11%	26,379 63.18%	21,165 50.69%	18,063 43.26%	15,670 37.53%	13,942 33.39%	12,027 28.81%
Number of genomes linked to a representative with at least one Candidatus entry	2,991 100%	2,988 99.90%	2,995 100.13%	3,027 101.20%	3,050 101.97%	3,006 100.50%	3,126 104.51%	3,224 107.79%
Number of genomes linked to a representative that is RefSeq-representative	134,851 100%	99,677 73.92%	140,556 104.23%	155,319 115.18%	159,235 118.08%	162,622 120.59%	164,555 122.03%	166,451 123.43%
Number of genomes that are singleton MPS-cluster	178,203 100%	48,296 27.1%	19,803 11.11%	9,882 5.55%	5,329 2.99%	2,768 1.55%	1,415 0.79%	742 0.42%

	Complete dataset	MPS-Sampling runs						
		Δ=1	Δ=0.9	Δ=0.8	Δ=0.7	Δ=0.6	Δ=0.5	Δ=0.4
Taxonomic diversity								
Number of phyla	122 100%	122 100%	122 100%	121 99.18%	119 97.54%	119 97.54%	119 97.54%	117 95.90%
Number of classes	114 100%	114 100%	112 98.25%	111 97.37%	111 97.37%	110 96.49%	108 94.74%	107 93.86%
Number of orders	263 100%	263 100%	262 99.62%	261 99.24%	259 98.48%	255 96.96%	251 95.44%	242 92.02%
Number of families	661 100%	661 100%	654 98.94%	648 98.03%	628 95.01%	607 91.83%	567 85.78%	519 78.52%
Number of genera	3,896 100%	3,885 99.72%	3,782 97.07%	3,491 89.60%	3,023 77.59%	2,404 61.70%	1,770 45.43%	1,251 32.11%
Number of species	16,814 100%	15,881 94.45%	11,858 70.52%	7,846 46.66%	5,063 30.11%	3,173 18.87%	1,985 11.81%	1,292 7.68%
Average number of genomes per phylum	1,459.98 100%	469.26 32.14%	246.93 16.91%	157.53 10.79%	107.00 7.33%	69.88 4.79%	44.71 3.06%	29.55 2.02%
Average number of genomes per class	1,501.99 100%	452.07 30.10%	233.31 15.53%	146.33 9.74%	96.02 6.39%	61.95 4.12%	39.31 2.62%	24.77 1.65%
Average number of genomes per order	641.43 100%	184.81 28.81%	90.28 14.08%	54.72 8.53%	35.05 5.46%	22.09 3.44%	13.65 2.13%	8.60 1.34%
Average number of genomes per family	249.00 100%	67.74 27.21%	31.74 12.75%	18.80 7.55%	11.97 4.81%	7.47 3%	4.72 1.90%	3.14 1.26%
Average number of genomes per genus	40.95 100%	10.29 25.14%	4.55 11.11%	2.79 6.81%	1.94 4.74%	1.47 3.59%	1.23 3%	1.12 2.72%
Average number of genomes per species	8.12 100%	1.85 22.71%	1.10 13.55%	1.02 12.56%	1.01 12.44%	1.01 12.41%	1.01 12.41%	1.01 12.40%
Phylogenetic diversity								
Length of the ML tree / Number of genomes	0.0159	0.0489	0.089	0.127	0.1637	0.2056	0.2534	0.3049

Supplementary Material S31 – The *Lactobacillaceae* family within the *Bacteria* reduction using MPS-Sampling

	Complete dataset	MPS-Sampling runs						
		Δ=1	Δ=0.9	Δ=0.8	Δ=0.7	Δ=0.6	Δ=0.5	Δ=0.4
Genomic size								
Number of genomes	6,401	1,094	310	147	80	45	24	12
	100%	17.07%	4.84%	2.29%	1.25%	0.70%	0.37%	0.19%
Number of genomes with complete taxonomic affiliation	6,268	1,039	298	142	77	44	24	12
	100%	16.58%	4.75%	2.27%	1.23%	0.70%	0.38%	0.19%
Number of genomes with partial taxonomic affiliation	142	55	12	5	3	1	0	0
	100%	38.73%	8.45%	3.52%	2.11%	0.70%	0%	0%
Number of genomes with no genus affiliation	2	1	1	1	1	0	0	0
	100%	50%	50%	50%	50%	0%	0%	0%
Number of genomes with no species affiliation	142	55	12	5	3	1	0	0
	100%	38.73%	8.45%	3.52%	2.11%	0.70%	0%	0%
Number of genomes with at least one Candidatus entry	1	1	1	1	1	0	0	0
	100%	100%	100%	100%	100%	0%	0%	0%
Number of genomes that are RefSeq-representative	366	332	236	132	73	44	24	12
	100%	90.71%	64.48%	36.07%	19.95%	12.02%	6.56%	3.28%

	Complete dataset	MPS-Sampling runs						
		Δ=1	Δ=0.9	Δ=0.8	Δ=0.7	Δ=0.6	Δ=0.5	Δ=0.4
Taxonomic diversity								
Number of genera	33	33	33	33	32	27	20	12
	100%	100%	100%	100%	96.97%	81.82%	60.61%	36.36%
Number of species	379	362	254	140	77	44	24	12
	100%	95.51%	67.02%	36.94%	20.32%	11.61%	6.33%	3.17%
Average number of genomes per genus	194.18	33.12	9.36	4.42	2.47	1.67	1.00	1.00
	100%	17.06%	4.82%	2.28%	1.27%	0.86%	0.62%	0.51%
Average number of genomes per species	16.54	2.87	1.17	1.01	1.00	1.00	1.00	1.00
	100%	17.35%	7.09%	6.13%	6.05%	6.05%	6.05%	6.05%
Phylogenetic diversity								
Length of the ML tree / Number of genomes	0.0026	0.0145	0.0469	0.0811	0.1174	0.1458	0.1896	0.2621

Supplementary Material S32 – The *Bacillaceae* family within the *Bacteria* reduction using MPS-Sampling

	Complete dataset	MPS-Sampling runs						
		Δ=1	Δ=0.9	Δ=0.8	Δ=0.7	Δ=0.6	Δ=0.5	Δ=0.4
Genomic size								
Number of genomes	7,113	1,622	688	417	249	142	78	39
	100%	22.80%	9.67%	5.86%	3.50%	2%	1.10%	0.55%
Number of genomes with complete taxonomic affiliation	6,160	1,258	551	370	236	138	76	39
	100%	20.42%	8.94%	6.01%	3.83%	2.24%	1.23%	0.63%
Number of genomes with partial taxonomic affiliation	953	364	137	47	13	4	2	0
	100%	38.20%	14.38%	4.93%	1.36%	0.42%	0.21%	0%
Number of genomes with no genus affiliation	17	13	8	5	4	1	1	0
	100%	76.47%	47.06%	29.41%	23.53%	5.88%	5.88%	0%
Number of genomes with no species affiliation	953	364	137	47	13	4	2	0
	100%	38.20%	14.38%	4.93%	1.36%	0.42%	0.21%	0%
Number of genomes with at least one Candidatus entry	0	0	0	0	0	0	0	0
	-	-	-	-	-	-	-	-
Number of genomes that are RefSeq-representative	625	585	494	362	234	136	74	37
	100%	93.60%	79.04%	57.92%	37.44%	21.76%	11.84%	5.92%

	Complete dataset	MPS-Sampling runs						
		Δ=1	Δ=0.9	Δ=0.8	Δ=0.7	Δ=0.6	Δ=0.5	Δ=0.4
Taxonomic diversity								
Number of genera	108	108	107	103	93	76	55	32
	100%	100%	99.07%	95.37%	86.11%	70.37%	50.93%	29.63%
Number of species	644	619	509	366	236	138	76	39
	100%	96.12%	79.04%	56.83%	36.65%	21.43%	11.80%	6.06%
Average number of genomes per genus	65.70	14.90	6.36	4.00	2.63	1.86	1.00	1.22
	100%	22.67%	9.67%	6.09%	4.01%	2.82%	2.13%	1.85%
Average number of genomes per species	9.57	2.03	1.08	1.01	1.00	1.00	1.00	1.00
	100%	21.25%	11.32%	10.57%	10.45%	10.45%	10.45%	10.45%
Phylogenetic diversity								
Length of the ML tree / Number of genomes	0.0046	0.0199	0.0447	0.0634	0.0840	0.1083	0.1341	0.1643

Supplementary Material S33 – The *Enterobacteriaceae* family within the *Bacteria* reduction using MPS-Sampling

	Complete dataset	MPS-Sampling runs						
		Δ=1	Δ=0.9	Δ=0.8	Δ=0.7	Δ=0.6	Δ=0.5	Δ=0.4
Genomic size								
Number of genomes	17,096	682	113	49	37	28	21	18
	100%	3.99%	0.66%	0.29%	0.22%	0.16%	0.12%	0.11%
Number of genomes with complete taxonomic affiliation	15,692	550	82	36	26	22	16	13
	100%	3.50%	0.52%	0.23%	0.17%	0.14%	0.10%	0.08%
Number of genomes with partial taxonomic affiliation	1,404	132	31	13	11	6	5	5
	100%	9.40%	2.21%	0.93%	0.78%	0.43%	0.36%	0.36%
Number of genomes with no genus affiliation	32	29	15	7	6	3	2	2
	100%	90.62%	46.88%	21.88%	18.75%	9.38%	6.25%	6.25%
Number of genomes with no species affiliation	1,404	132	31	13	11	6	5	5
	100%	9.40%	2.21%	0.93%	0.78%	0.43%	0.36%	0.36%
Number of genomes with at least one Candidatus entry	80	71	40	30	27	23	17	14
	100%	88.75%	50%	37.50%	33.75%	28.75%	21.25%	17.50%
Number of genomes that are RefSeq-representative	204	130	74	40	31	25	19	16
	100%	63.73%	36.27%	19.61%	15.20%	12.25%	9.31%	7.84%

	Complete dataset	MPS-Sampling runs						
		Δ=1	Δ=0.9	Δ=0.8	Δ=0.7	Δ=0.6	Δ=0.5	Δ=0.4
Taxonomic diversity								
Number of genera	58	52	38	27	22	20	16	13
	100%	89.66%	65.52%	46.55%	37.93%	34.48%	27.59%	22.41%
Number of species	178	139	67	33	24	22	16	13
	100%	78.09%	37.64%	18.54%	13.48%	12.36%	8.99%	7.30%
Average number of genomes per genus	294.21	12.56	2.58	1.56	1.41	1.25	1.19	1.23
	100%	4.27%	0.88%	0.53%	0.48%	0.42%	0.40%	0.42%
Average number of genomes per species	88.16	3.96	1.22	1.09	1.08	1.00	1.00	1.00
	100%	4.49%	1.39%	1.24%	1.23%	1.13%	1.13%	1.13%
Phylogenetic diversity								
Length of the ML tree / Number of genomes	0.0008	0.0186	0.1002	0.2016	0.2525	0.2912	0.3311	0.3580

Supplementary Material S34 – Phylogenetic statistics about MPS-samples

For each studied subset, a phylogenetic tree was reconstructed with the input genomes. The length of all branches divided by the number of leaves was computed on the whole tree and on each subtree of the MPS-sample.

For example, for the *Lactobacillaceae* family, the initial tree encompassed 6,410 genomes and was reduced from 1,094 to 12 MPS-representatives. The phylogenetic diversity starts from 0.0026 and decreased from 0.0145 down to 0.2621.

Studied subset	Initial data	MPS-Sampling runs						
		Δ=1	Δ=0.9	Δ=0.8	Δ=0.7	Δ=0.6	Δ=0.5	Δ=0.4
Bacteria								
Number of leaves	178,203	57,332	30,196	19,117	12,775	8,352	5,347	3,474
Length of the ML tree / Number of leaves	0.0159	0.0489	0.089	0.127	0.1637	0.2056	0.2534	0.3049
Bacterial backbone								
Number of leaves	35,103	31,631	24,248	17,547	12,775	8,352	5,347	3,474
Length of the ML tree / Number of leaves	0.0716	0.0793	0.1009	0.1299	0.1590	0.2003	0.2473	0.2983
Lactobacillaceae								
Number of leaves	6,410	1,094	310	147	80	45	24	12
Length of the ML tree / Number of leaves	0.0026	0.0145	0.0469	0.0811	0.1174	0.1458	0.1896	0.2621
Bacillaceae								
Number of leaves	7,113	1,622	688	417	249	142	78	39
Length of the ML tree / Number of leaves	0.0046	0.0199	0.0447	0.0634	0.0840	0.1083	0.1341	0.1643
Enterobacteriaceae								
Number of leaves	17,096	682	113	49	37	28	21	18
Length of the ML tree / Number of leaves	0.0008	0.0186	0.1002	0.2016	0.2525	0.2912	0.3311	0.3580

Supplementary Material S35 – Tags frequency among the 178,203 genomes of the bacterial dataset

The seven tags are standing for :

- R: RefSeq-representative.
- T: Type strain.
- E: Ensembl! Bacteria
- C: Complete assembly
- S: Scaffold
- U: Unassembled
- d: doubtful

For instance, 16,135 genomes are tagged as RefSeq-representative (R). 12,410 genomes are tagged as RefSeq-representative (R) and type strain (T). 5,859 genomes are tagged as RefSeq-representative and Ensembl! Bacteria. Etc...

tags	R	T	E	C	S	U	d
R	16135	12410	5859	4681	4706	6748	1
T	12410	18664	6449	3996	6049	8569	68
E	5859	6449	22239	4869	7721	8927	1734
C	4681	3996	4869	33785	0	0	474
S	4706	6049	7721	0	58868	0	7142
U	6748	8569	8927	0	0	72067	0
d	1	68	1734	474	7142	0	21099

Supplementary Material S36 – Taxonomic statistics about each investigated subset

MPS-Sampling was launched on the bacterial dataset, encompassing 178,203 bacterial genomes. In addition to the complete dataset, some subsets were investigated. The constitution of the bacterial backbone, a subset encompassing 35,103 genomes, is described in the Supplementary Material S38. Three taxonomic families have also been investigated, respectively *Lactobacillaceae Bacillaceae* and *Enterobacteriaceae*.

For example, the family *Lactobacillaceae* encompassed 6,410 genomes, including 4,992 (98%) with a complete nomenclature. It involves 361 species and 32 genera, with 11.28 species per genus in average. Concerning the taxonomic density, there were 13.83 and 159.72 genomes per species and genus in average, respectively.

Subset		Genome Number			Taxonomic Diversity		Taxonomic Density		
Level	**Name**	**Total**	**Having a complete nomenclature**		**Number of species** *(without sp. and subsp.)*	**Number of genera**	**Species per genus**	**Genomes per species** *(from complete nomenclature)*	**Genomes per genus** *(from complete nomenclature)*
domain	bacterial dataset	178,203	135,315	76%	16,814	3,896	4.32	8.12	40.95
domain	bacterial backbone	35,103	16,836	48%	16,228	3,787	4.29	1.07	5.13
family	*Lactobacillaceae*	6,410	6,268	98%	379	33	11.48	16.56	194.18
family	*Bacillaceae*	7,113	6,160	87%	644	108	5.96	9.57	65.70
family	*Enterobacteriaceae*	17,096	15,692	92%	178	58	3.07	88.16	294.21

Supplementary Material S37 – Phylogenic statistics about each phylogenetic inference

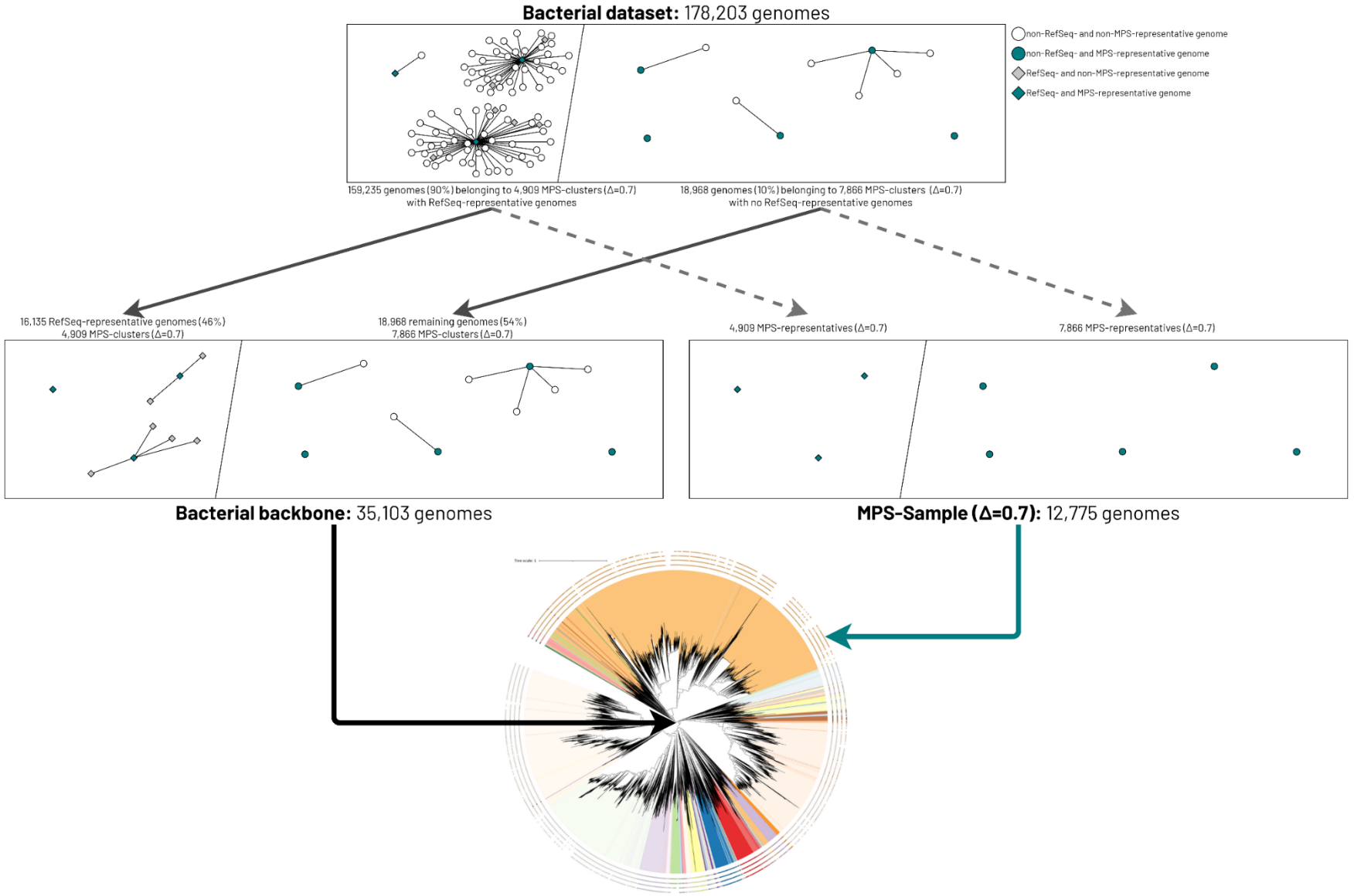
For each investigated group (level + group name), a set of genomes was used for phylogenetic inference (genomes number). After recruitment, alignment and trimming, a supermatrix was built with a given number of rows (sequences number) and columns (positions number); moreover, it had a given number of missing values (gaps number) and a given proportion of missing values among all cells (gaps ratio). From the computed phylogenetic tree, the sum of all branch lengths was divided by the number of tips; this quotient gave an indication about the phylogenetic diversity of the genomic set (total length / number of tips).

For example, the family *Lactobacillaceae* encompassed 6,410 genomes, i.e. 6,410 sequences (one sequence per genome) and 6,105 positions. So, the supermatrix had 6,410 rows and 6,105 columns. There were 1,020,435 gaps in the supermatrix, i.e. 2.61% of it. After the phylogenetic inference, the sum of all branch lengths divided by the number of tips was equal to 0.0026. This provides an indicator about the phylogenetic inference.

Subset		Supermatrix					Tree
Level	**Group Name**	**Genomes Number**	**Sequences Number**	**Positions Number**	**Gaps Number**	**Gaps Ratio**	**Total length / Number of tips**
domain	bacterial dataset	178,203	178,203	5,874	30,880,709	2.95%	0.0159
domain	bacterial backbone	35,103	35,103	5,819	10,386,208	5.08%	0.0716
family	*Lactobacillaceae*	6,410	6,410	6,105	1,020,435	2.61%	0.0026
family	*Bacillaceae*	7,113	7,113	6,138	434,529	1.00%	0.0046
family	*Enterobacteriaceae*	17,096	17,096	6,188	514,233	0.49%	0.0008

Supplementary Material S38 – Construction of the bacterial backbone

For visualization, the bacterial dataset of 178,203 genomes was reduced to a bacterial backbone of 35,103 genomes. More precisely, all the 16,135 RefSeq representative genomes were kept, as they were considered as a standard against which other data should be compared. When Δ = 0.7, these genomes were MPS-representatives and/or belonged to 4,909 MPS clusters, that together represent 159,235 genomes (90%). It was assumed that these 16,135 RefSeq-representative genomes were a reliable reference to represent these 159,235 genomes. The remaining 18,968 genomes (10%) were distributed across the 7,866 MPS clusters that did not contain any RefSeq representative genomes. Because these genomes could not be linked to any reference external to our analysis, all of them were kept for the phylogenetic analysis. Thus, in total, 16,135 + 18,968 = 35,103 genomes were used to infer a reference phylogeny from the bacterial dataset.



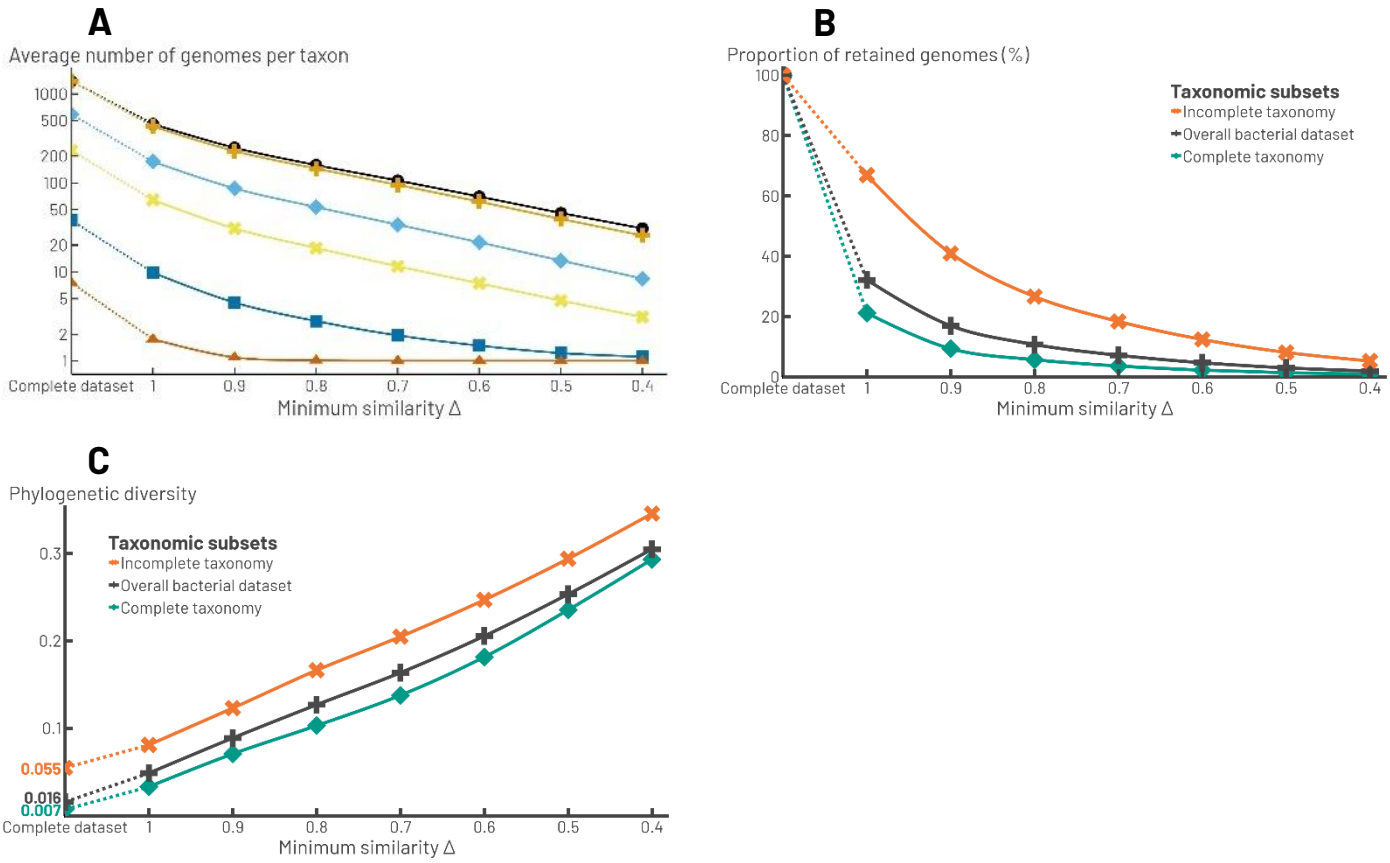
Supplementary Material S39 – Reduction and taxonomic affiliation

A: Average number of genomes representing each taxonomic level in the samples.

B: Genomic reduction of the 135,315 genomes and the 42,888 genomes with a complete and incomplete taxonomic affiliation, respectively, with a normalized scale.

C: Phylogenetic diversity of three subsets: the overall bacterial dataset, the subset of genomes with complete taxonomic affiliation and the subset of genomes with incomplete taxonomic affiliation. The phylogenetic diversity was computed by the length of all branches divided by the number of leaves.

All precise statistics are available in Supplementary Material S30.



Supplementary Material S40 – Inspection of *Bacteria* samplings at a local scale

Sampling of *Lactobacillaceae* family (*Firmicutes* phylum)

The *Lactobacillaceae* family included 6,401 genomes from 33 genera and 379 species (Supplementary Material S31). The taxonomic distribution was balanced, with on average 17 genomes per species. Because the *Lactobacillaceae* was well studied and characterized, a high level of redundancy was observed at both taxonomic and genetic levels: the average number of genomes per genus and species was more than four times higher in this taxonomic family than in the complete bacterial dataset (194 and 41 genomes per genus, and 17 and 8 genomes per species, respectively). This redundancy was also obvious at the phylogenetic level, as the diversity of *Lactobacillaceae* was 3 times lower than for the bacterial dataset (0.0047 and 0.0159, respectively) (Supplementary Material S41B).

Consistently, the dereplication of *Lactobacillaceae* genomes was more intense than for *Bacteria* (Supplementary Material S41A). For instance, when $\Delta = 1$, even if 100% of the genera and 96% of the species were conserved, only 17% of the genomes were kept. As Δ decreased, the number of genomes per genus and per species gradually decreased to 1 (for $\Delta \leq 0.9$ and $\Delta \leq 0.5$ respectively), meaning that all the intra-species and intra-genus redundancy was eliminated.

Compared to *Bacteria*, the reduction of *Lactobacillaceae* was 2 to 8 times higher (Supplementary Material S41A), while the genetic diversity of the *Lactobacillaceae* gradually increased to reach a comparable level when $\Delta = 0.4$ (Supplementary Material S41B). The phylogenetic mapping of MPS-representatives was also reliable, with a denser sampling in regions of the tree with greater phylogenetic diversity (Supplementary Material S42A).This illustrated the ability of MPS-Sampling to adjust sampling intensity according to taxonomic and genetic redundancy in the data, but also to homogenize the taxonomic and genetic diversity of a data set at different evolutionary scales.

Sampling of *Bacillaceae* family (*Firmicutes* phylum)

The case of the *Bacillaceae* was more complex, because this family may not be monophyletic (Maayer *et al.*, 2019). In this context, it was quite possible that part of the reconstructed MPS-clusters mixed genomes from *Bacillaceae* and other families, leading to sampling biases. In the bacterial dataset, the *Bacillaceae* family included 7,113 genomes from 108 genera and 644 species (Supplementary Material S32). However, their taxonomic distribution was unbalanced (Supplementary Material S42B), with the genus *Bacillus* alone comprising three-quarters of the *Bacillaceae* genomes (5,260 genomes out of 7,113). Despite this, the samplings were reliable regarding the phylogenetic distribution of the MPS-representatives (Supplementary Material S42B). In particular, the overrepresentation of *Bacillus* was successfully reduced as they account for 18% to 10% of the samples, 44 out of 249 when $\Delta = 0.7$ and 4 out of 39 when $\Delta = 0.4$, respectively. In fact, most MPS-representatives belonged to genera other than *Bacillus*, demonstrating the ability of MPS-Sampling to capture the genetic diversity of *Bacillaceae*. It also balanced the representativeness of *Bacillaceae*, whose taxonomic and genetic diversity in the samples became comparable to that of *Lactobacillaceae* and *Bacteria* (Supplementary Material S41B).

Through the example of *Bacillaceae*, MPS-Sampling showed its ability to produce relevant samples even in cases where the initial level of redundancy was very high and unbalanced, but also when taxonomy and phylogeny disagreed.

Sampling of the *Enterobacteriaceae* family (*Proteobacteria* phylum)

The *Enterobacteriaceae* represented another interesting case because their taxonomy was more unbalanced than that of the *Bacillaceae*. The *Enterobacteriaceae* family included 17,096 genomes from 178 species and 58 genera, which theoretically corresponded to an average of 88 and 294 genomes per species and per genus, respectively (Supplementary Material S33). However, this was far from the case

as six genera (*Klebsiella*, *Enterobacter*, *Salmonella*, *Shigella*, *Escherichia*, and *Citrobacter)* accounted for 92% of the genomes (15,728 out of 17,096). Consistently, even with the densest sampling ($\Delta = 1$), a few genomes (3.99%) were kept, compared to 32.17%, 17.07%, and 22.80% for *Bacteria*, *Lactobacillaceae*, and *Bacillaceae*, respectively (Supplementary Material S41A). The situation was even more extreme when $\Delta = 0.4$: as 0.11% of the genomes were retained, compared to 1.95%, 0.19%, and 0.55% for *Bacteria*, *Lactobacillaceae*, and *Bacillaceae*, respectively (Supplementary Material S41A). From $\Delta \leq 0.7$, the six most represented genera were reduced to only 1 MPS-representative. Regarding the taxonomic density, one genome per species and genus was kept for *Enterobacteriaceae*, from $\Delta \leq 0.9$ and $\Delta \leq 0.7$ respectively, indicating that as for *Bacteria*, *Lactobacillaceae*, and *Bacillaceae*, most of the redundancy within genera and species was eliminated. However, dereplication was much higher, as 37.93% of the genera conserved when $\Delta = 0.7$, compared to 96.97%, and 86.11% for *Lactobacillaceae*, and *Bacillaceae*, respectively. This reflected a much lower inter-genera and inter-species diversity in *Enterobacteriaceae* than in the two others families. This might be due to biases in the delineation of taxa, which might reflect historical legacy and/or practical convenience, as in the case of the genera *Shigella* and *Escherichia* (Lan and Reeves, 2002). Regardless of the origin of these biases, taxonomy-based sampling would have led to an over-representation of *Enterobacteriaceae*, but also of higher taxa (i.e. *Enterobacterales* and *Gammaproteobacteria*) in the samples.

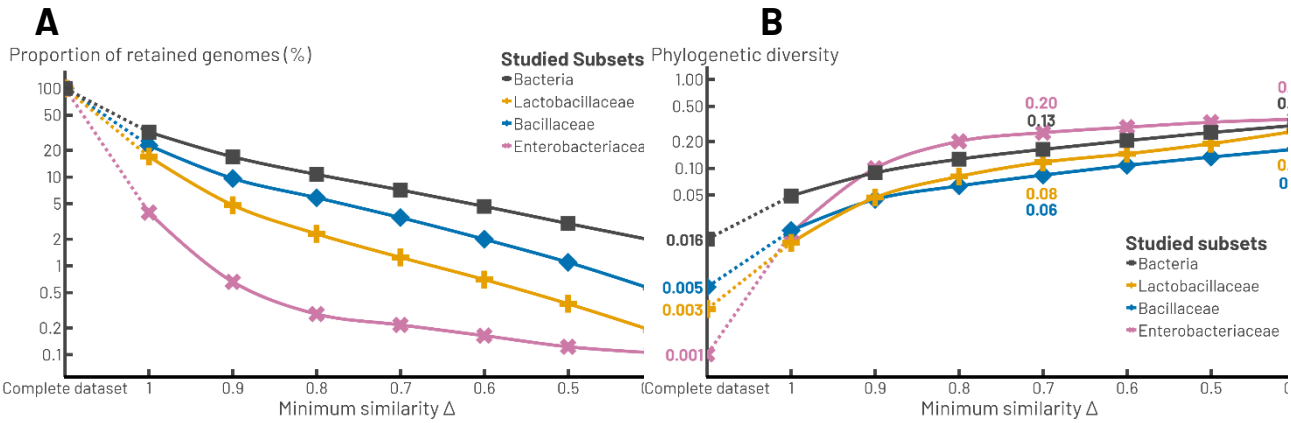
Considering phylogenetic diversity, *Enterobacteriaceae* were initially less diversified than *Bacteria*, *Lactobacillaceae*, and *Bacillaceae* (0.0008, compared to 0.0159, 0.0026, and 0.0046, respectively (Supplementary Material S42C). As with other taxonomic groups, the phylogenetic diversity of *Enterobacteriaceae* increased as sampling density decreased. More precisely, although *Enterobacteriaceae* has the lowest initial phylogenetic diversity, the diversity of its MPS-samples is higher than for the other groups from $\Delta = 0.8$ to $\Delta = 0.4$ (Supplementary Material S42C). This indicated that the true phylogenetic diversity of *Enterobacteriaceae* was hidden by its extreme genetic redundancy. And indeed, the *Enterobacteriaceae* included highly divergent strains and species, with very diverse lifestyles and habitats (e.g., insect endosymbionts, plant and animal pathogens, extremophiles…), some of which might have high rates of evolution. While being phylogenetically agnostic, but thanks to its ability to adapt the sampling density, MPS-Sampling succeeded in revealing the true phylogenetic diversity of *Enterobacteriaceae*, despite a very high initial redundancy.

Supplementary Material S41 – Reduction of three taxonomic families

A: Genomic reduction of the three taxonomic families *Lactobacillaceae*, *Bacillaceae* and *Enterobacteriaceae*, in comparison with *Bacteria*.

C: Phylogenetic diversity of the three taxonomic families *Lactobacillaceae*, *Bacillaceae* and *Enterobacteriaceae*, in comparison with *Bacteria*, computed by the length of all branches divided by the number of leaves,

All precise statistics are available in Supplementary Material S30-S25.



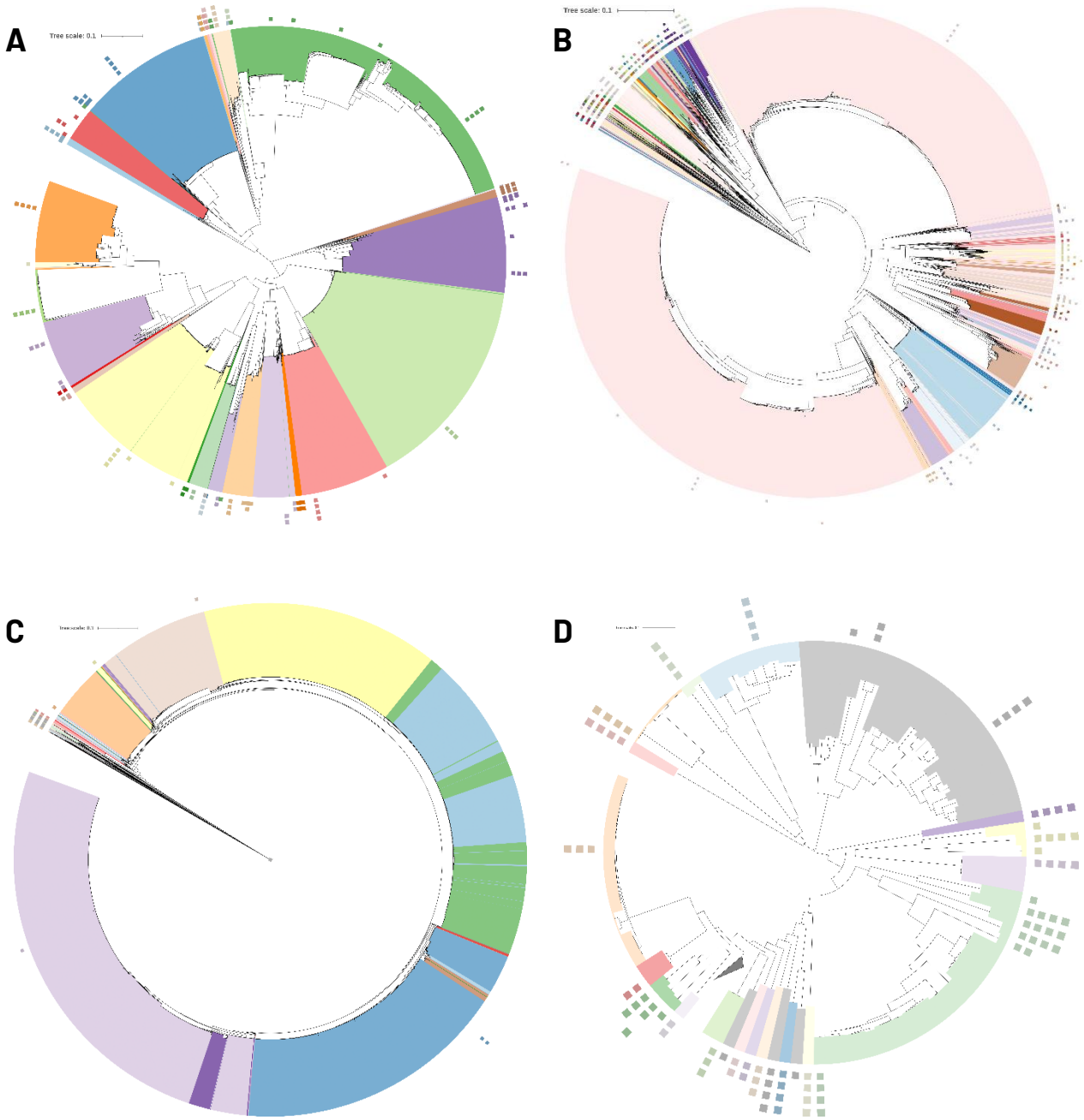
Supplementary Material S42 – Phylogenetic distribution of MPS-representatives of three taxonomic families

A: ML tree of the 6,410 *Lactobacillaceae* genomes present in the bacterial dataset. The tree has been inferred with r-prot sequences. From the innermost to the outermost circle: MPS-representatives corresponding to Δ=0.7 (80 MPS-representatives), Δ=0.6 (45 MPS-representatives), Δ=0.5 (24 MPS-representatives), and Δ=0.4 (12 MPS-representatives). Colors correspond to the 33 genera of *Lactobacillaceae*.

B: ML tree of the 7,113 *Bacillaceae* genomes present in the bacterial dataset. The tree has been inferred with r-prot sequences. From the innermost to the outermost circle: MPS-representatives corresponding to Δ=0.7 (249 MPS-representatives), Δ=0.6 (142 MPS-representatives), Δ=0.5 (78 MPS-representatives), and Δ=0.4 (39 MPS-representatives). Colors correspond to the 108 genera of *Bacillaceae*.

C: ML tree of the 17,096 *Enterobacteriaceae* genomes present in the bacterial dataset. The tree has been inferred with r-prot sequences. From the innermost to the outermost circle: MPS-representatives corresponding to Δ=0.7 (37 MPS-representatives), Δ=0.6 (28 MPS-representatives), Δ=0.5 (21 MPS-representatives), and Δ=0.4 (18 MPS-representatives). Colors correspond to the 58 genera of *Enterobacteriaceae*.

D: ML tree of the 17,096 *Enterobacteriaceae* genomes with 16,987 leaves collapsed. It highlights the 18 Candidatus genera contained within the *Enterobacteriaceae* genomes, representing the large majority of the MPS-samples: 27 out of 37 and 14 out of 18 MPS-representatives are Candidatus genomes, for Δ=0.7 and Δ=0.4 respectively.



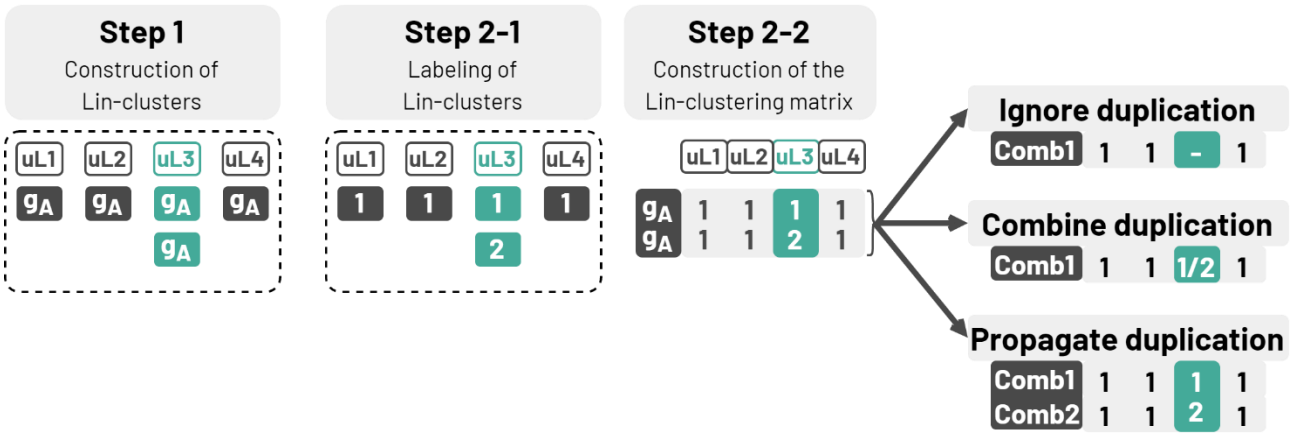
Supplementary Material S43 – Handling duplicated sequences

In this example, there is a duplication case for genome $\mathbf{g_A}$: two duplicated copies have been found for the third protein family **uL3**. The first copy laid in the Lin-cluster 1 and the second copy laid in the Lin-cluster 2. This leads to consider two Lin-combinations for the genome $\mathbf{g_A}$: **(1111)** and **(1121)**.

The first possibility is to ignore duplication cases. This is what is currently done in MPS-Sampling. As the duplication case involved the third protein family **uL3**, it leads to consider a gap for this third protein family; the genome $\mathbf{g_A}$ will be encoded as the Lin-combination **(11-1)**.

The second possibility is to combine the Lin-clusters of duplication cases. In this case, the Lin-clusters of the two duplicated copies lead to a new label (1/2). This is the easiest way to handle duplication cases, because it does not increase the number of Lin-combinations. However, it is very stringent because this duplication case (1/2) will be considered and matched only with exactly the same duplication case (1/2).

The third possibility is to create a new Lin-combination for each duplicated copy. It is the most exhaustive but also the most complex way to handle duplication cases. However, it has two major issues. First, a genome could now be linked to several different Lin-combinations. It complexifies the encoding of genomes into Lin-combinations and its interpretation. Second, it may largely increase the number of Lin-combinations. Let's take an example. With 3 cases of 2-copies duplication, a genome **g** will be associated to $2^3 = 8$ different Lin-combinations. First, the 8 different Lin-combinations could potentially lay in 8 different MPS-clusters, so how to propagate this result to the genome **g**. Second, it could rapidly increase the number of Lin-combinations to handle.



Supplementary Material S44 – Gain of MPS-Sampling compared to quadratic complexity

The bacterial dataset encompassed 178,203 genomes. Processing its complete similarity matrix would require N = 178,203 × (178,203 - 1) / 2 = 15,878,065,503 pairwise comparisons. Through the first dereplication step (i.e. step 2: construction of the EGG), the 178,027 genomes were reduced to 57,332 Lin-combinations. The 57,332 Lin-combinations should induce 57,332 × (57,332 - 1) / 2 = 1,643,450,446 pairwise comparisons, which was 10% of N and a gain of 90%. The second clustering step (i.e. step 3: pre-connection) provided 488 pre-connected components. By computing similarity submatrix among Lin-combinations within each pre-connected component, the number of pairwise comparisons was reduced to 520,218,712, which was 3% of N and a gain of 97%. Last, the computing time of the submatrices was 26 min, which should take 26 / 0.03 = 867 min = 14h without the gain of 97%. Regarding the memory usage, the largest similarity submatrix contained 25,351 Lin-combinations, which required 321,323,925 pairwise comparisons, which was 2% of N and a gain of 98%. As the memory usage was 40 GB, it should lead to 40 / 0.02 = 2,000 TB = 40 GB.

Supplementary Material S45 – Computational times of alternatives approaches

FastANI (Jain *et al.*, 2018) would take ~153,000 CPU hours to compute a whole pairwise distance matrix for 178,203 genomes. A pairwise comparison using FastANI takes around ~5s according to Palmer et al. (Palmer *et al.*, 2020), but it's potentially when only one unique computation is done. In the original article (Jain *et al.*, 2018), it is said that "took 77 K CPU hours for all 8.01 billion comparisons". Assuming that 84,499 genomes have been used 91,761 - 2,262) and that pairs are examined twice, it leads to an exact number of n_1 = 8,009,981,502 comparisons. Then the time in seconds per comparison is given by t = 77,000 * 3600 / 8,009,981,502 ≈ 0.0346. For 178,203 genomes and without doubling the pairs, n_2 = 15,878,065,503 comparisons are needed, leading to an anticipated computing time of $t \times n_2 \approx 549{,}489{,}379$ seconds, i.e. ~153K CPU hours.

Mash (Ondov *et al.*, 2016) would take ~239 CPU hours to compute a whole pairwise distance matrix for 178,203 genomes. With 54,118 genomes, the sketch time is 31.3 hours and the dist time 17.4 hours, with the default parameters (k=21 and s=1,000) (Ondov *et al.*, 2016). Assuming that the sketch time is linear and the dist time quadratic, the computational time for 178,203 genomes would be respectively 103 hours and 189 hours, for a total of 239 CPU hours. The first result is given by :

$$31.3 \times (\frac{178{,}203}{54{,}118}) \approx 103.07$$

And the second by :

$$\frac{178{,}203 \times (178{,}203 - 1)}{2} \times 17.4 \times \frac{2}{54{,}118 \times (54{,}118 - 1)} = \frac{178{,}203 \times (178{,}203 - 1)}{54{,}118 \times (54{,}118 - 1)} \times 17.4 \approx 188.67$$

Treemmer (Menardo *et al.*, 2018) would take ~721 CPU hours to reduce a tree with 178,203 leaves to 3,474 representatives (the same number as for $\Delta = 0.4$). Treemmer being quadratic, the total number of computed pairwise comparisons is N = 178,203 * (178,0203 - 1) / 2 = 15,878,065,503. Treemmer has processed 15,500 tips in 5 days, so the remaining number of comparisons to compute is (178,203 – 15,500) * (178,203 - 15,500 – 1) = 13,236,051,753, which represents ~83.36% of the proportion of N. Then the total anticipated time is 721 K CPU hours, given by :

$$120/(1 - \frac{13{,}236{,}051{,}753}{15{,}878{,}065{,}503}) \approx 721$$

Last, FastTree lasted 51.50 CPU hours to generate a tree of 178,203 genomes and Treemmer should last 721 CPU hours to reduce it. Thus the total processing time for this phylogeny method is 51.50 + 721 = 772.50 CPU hours.

Supplementary Material S46 – Single-linkage VS Complete-linkage (example)

Educational example of a sampling based on hierarchical clustering of a simulated dataset of 131 points in a 2-dimensions space. To begin with, clusters were built according to hierarchical clustering, up to a given threshold. Then, a representative was chosen per cluster, most central as possible. The most central point was the point whose average Euclidian distant to other points within the cluster was minimal.

On the left, three runs using single-linkage with different linkage thresholds. On the right, three runs using complete-linkage with different group limits. The hierarchical thresholds were set to provide around the same number of representatives for single-linkage and complete-linkage.

Supplementary Material S47 – Single-linkage VS Complete-linkage (text)

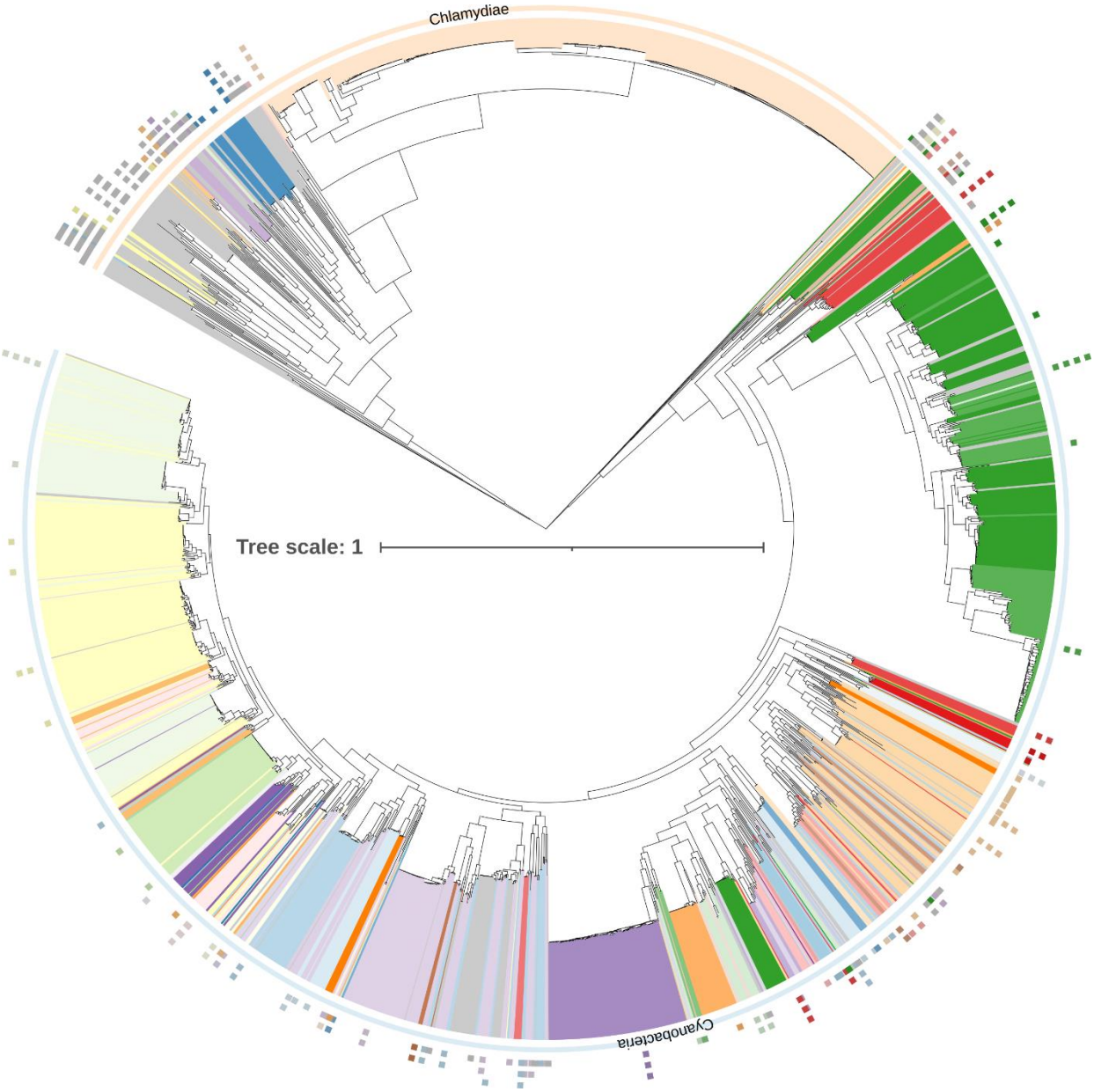
On the one hand, single-linkage provided unbalanced clusters. Some clusters covered a lot of space, potentially due to snake-effects, while other clusters were very narrow. On the first run (A1), some large clusters (blue, orange or light orange) covered a lot of space while other isolated points could have been grouped, for example in the middle or the top-right corner. On the second run (A2), the disparity became more pronounced. The largest cluster (blue) covered most of the space while some isolated points on the top and the top-right corner could have been grouped. On the third run (A3), the disparity was even more visible. The largest cluster (blue) covered almost all the data while other clusters were in comparison very small. As a result, there was only one representative in the center, coming from the largest cluster (blue). At the same time, there were 4 peripherical representatives at the top, 4 on the top-right corner and 2 at the bottom. Peripherical areas were over-sampled. The sampling density on the top was higher than in the middle. This was contrary to the initial data distribution, because the initial data was more dense in the middle than in the top. In this case, sampling based on single-linkage completely inversed the distribution of the initial data.

On the other hand, complete-linkage provided more balanced clusters, mainly because the diversity of each group was equally bounded. On the first run (B1), built clusters were quite small. However, the sampling density was regular all over the data. On the second run (B2), built clusters were still rather small. Even if some clusters are larger than other clusters, all clusters have more or less the same size, according to the diversity. The cluster diversity and the sampling density were much more regular for complete-linkage (B2) than for single-linkage (A2), because clusters have the same size in B2 unlike the clusters with unbalanced size of A2. On the third run (B3), clusters were larger. In the sample, the central essential part was represented by 9 points while 2 representatives remain for outliers, at the top and at the right. In comparison, the third run of single-linkage (A3) completely sub-sample the central part of the data with only 1 representative while over-representing the peripherical parts at the top, the top-right and the bottom with 10 representatives, which is almost all the sample (10 out of 11).

Supplementary Material S48 – Phylogenic distribution of MPS-representatives for the two phyla : *Chlamydiae* and *Cyanobacteria*.

The two phyla *Chlamydiae* and *Cyanobacteria* encompassed respectively 568 and 1,363 genomes, all represented in this phylogenetic tree of 1,932 leaves.

Four circles of MPS-representatives were tagged, chosen with Δ=0.7, Δ=0.6, Δ=0.5 and Δ=0.4. On the top, the phylum *Chlamydiae* was reduced from 51 to 13 MPS-representatives. On the bottom, the phylum *Cyanobacteria* was reduced from 119 to 8 MPS-representatives.

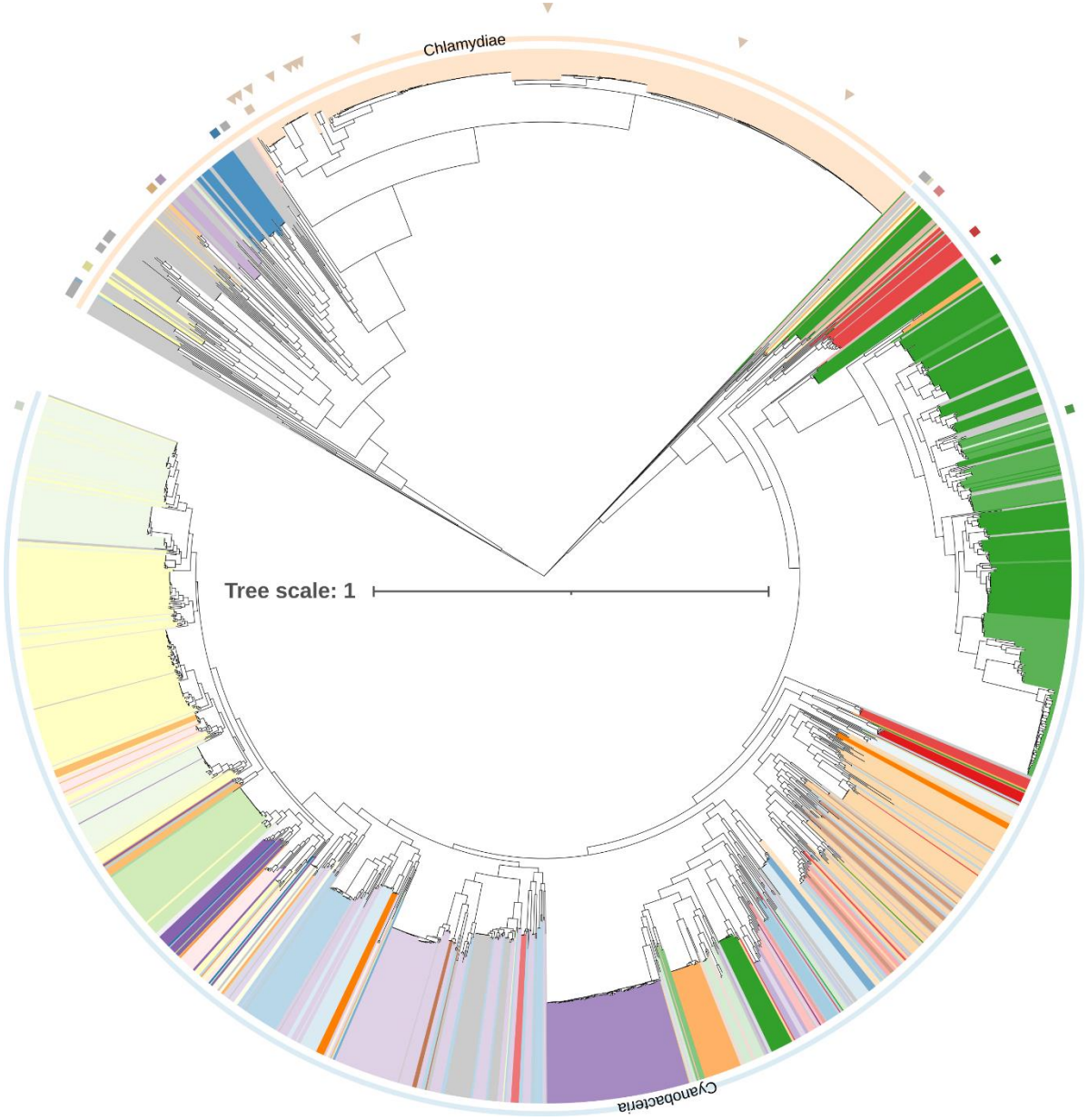


Supplementary Material S49 – MPS-representatives and TQMD-representatives for the two phyla : *Chlamydiae* and *Cyanobacteria*.

The two phyla *Chlamydiae* and *Cyanobacteria* encompassed respectively 568 and 1,363 genomes, all represented in this phylogenetic tree of 1,932 leaves.

On the first circle with squares, the 21 MPS-representatives chosen with Δ=0.4 were tagged, respectively 13 for *Chlamydiae* and 8 for *Cyanobacteria*.

On the second circle with triangles, the 11 TQMD-representatives were tagged, respectively 11 for *Chlamydiae* and 0 for *Cyanobacteria*. TQMD was run to reduce a dataset of 63,863 genomes to a sample 151 TQMD-representatives (Léonard *et al.*, 2021). Among these 63,863 genomes, the phylum *Chlamydiae* encompassed 360 genomes and was reduced to 11 TQMD-representatives while the phylum *Cyanobacteria* encompassed 428 genomes and was reduced to 0 TQMD-representative.

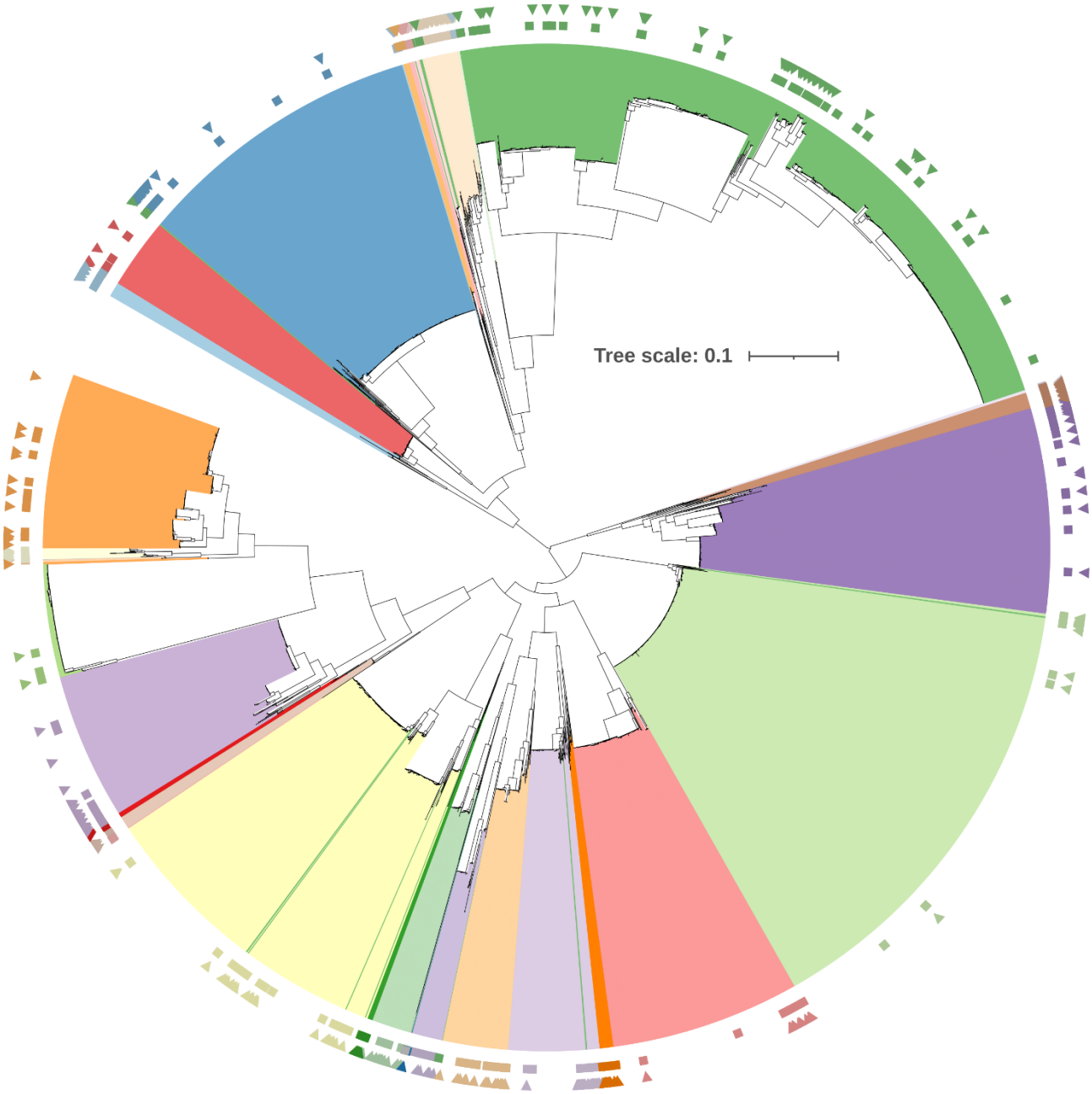


Supplementary Material S50 – MPS-representatives and RefSeq-representatives for *Lactobacillaceae*

Among the 6,401 *Lactobacillaceae* genomes, the 310 MPS-representative genomes selected with Δ=0.9 (squares, inner circle) were compared to the 366 RefSeq-representatives (triangles, outer circle).

The two circles of squares and triangles were quite regular. It demonstrated that both MPS-representatives and RefSeq-representatives were well-distributed along the phylogenetic tree. Thanks to the priority rules, the intersection of the 366 RefSeq-representatives and the 310 MPS-representatives was made of 236 genomes. It means that RefSeq-sample and this MPS-sample shared a similarity of 70%, regarding the Dice index.

To bring more insight, a zoom in the *Bacillus* genus is provided in the Supplementary Material S51, to compare the seven MPS-samples with the unique RefSeq-sample.



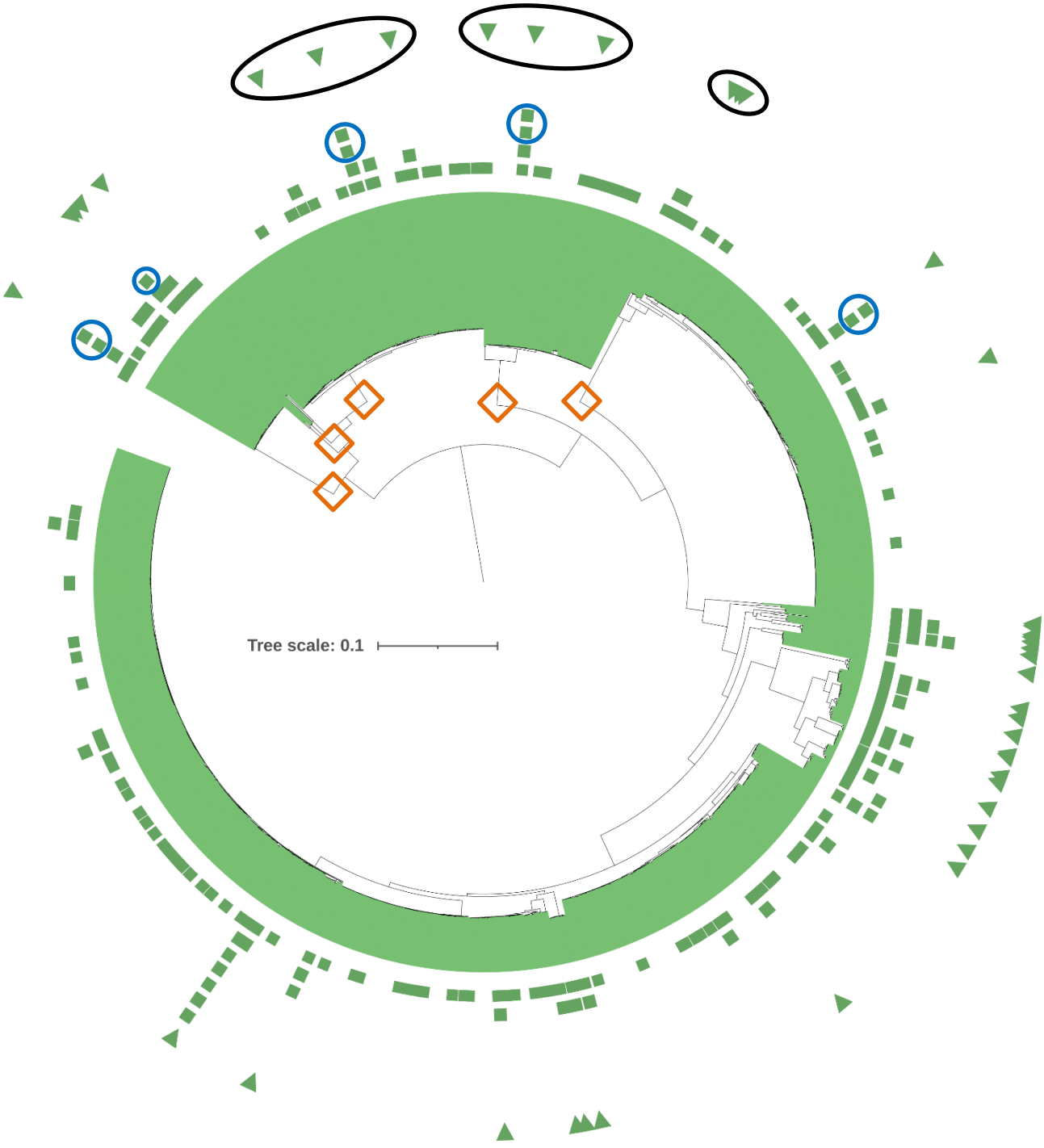
Supplementary Material S51 – MPS-representatives and RefSeq-representatives for *Lactobacillus*

Among the 1,517 *Lactobacillus* genomes, MPS-representatives (squares) and RefSeq-representatives (triangles) were compared.

Seven circles corresponded to seven samples of *Lactobacillus*, encompassing 242, 50, 19, 9, 2, 1 and 1 MPS-representatives, chosen respectively with $\Delta \in \{1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4\}$. Regarding RefSeq-representatives, an unique circle corresponded to 44 RefSeq-representatives.

As shown, only one RefSeq-sample was available with 44 genomes while MPS-Sampling provided seven different samples, from 242 to 1 genomes, the closest size being the sample of 50 genomes with $\Delta = 0.9$. Moreover, new samples could be easily generated using new values of Δ (i.e. $\Delta = 0.85$ for a sample between 242 and 50 genomes).

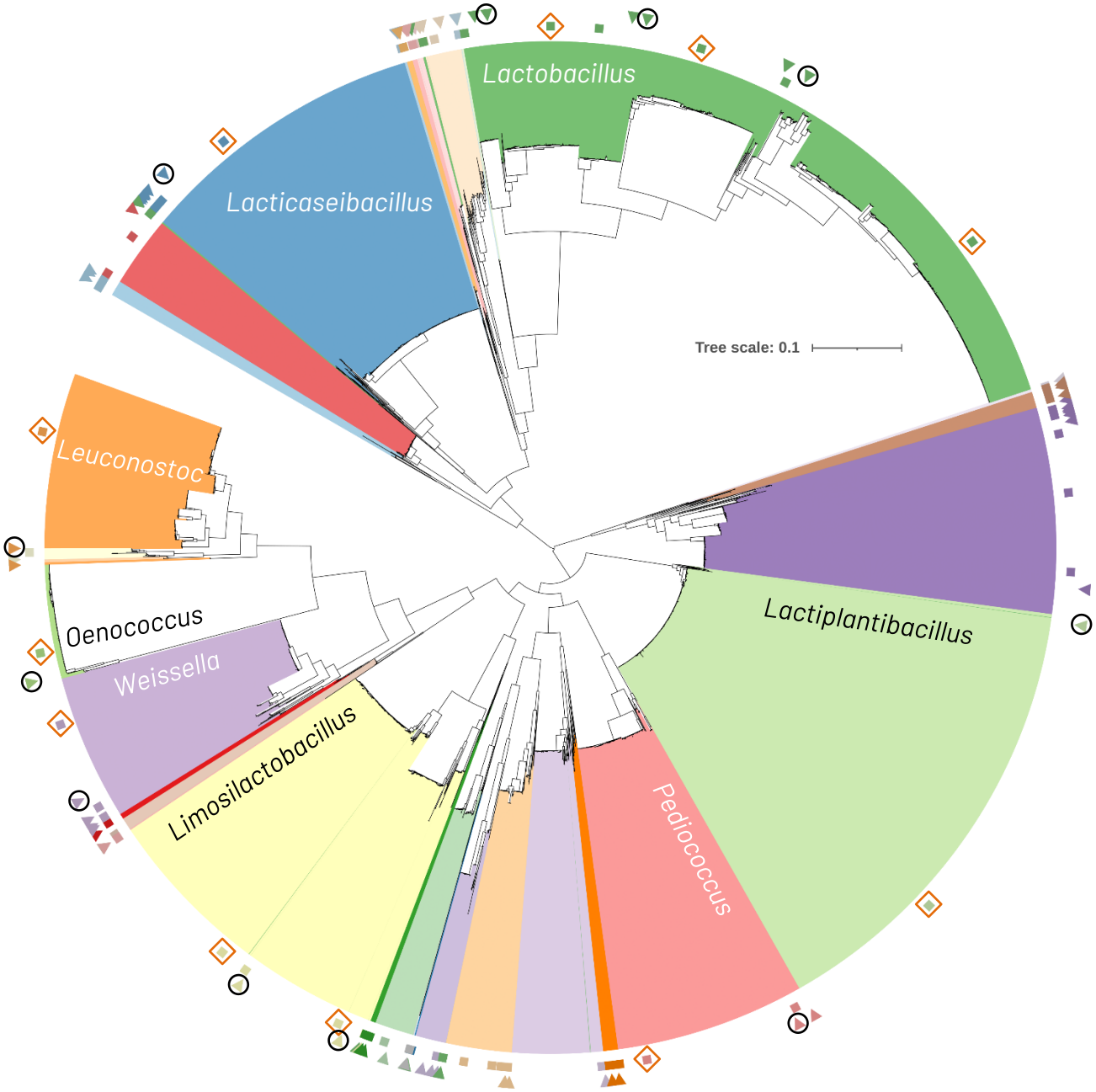
The RefSeq-representatives could be quite redundant. For example, at the top, some RefSeq-representatives may be rather redundant, because being in the same rake (black ovals). This redundancy could be dereplicated using $\Delta = 0.8$ or $\Delta = 0.7$, providing respectively 5 or 4 MPS-representatives (blue circles), corresponding to the 5 large phylogenetic clades at the top (orange rhombuses). It illustrated that the flexibility of MPS-Sampling to provide samples of variable density, while there was only one RefSeq-sample of fixed density. Furthermore, it also showed that MPS-Sampling was consistent with phylogeny while being phylogeny-agnostic.



Supplementary Material S52 – MPS-representatives and Treemmer-representatives for *Lactobacillaceae*

Among the 1,517 *Lactobacillus* genomes, MPS-representatives (squares) and Treemmer-representatives (triangles) were compared. Both samples encompassed 80 genomes, using MPS-Sampling with $\Delta = 0.7$ and Treemmer asking a sample of exactly 80 genomes.

Both samples were well-distributed along the phylogenetic tree. However, 11 Treemmer-representatives involving 8 genera were long branches (black circles), totally ignoring neighboring rakes where MPS-representatives were chosen (orange rhombuses). MPS-Sampling used centrality criteria while Treemmer used a sequential merging with random choice. After many dereplication steps, Treemmer managed to reduce rakes to single leaves, eliminating most of the redundancy. The downside was that, each rake was weighted equally as long branches: for instance, if there is a rake with 10 long branches alongside, there should be 10 chances out of 11 that a long branch should be selected, against 1 out of 11 that the rake should be kept. Treemmer is also high-sensible to long branches in another point: Treemmer uses either the number of representatives or the relative tree length compared to original tree (RTL). It implies that the inclusion of long branches within the phylogeny should disrupt the sampling of the rest of the tree.



Supplementary Material S53 – Non reproducibility of Treemmer for *Lactobacillaceae*

From the same phylogenetic tree of 1,517 *Lactobacillus* genomes, 4 replicates of Treemmer were run to generate 4 samples of 80 genomes.

First, Treemmer often chose long branches (black ovals). It happened for all the 4 replicates for 6 long branches and for few replicates only for 4 long branches. It induced that neighbooring rakes were missed by the 4 replicates (orange crosses) or retained only in few replicates (orange rhombuses). Second, the 4 samples are clearly different. Note that the chosen long branches are always different from the 4 replicates. Some rakes can be retained or completely eliminated from one replicate to another (orange rhombuses and crosses). When the phylogeny is "balanced", the retained genomes could be completely different (blue rhombuses). It demonstrated that Treemmer was not reproducible, generating quite different samplings from a replicate to another.

Moreover, Treemmer did not have any priority rules. First, Treemmer did not favour high-quality genomes from others. Second, priority rules could help Treemmer maintain the stability of the generated samples.