Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A pipeline for improved 16S rRNA gene-based phylogeny and diversity analyses of Bacteria and Archaea

Ramon Rossello-Mora

ramon@imedea.uib-csic.es

Agencia Consejo Superior de Investigaciones Científicas https://orcid.org/0000-0001-8253-3107

Tomeu Viver

Max-Planck-Institut for Marine Microbiology https://orcid.org/0000-0001-8868-9292

Pere Palmer-Rodríguez

University of the Balearic Islands https://orcid.org/0000-0001-5794-6858

Ralf Westram

Ribocon GmbH https://orcid.org/0000-0001-8243-3456

Esteban Bustos-Caparros

CSIC

Joan Francisco Gago

CSIC

Jaime Gonzalez-Mendez

CSIC https://orcid.org/0009-0000-6028-1316

Mercè Llabrés

University of the Balearic Islands https://orcid.org/0000-0002-5358-6768

Wolfgang Ludwig

Max Planck Institute for Marine Microbiology

Rudolf Amann

Max-Planck-Institut for Marine Microbiology https://orcid.org/0000-0002-0846-7372

Article

Keywords:

Posted Date: April 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9370187/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

A pipeline for improved 16S rRNA gene-based phylogeny and diversity analyses of Bacteria and Archaea

Tomeu Viver[1], Pere Palmer[2], Ralf Westram[3], Esteban Bustos-Caparros[1,4], Juan Francisco Gago[1], Jaime González-Méndez[1], Mercè Llabrés[2], Wolfgang Ludwig[5], Rudolf Amann[5], Ramon Rossello-Mora[1*]

1- Marine Microbiology Group, Department of Animal and Microbial Biodiversity, Mediterranean Institute for Advanced Studies (IMEDEA, CSIC-UIB), Esporles, Spain.
2- Department of Mathematics and Computer Science, University of the Balearic Islands, Palma 07122, Spain.
3- Ribocon GmbH, Fahrenheitstraße. 1, D-28359 Bremen, Germany
4- Nordcee, Department of Biology, University of Southern Denmark, 5230 Odense, Denmark
5- Department of Molecular Ecology, Max Planck Institute for Marine Microbiology, Bremen, Germany

Corresponding author:

*Ramon Rossello-Mora

Marine Microbiology Group, Department of Animal and Microbial Biodiversity

Mediterranean Institute for Advanced Studies (IMEDEA, CSIC-UIB)

E-07190, Esporles, Spain

Tel: +34 696488652

Email: ramon@imedea.uib-csic.es

Abstract

The LTP database compiling 16S rRNA gene sequences of all type strains of Bacteria and Archaea has served taxonomists for two decades. Here we report the LTPlus, a manually curated and aligned database expanding LTP with high quality sequences covering most of the phylogenetic diversity hitherto detected. To this purpose, the sequences were selected from the SILVA non-redundant and the GTDB databases, together with the highest quality 16S rRNA gene sequences deposited at the NCBI in the timeframe from 2019 to 2025. The LTPlus was kept compact by just integrating non-redundant sequences, applying a 98.7% identity threshold. The LTPlus was used to reconstruct a phylogenetic tree, which we believe is currently offering the most robust topology of prokaryotic lineages based on 16S rRNA gene sequences. Using the LTPlus as reference enables more accurate identification of new sequences and improved phylogenetic

inferences. LTPlus, together with the newly developed OPUcheck tool, serves as a pipeline for fast diversity analysis of environmental microbiomes based on amplicon and 16S rRNA metagenome reads.

Introduction

The 16S rRNA gene sequence encoding for the small subunit of the ribosome of Archaea and Bacteria is one of the most extensively represented genetic markers in public sequence repositories with tens of millions of deposits in GenBank[1] and its partner databases in the International Sequence Database Collaboration (INSDC)[2]. Analyses based on this marker have substantially advanced our understanding of microbial diversity, taxonomy, systematics, ecology, epidemiology, and a wide range of related fields. Since the first sequencing solely for phylogenetic analyses[3], its use quickly expanded to the discovery of new unclassified taxa, unveiling the vast diversity of uncultivated prokaryotes[4], and to the development of phylogenetic probes and strategies to quantify and assign ecological roles of different members of microbial ecosystems[5]. The relative easiness of its determination and analysis, together with declining sequencing costs promoted a fast exponential increase of the deposition of 16S rRNA gene sequences in public repositories together with an arithmetical increase of new species classifications[6] (Figure 1).

The 16S rRNA gene is central to prokaryotic systematics, diversity, and molecular ecology[4] because it is universally present and highly conserved in cellular organisms. Yet it is sufficiently variable to resolve relationships from broad taxonomic groups to species in many organisms[6]. In some cases, closely related species cannot be discerned using this sequence and only whole genome comparisons would render clear cut discriminative thresholds for taxonomic purposes[7]. For an accurate and precise analysis, the use of the almost full gene sequence is desired[6] especially because the variability of the homologous positions is not homogeneous along the gene[8]. Due to the secondary and tertiary structure that confers its functionality, along the stretch there is a succession of highly conserved to hypervariable regions with different phylogenetic information.

The relevance of this gene promoted the development of highly curated databases, often offering bioinformatic tools for analysis, and regularly updated as an important service for microbiologists. Platforms such as the Ribosomal Database Project[9] (RDP), SILVA[10], Greengenes[11], EzBioCloud[12], or LTP[13] have been key for the community for their reliable analyses in the last years. LTP RDP, EzBioCloud, and Greengenes are kept updated[13, 14, 15], and SILVA interrupted in 2019, and fortunately just recently retook their activity[16]. Among all these databases, SILVA and LTP are structured to be manipulated also using the ARB program package[17], which among the many implemented tools and features, includes a highly curated and continuously improved universal alignment[10, 13]. The latter, despite not covering intraspecies diversity, has shown to optimally perform in microbiome analyses[17]. The current difference between both databases is the length of the alignment. While SILVA still maintains an alignment with 50K positions[16], the LTP extended to 100K sites[13], facilitating the placement of indels and sequencing errors to minimize their influence on the analysis, and therefore the increasing the quality of the phylogenetic inference.

16S rRNA gene tree topologies are highly dependent on the quality of the entries and their alignment. Many of the sequences deposited in the last two decades were of low quality, short in length and with many dubious indels, generating noise making and hindering appropriate analyses[6, 10]. Amplicon surveys to determine microbial diversity in naturally occurring ecosystems are generally based on short, PCR-amplified sequences covering different hypervariable regions with different resolution power[6]. The big datasets of hundreds of thousands to millions of reads are grouped in operational taxonomic units[19, 20] (OTUs) based on a relaxed identity threshold (generally 97%), or more recently using an improved clustering version based on amplicon sequence variants[21] (ASVs). Representative sequences of each group are recovered and generally blasted against 16S rRNA gene databases to assign their identities, and in rare cases there is a previous alignment, though never manually supervised that would render slightly more precise results. Altogether, the sum of short read amplicons with imprecise reference databases may produce an important degree of identification inaccuracy that could compromise the results.

In the last 10 years, we experimented with the operational phylogenetic unit approach[21] (OPU), in where presentative sequences of clusters built with very strict thresholds were aligned and inserted in a preexisting tree, using the parsimony insertion tool implemented in the ARB program parckage which allows adding the sequences in a tree minimizing the topology changes [13, 17]. The tree was then manually supervised and each of the OPUs was manually created. The resulting OPUs represented the smallest lineages containing query and reference sequences[22]. This approach based on phylogenetic inference allowed a more precise identification, often at the species level. The OPU approach not only enables fine-tuned identification, but also benefits from the use of near full-length 16S rRNA reference sequences. This allows for comparable results regardless of the specific region selected for amplification, and even when using reads or contigs derived from metagenomic surveys targeting different regions of the gene.

In the current manuscript we present an automatization of the OPU analyses (OPUcheck) using the ARB program package, making OPU clustering faster and more objective, and thereby increasing the accuracy of the identification. Combining OPUcheck tools with the LTPlus, a compact database covering most of the genealogical diversity known, the quality of phylogenetic reconstructions and the assignment of environmental sequences are much improved. This pipeline will benefit the many disciplines depending on 16S rRNA-based analyses of microbiomes.

Results & discussion:

LTPlus: a broad, yet compact 16S rRNA gene database

The LTPlus database comprises 87,282 16S rRNA gene sequences, including ~32K from the GTDB[23], ~35K from the latest release of the SILVA database, ~1K entries corresponding to the type strains classified until the year 2025 which sequences qualified for the LTP, and the ~23K best dereplicated sequences uploaded in the NCBI database in the years 2020 to 2025 (Figure 2). As the selection covered all groups in the GTDB and 99.2% (7,890 included and 65 excluded) of the groups present in the SILVA database (see methods below) we feel confident that this selection represents the best coverage of the hitherto described diversity. As in the past[13], and when necessary, LTP type strain 16S rRNA gene sequences used by the authors in describing new

taxa were substituted by other of higher quality, mostly recovered from the type strain genome. From them 83,369 sequences affiliated with Bacteria and 3,913 with Archaea. 4,350 of the sequences corresponded to type species representative of almost all validly published genera names. 94.4% of the sequences were larger than >1400 nucleotides, and only 0.1% of the sequences were smaller than <900 nucleotides (Table 1). In accordance with the internal quality check[13] that qualifies the sequences from 0 (highest quality) to 16 the lowest quality, the mean of quality was 4.4 ±5.2 (median 2) for the complete dataset of sequences with lengths >900 nucleotides, reaching a low-quality value of 15.9 ±4.5 for sequences <900 nucleotides. When not considering the sequences <1400 nucleotides the quality values were exceptionally high with scores 3.7 ±4.5 (median 2). However, the ~4900 sequences with length <1400 nucleotides were necessarily kept as most of them were from type strains listed in the LTP and would probably be substituted in the future when their genomes are sequenced.

This ~87K sequence collection is probably the best curated dataset of publicly accessible 16S rRNA gene sequences in terms of quality and diversity spectrum. With this, we reconstructed the LTPlus tree, available in the LTPlus website (https://bioinfo.uib.es/ltp), which, given the quality of the sequences and alignment, offers the most reliable tree topology to date. We have transferred the updated taxonomy up to December 2025 that could be retrieved from the published Notification and Validation Lists in the International Journal of Systematic and Evolutionary Microbiology (IJSEM). When necessary, taxonomic accuracy was tested by crosschecking with the LPSN website[24]. For groups, for which no clear taxonomy could be assigned, we superimposed the taxonomy available in GTDB[23]. In addition to the newly reconstructed tree_LTPlus, we have updated the tree_LTP_all with the best quality type strain sequences. The tree_LTP_core remained unmodified. The tree_LTPlus currently contains 7,941 clades (7537 of Bacteria and 404 of Archaea)

In general, the tree topology agreed with the taxonomic ranks that the sequences represented. However, as it can be seen in Figure 3 (edited with the itol[25]) in some cases large taxonomic groups appeared para- or polyphyletic as for example, the *Bacillota* that appear split in two major distant branches, and the *Pseudomonadota* that appears to be in especially two major monophyletic branches one harboring most of the alphaproteobacterial taxa and the other most of the gammaproteobacterial counterparts. The branches representing the taxon hosting the type taxon (i.e. type species for genera and type genus for higher taxa) are labelled with the correct name. Those branches hosting species or higher taxa not affiliating with the corresponding type taxon, were labeled with a "~" after the name and numbered. Remarkable were cases such as the genera *Pseudomonas* with 27 clades and *Clostridium* with 33 clades or the family *Lachnospiraceae* with 32 clades. These massively para- or polyphyletic lineages require the future attention of taxonomists. The ~1420 clades named as noname~XXX corresponded to lineages that did not show known taxonomic rank assignation.

LTPlus to identify new taxa and reconstruct phylogenies

The LTPlus covers most of the taxonomic diversity yet described. It contains reference 16S rRNA gene sequences of the taxonomic groups in GTDB[23], and ~99% of the lineages of the SILVA database. It was deliberately kept very compact, and has only ~1/5 of the size of the SILVA_non_redundant dataset. It's almost complete coverage of current diversity guarantees a

reliable identification of new taxa and the affiliation of sequences to known lineages, resulting in robust phylogenetic topologies. The LTPlus website offers the database in different formats that can be used for identity inferences. We recommend the use of the ARB program package because the use of the pre-aligned dataset to align the new sequences facilitates the recognition of the homologous positions, that together with the implemented filters allows the removal of dubious bases for reliable phylogenetic inferences. In the Supplementary Text Stxt1 the reader will find our specific recommendations for accurate tree reconstructions.

LTPlus to assign sequences to Operational Phylogenetic Units (OPUs):

Diversity studies of microbiomes based on 16S rRNA gene amplicons usually cluster the sequences in groups of high identity, before analyzing selected representatives using a blast algorithm against a reference dataset. Originally, sequences were often clustered at ~97% identity, referring to the resulting groups as operational taxonomic units[19, 20] (OTUs). Current studies often use amplicon sequence variants (ASVs) for clustering, which are not based on a strict threshold, but allow broader variability in sequence identity[21], and produce more accurate results. Instead, our group has been using a different approach since more than a decade. We developed a taxonomic identification of the amplicon sequences based on a phylogenetic inference[22]. To this purpose, the reference sequences based on OTUs at 99% identity or later on ASVs, were aligned and inserted in a preexisting tree using the parsimony tool implemented in the ARB program package. The approach differs from the blast assignment because it uses the parsimony algorithm to insert each sequence in the position requiring the least number of mutations, while keeping the tree topology. The placement of the sequences within the preexisting clades permits transferring the cluster taxonomic hierarchies to the query sequences at least from the genus level and above. We always decided where to establish the limits of each clade by visually inspecting the tree topology, but the intra-clade identities were in all cases close to single species thresholds. We referred to these clusters as Operational Phylogenetic Units (OPUs). An OPU represents the smallest independent clade of inserted sequences, affiliating (in most cases) with the tree reference sequences, and separated from other similar OPUs by preexisting lineages in the tree. The approach was tedious, and arguably also subjective in defining the limits of each OPU. However, to our experience the approach was very accurate and reproducible.

The OPU approach has three major advantages: (i) It offers robust identification relying on the taxonomy of the tree clades containing the query sequences and not on the taxonomic information present in the sequence voucher. (ii) When inserting OPUs in the reference tree by the parsimony algorithm phylogenetic filters are applied keeping highly frequent homologous positions and removing extra bases from the analysis that could possibly have resulted from technical errors. This point is very important as the SINA aligner[26] allows the recognition of homologous sites within the universal alignment, placing insertions in the gaps that are latter not taken into account in the analysis with the use of the filter. Illumina sequences are generally of high quality. (iii) It is fairly independent of the length and position of the query sequences as it relies on the affiliation within a tree and not just on an identity threshold matching the closest sequence. Illumina sequences are short and therefore with low information content. Longer reads as provided by the PacBio and Nanopore technologies contain after filtering many more

informative positions. Applying OPUs, the placement within the tree is independent of the length, unless the amplicon corresponds only to highly conserved regions. This allows to use metagenomic short reads, which correspond to different regions of the 16S rRNA gene (see Supplementary Figure SF1).

To facilitate the approach, we have automatized the OPU design (Figure 4) using macros for the ARB program[17]. The OPU design starts with the import of the query sequences following their alignment and insertion into the LTPlus tree. Then, the program performs branch-length analysis and clusters the sequences in clades from 2% to 14% internal distance to further remove from the tree long branches and unclustered sequences or singletons that may result from sequence noise. In our pilot studies, a proportion ranging between ~0.3% to ~25% for amplicon sequencing and up to ~30% for metagenomic reads, but in general < 14% of the reads are discarded due to their suspicious position in the tree topology, which was reflected by the large number of misplaced nucleotides and/or the abundance of indels (Supplementary Table ST1 and supplementary Figures SF2 to SF10). Most of the discarded sequences were clearly not of 16S rRNA genes, but were either eukaryotic rRNA or even non-ribosomal sequences. However, there were still some discarded sequences that after visually inspecting the resulting tree could be reincluded in the study if relevant as we indicate in the supplementary Text Stxt2.

To evaluate the quality of the automatized OPU approach, we reanalyzed studies which had used either 16S rRNA amplicon datasets or 16S rRNA reads retrieved from metagenomes for assessing microbiome diversity. We re-evaluated both, some of our own studies[27-32] and some of colleagues[33-41]. The results given in supplementary texts Stxt3 validate our published data based on visual inspection of the tree, diminishing worries regarding potential subjectivity of manual OPU assignment. In general, most of the sequences generated by the Illumina platform qualified for the analyses, whereas the long-read platforms generated higher numbers of discarded sequences (Supplementary Table ST1). The sequences generated with the latter platforms (even the ones that were kept) often exhibited relatively long branches indicating that they contained indels could be a result from technical errors or misplaced nucleotides that modified their phylogenetic signal (supplementary Figures SF2 to SF10), but the use of conservational filters minimized the effect of the added errors (especially insertions). In general, the diversity assessment obtained by the automatized OPU inference also agreed well with the results published by our colleagues, yet the OPU-based identification is achieving species level accuracy. As indicated in Figure 4, the analysis generates a series of documents that the user can use to evaluate the results. Besides the resulting ARB document (OUTPUT.arb) with the trees and selections generated, the system exports a summary of the OPUs assignment of OTUs or ASVs (TABLE_KEPT.txt) and their taxonomy, together with the lists of the kept and discarded OTUs or ASVs, and a list of the putatively eukaryotic sequences. As indicated in the Supplementary Text Stxt2, the results are ready to be used. We still recommend, however, a visual check of the discarded OTUs or ASVs, as well on the tree topology to evaluate additional discardable sequences.

Based on the comparison of different amplicon and metagenomic 16S rRNA surveys, we conclude that almost full length 16S rRNA gene sequences generated by PacBio long read technology and evaluated with an automated OPU approach referencing to the LTPlus tree yields superior results. The increased information content together with the alignment-based filtering

misaligned nucleotides places query amplicons or metagenomic reads in highly stable, short branches indicating a high accuracy of the procedure.

Features available on the website

In the website hosting the LTPlus (https://bioinfo.uib.es/ltp) the user can download the LTPlus database in ARB format (ltplus.arb) and csv format, the LTPlus and LTP trees in newick, xfig and svg formats, the compressed and aligned databases, both in fasta format, and the related literature. There is a direct link to the OPUcheck tool for users to access to the online analyses of 16S rRNA gene amplicon or metagenomic reads or contigs. Finally, the user will also have access to a ready-to-use virtual machine that can be added to the most common operational systems (Windows, Linux and macOS) excepting the Mac computers with configuration macOS on ARM (Apple Silicon) architecture.

Final remarks

The LTPlus database, despite being highly compact, embraces most of the known diversity hitherto described. A careful selection of high-quality sequences allowed constructing a robust universal alignment that forms the basis for accurate phylogenetic reconstructions. The topology of the resulting LTPlus tree will guide taxonomic studies that may lead to reclassifications. After nearly 20 years of experience in curating the LTP database[13] and analyzing environmental sequences using the OPU approach[22, 28] we feel confident that the identification of sequences after inferring phylogenetic placements within a preexisting tree is a very precise procedure to analyze environmental diversity. The compact LTPlus database combined with the automatized OPU approach offers a fast, reliable and cost-efficient method for analyzing the diversity of microbiomes.

Methods:

Selection of sequences to be kept in the LTPlus

To construct the LTPlus database we used the four main sources of sequences: the LTP[13] database, the GTDB[23] genome compilation, the latest sequence package of SILVA_138.1_SSUREF_N99 released in June 2020[10], and sequences retrieved from NCBI deposited between June 2020 and February 2024. All sequences were evaluated using our internal quality check[13] given in the 'qual_ltp' database field that indicates how well the individual seq data fit into our current view of alignment, tree, taxonomy at the different phylogenetic (taxonomic) levels (groups). From NCBI, 368,291 16S rRNA gene sequences were downloaded, of which 295,155 ranged between 1,000 and 1,600 nucleotides in length. Also, 1,372,549 genomes had been downloaded from NCBI and extracted their 16S rRNA gene sequences using Barrnap (available at https://github.com/tseemann/barrnap), recovering 913,183 sequences with length > 1,000 nucleotides. On the one hand, we kept the ~19.8K sequences of the LTP as these represented type strain sequences previously curated and with its manually improved alignment[13]. During the evaluation of the sequence quality, many of the LTP type strain sequences referenced by the authors describing new taxa were substituted by other

of higher quality mostly recovered from the type strain genome. Only if the strain collection numbers given the entries coincided, the sequences were considered to be substitutable. Also, ~32K almost full sequences were extracted from the GTDB[23] database considered of high quality to be included in the LTPlus. Finally, all non-redundant sequences (≥99% identity) with length >1300 nucleotides were selected from the SILVA_138.1_SSUREF_N99 dataset to keep ~35.4K sequences covering almost all taxonomic groups not included in the LTP + GTDB. A total of 7,954 SILVA clades were kept in our sequence selection, and for the non-represented lineages, lower quality sequences had to be selected. According to SILVA standards the mean and median of the quality of the selected sequences were 93.4± 3.8 and 94; of the alignment were 97.2± 2.9 and 98; and that of the pintail 95.6± 14.8 and 100. Finally, from the 1,208,338 sequences with length >1000 recovered from the NCBI deposited after 2019, and clustered at 99% identity to remove redundancies we kept 23,658 of high quality. The complete dataset of 87,282 sequences covered 7,890 groups in the SILVA taxonomy and left out just 65 that were represented by the 302 sequences listed in the Supplementary Tables ST2 and ST3, all formed by sequences that did not qualify for the LTPlus.

De novo tree reconstruction and insertion of sequences of low quality

All sequences included were aligned using the latest update of the universal alignment implemented in the LTP[13] currently containing ~100K homologous positions. All sequences were manually checked and the alignments improved for reliable phylogenetic inferences. For the tree reconstruction, only seqs with the quality range (field 'qual_ltp') from 0 to 5 were used. Sequences with lower quality were inserted in the tree once ready using the parsimony tool implemented in the ARB program package. The LTPlus has also implemented filters to remove homologous positions covering distinct phylogenetic divergences, i.e. from hypervariable increasing towards more conserved homologous sites, and also for sequence stretches that should not be considered. For it we have implemented the filters:

(i) Termini, to remove the flanking regions of the gene
(ii) gap_q5_95_08may24, to remove all positions in where 95% of the high-quality sequences (quality 0 to 5) show a gap.
(iii) ssu_q5_*_08may 24, a set of 9 different positional conservational filters from 10% to 90% (* indicates the conservational percentage used), each filter removes the homologous positions with conservation equal or below the threshold.

All individual trees were generated using 'termini' and 'gap_q5_95_08may24' filters, with different conservational filter combinations, and applying each of both ARB-parsimony and FastTree algorithms. Given the large size of the dataset, maximum likelihood algorithms to reconstruct the phylogeny were not applied given the high time and computer resources consume, but these may be used in the future. Finally, a consensus tree was generated applying the ARB consensus tree tool. To the high-quality tree, all remaining sequences with quality score > 6 were inserted using the ARB-parsimony tool applying the filters 'termini' and 'gap_q5_95_08may24' but no conservation filter.

Clades and nomenclature:

After determining the best topology, the taxonomic nomenclature of the clades generated was checked, and the groups curated and formed in accordance with their taxonomic ranking. The circumscription was primarily based on the valid published taxonomic names given in the LTP and secondarily the nomenclature and ranking extracted from the GTDB taxonomy[23]. All groups with unclear taxonomic rank assignation were grouped under the denomination "noname". All groups containing a type species, were named in accordance to the valid nomenclature. Groups not enclosing a type taxon and considered polyphyletic were named with their correct taxon name, but adding an "~" to indicate polyphyly. All polyphyletic groups had been numerated after the ~ (e.g. *Pseudomonas*~1 to *Pseudomonas*~26). None with ~ contain a type species.

Automatized Operational Phylogenetic Unit (OPU) generation using amplicon sequences and ARB macros

After trimming, the amplicon sequences were grouped in different variants (ASVs) with the DADA2 software implemented in QIIME2[42]. One representative of each reference sequence is selected and used for the phylogenetic inference. To facilitate the analyses, we have developed an automatized procedure to cluster and detect OPUs (Supplementary Text Stxt4). The analyses can be done either by adding the macros to the own ARB package, install a precompiled ARB program in a virtual box (only usable in Linux, Windows environments, and for IOS versions with Intel x86-64 processors, or use the OPUcheck online tool. All this information is available in https://bioinfo.uib.es/ltp.

Data availability: most analyses shown in the supplementary material have been generated using the publicly available data from our own and other colleagues' publications and previously uploaded in the public repositories.

Code availability: scripts, precompiled programs and databases are available at the LTPlus website: https://bioinfo.uib.es/ltp

References:

[1] Sayers, E.W. et al. IGenBank 2025 update. Nuc. Acid Res. 53, D56-D61 (2025) https://doi.org/10.1093/nar/gkae1114

[2] Karsch-Mizrachi, I. The international nucleotide sequence database collaboration (INSDC): enhancing global participation. Nuc. Acid Res. 53, D62-D66 (2025) https://doi.org/10.1093/nar/gkae1058

[3] Woese, C.R. & Fox, G.E., Phylogenetic structure of the prokaryotic domain: The primary kingdoms. PNAS 74, 5088-5090 (1977) https://doi.org/10.1073/pnas.74.11.508

[4] Ludwig, W. & Klenk, H.-P. Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Boone, D.R., Castenholz, R.W., Garrity, G.M. (Eds.), Bergey's Manual of Systematic Bacteriology, seconded., Springer, New York, NY, pp. 49–65. (2001) https://doi.org/10.1002/9781118960608.bm00020

[5] Amann, R., Ludwig, W. & Schleifer, R. Phylogenetic identification and detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169 (1995) https://doi.org/10.1128/mr.59.1.143-169.1995

[6] Yarza, P., Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Revs Microbiol. 12, 635-645 (2014) https://doi.org/10.1038/nrmicro3330

[7] Stackebrandt, E. & Ebers, J. Taxonomic parameter revisited: tarnished gold standards. Microbiol. Today 33, 152–155. (2006)

[8] Andersson, A.F. et al. Comparative analysis of human gut microbiota by barcoded pyrosequencing. PlosONE 3, e2836 (2008) https://doi.org/10.1371/journal.pone.0002836

[9] Maidak B.L. The Ribosomal Database Project (RDP), *Nuc Acid Res*, 24, 82–85 (1996) https://doi.org/10.1093/nar/24.1.82

[10] Quast, C. et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nuc Acid Res, 41, D590–D596 (2013) https://doi.org/10.1093/nar/gks1219

[11] McDonald, D. et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of Bacteria and Archaea. ISME J. 6, 610–618 (2012) https://doi.org/10.1038/ismej.2011.139

[12] Yoon, S. H. et al. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int J Syst Evol Microbiol* 67, 1613-1617 (2017) https://doi.org/10.1099/ijsem.0.001755

[13] Ludwig, W. et al. Release LTP_12_2020, featuring a new ARB alignment and improved 16S rRNA tree for prokaryotic type strains. Syst. Appl. Microbiol 44, 126218. (2021) https://doi.org/10.1016/j.syapm.2021.126218

[14] McDonald, D. et al. Greengenes2 unifies microbial data in a single reference tree. *Nat. Biotechnol.* **42**, 715–718 (2024) https://doi.org/10.1038/s41587-023-01845-1

[15] Wang, Q. & Cole, J.R. Updated RDP taxonomy and RDP Classifier for more accurate taxonomic classification. Microbiol. Resour. Announc. 13, e01063-23 (2024) https://doi.org/10.1128/mra.01063-23

[16] Chuvochina, M. et al. SILVA in 2026: a global core biodata resource for rRNA within the DSMZ digital diversity. Nuc. Acid Res. 54, D334–D341. (2025) https://doi.org/10.1093/nar/gkaf1247

[17] Ludwig, W. et al., ARB: A software environment for sequence data. Nucleic Acids Research, 32, 1363–1371 (2004) https://doi.org/10.1093/nar/gkh293

[18] Hiergeist, A. et al. Reliability of species detection in 16S microbiome analysis: Comparison of five widely used pipelines and recommendations for a more standardized approach. PlosOne 18, e0280870 (2023) https://doi.org/10.1371/journal.pone.0280870

[19] Fu, L., et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012). https://doi.org/10.1093/bioinformatics/bts565

[20] Edgar RC. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinforma Oxf Engl. 34, 2371–5 (2018) https://doi.org/10.1093/bioinformatics/bty113

[21] Eren AM, et. al. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol. 4, 1111–1119 (2013) https://doi.org/10.1111/2041-210X.12114

[22] Mora-Ruiz, M. D. R. et al. Endophytic microbial diversity of the halophyte *Arthrocnemum macrostachyum* across plant compartments. FEMS Microbiology Ecology, 92, fiw145. (2016) https://doi.org/10.1093/femsec/fiw145

[23] Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 50, D785–D794 (2022). https://doi.org/10.1093/nar/gkab776

[24] Göker, M. et al. List of Recommended Names for bacteria of medical importance: report of the Ad Hoc Committee on Mitigating Changes in Prokaryotic Nomenclature. *Int J Syst Evol Microbiol*, **75**, 6943 (2025) https://doi.org/10.1099/ijsem.0.006943

[25] Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nuc. Acids Res. 52, W78–W82. (2024) https://doi.org/10.1093/nar/gkae268

[26] Pruesse, E., Peplies, J., & Glöckner, F. O. SINA: Accurate high throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics, 28, 1823–1829. (2012) https://doi.org/10.1093/bioinformatics/bts252

[27] Bustos-Caparros, E. et al. Ecological success of extreme halophiles subjected to recurrent osmotic disturbances is primarily driven by congeneric species replacement. ISME J. 18, wrae215. (2024) https://doi.org/10.1093/ismejo/wrae215

[28] Font-Verdera, F. et al. Inverted microbial community stratification and spatial–temporal stability in hypersaline anaerobic sediments from the S'Avall solar salterns. Syst. Appl. Microbiol. 44, 126231 (2021) https://doi.org/10.1016/j.syapm.2021.126231

[29] Ivulic, D. et al. Litter Management Strategies and Their Impact on the Environmental and Respiratory Microbiome Might Influence Health in Poultry. Microorganisms. 10, 878. (2022) https://doi.org/10.3390/microorganisms10050878

[30] Pomar, C.A. et al. Western diet strongly influences gut microbiomes in nursing rat dams but translates modest long-term consequences to offspring modulated by leptin supplementation during suckling. Food Frontiers. 5, 108–129. (2023) https://doi.org/10.1002/fft2.322

[31] Viver, T. et al. Food determines ephemerous and non-stable gut microbiome communities in juvenile wild and farmed Mediterranean fish. Science of the Total Environment 889, 164080 (2023) https://doi.org/10.1016/j.scitotenv.2023.164080

[32] Gago, J.F. et al. Metagenomics of two aquifers with thermal anomalies in Mallorca Island, and proposal of new uncultivated taxa named following the rules of SeqCode Syst. Appl. Microbiol. 47, 126506 (2024) https://doi.org/10.1016/j.syapm.2024.126506

[33] Anderson, M.A.J. et al. Unlocking river biofilm microbial diversity: a comparative analysis of sequencing technologies. Mol. Ecol. Res. 26, e70075. 1 of 12. (2026) https://doi.org/10.1111/1755-0998.70075

[34] Biada, I. et al. Comparative analysis of Illumina, PacBio, and nanopore for 16S rRNA gene sequencing of rabbit's gut microbiota. Front. Microbiomes. 4, 1587712. (2025) https://doi.org/10.3389/frmbi.2025.1587712

[35] Buetas, E. et al. Full-length 16S rRNA gene sequencing by PacBio improves taxonomic resolution in human microbiome samples. BMC Genomics 25, 310. (2024) https://doi.org/10.1186/s12864-024-10213-5

[36] Curry, K.D., et al. Emu: Species-level microbial community profiling for full-length Nanopore 16S reads. Nat. Methods. 19, 845-853. (2022) https://www.nature.com/articles/s41592-022-01520-4

[37] Glasl, B., et al. Co-occurring nitrifying symbiont lineages are vertically inherited and widespread in marine sponges, The ISME Journal, 18, wrae069 (2024) https://doi.org/10.1093/ismejo/wrae069

[38] Heritier-Robbins, P. et al. Beach sand oil spills select for generalist microbial populations, ISME Journal, 15, 3418–3422 (2021) https://doi.org/10.1038/s41396-021-01017-6

[39] Maritan, A.J. Sea cucumber grazing linked to enrichment of anaerobic microbial metabolisms in coral reef sediments. The ISME Journal 19, wraf088 (2025) https://doi.org/10.1093/ismejo/wraf088

[40] Miksch, S. et al. Bacterial communities in temperate and polar coastal sands are seasonally stable. ISME Comms. 1, 29 (2021) https://doi.org/10.1038/s43705-021-00028-w

[41] Veselovsky, V. et al. Comparative evaluation of sequencing platforms: Pacific Biosciences, Oxford Nanopore Technologies, and Illumina for 16S rRNA-based soil microbiome profiling. Front. Microbiol. 16, 1633360. (2025) https://doi.org/10.3389/fmicb.2025.1633360

[42] Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology, 37, 852–857. (2019) https://doi.org/10.1038/s41587-019-0209-9

Acknowledgements

The authors want to thank Conrad Schoch for his unvaluable help in retrieving the millions of sequences from the NCBI which had been deposited in the last years, as well as for his suggestions and comments on the manuscript.

Author contributions:

R.R-M., Conceptualization, Resources, Supervision, Writing – review and editing.

T.V., Conceptualization, analysis, methodology, supervision, writing – review and editing.

P.P., Software, Validation, methodology, review and editing.

R.W., Software, Validation, methodology, review and editing.

E.B.-C., data analysis, review and editing.

J.F.G., data analysis, review and editing.

J.G.M., data analysis, review and editing.

M.LL., Software, Validation, methodology, review and editing.

W.L., Conceptualization, analysis, supervision, validation, review and editing.

R.A., Resources, Supervision, Writing – review and editing.

Funding:

This study was funded by the Spanish Ministry of Science, Innovation and Universities projects PGC2018-096956-B-C41, RTC-2017-6405-1, PID2021-126114NB-C42 and PID2024-158829NB-C42 (to RRM), and PGC2018-096956-B-C44 and PID2021-126114NB-C44 (to MLL and PP), which were also supported by the European Regional Development Fund (FEDER). Further funding was supplied by the Max Planck Society. This research was carried out within the framework of the activities of the Spanish Government through the "Maria de Maeztu Centre of Excellence" accreditation to IMEDEA (CSIC-UIB) (CEX2021-001198). TV acknowledges the "Margarita Salas" postdoctoral grant, funded by the Spanish Ministry of Universities, within the framework of Recovery, Transformation and Resilience Plan, and funded by the European Union (NextGenerationEU), with the participation of the University of Balearic Islands (UIB), and also the "Vicenç Mut" postdoctoral grant founded by the Ministry of Education and Universities of the Government of the Balearic Islands, and the IMEDEA-CSIC-UIB.

Competing interests:

The authors declare no competing interests.

Additional information:

Supplementary information: The online version contains supplementary material available at XXXX

Correspondence and requests for materials should be addressed to Ramon Rossello-Mora

Tables and Figure legends:

Table 1: quality values of the LTPlus

The values in the table represent the number of sequences selected and the statistics on the sequence length and quality of the different categories all, >1400 nucleotides, <1400 nucleotides and <900 nucleotides. The quality check is based on an internal ranking from 0 to 16 the highest quality given the gene completeness and alignment fit. Out of the 87282 sequences only 1076 were ranked as 0, i.e. the complete sequence and perfect alignment fit.

		NUCLEOTIDES			QUALITY (qual_ltp field)		
	Number	Mean nucleotides	sd	Median nucleotides	Mean quality	sd	Median quality
All	87282	1491.44	84.48	1504	4.39	5.17	2
>1400 nuc	82372 (94.4%)	1503.85	56.82	1509	3.73	4.50	2
<1400 nuc	4910 (5.6%)	1282.66	162.86	1348	15.79	1.23	16
<900 nuc	105 (0.1%)	603.70	155.36	574	15.98	0.14	16

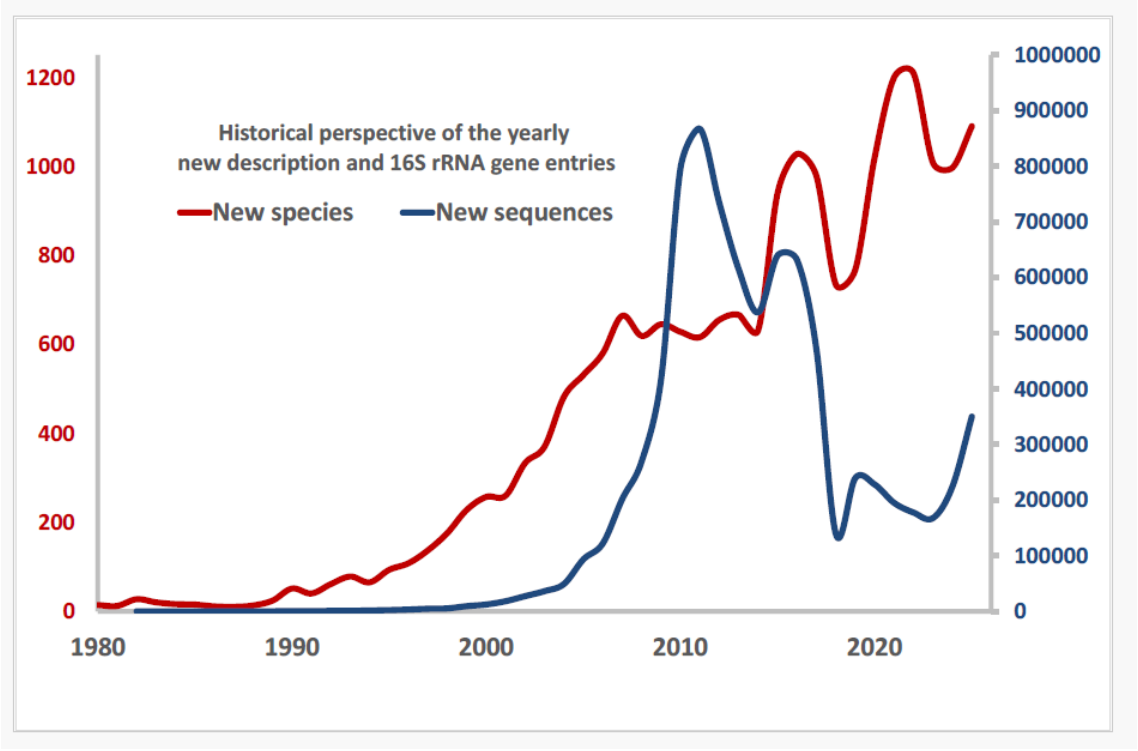

Figure 1: Chronology of the number of new species (red color and left Y axis) and new 16S rRNA gene entries (blue color and right Y axis) yearly classified or deposited (respectively) since 1980 to date.

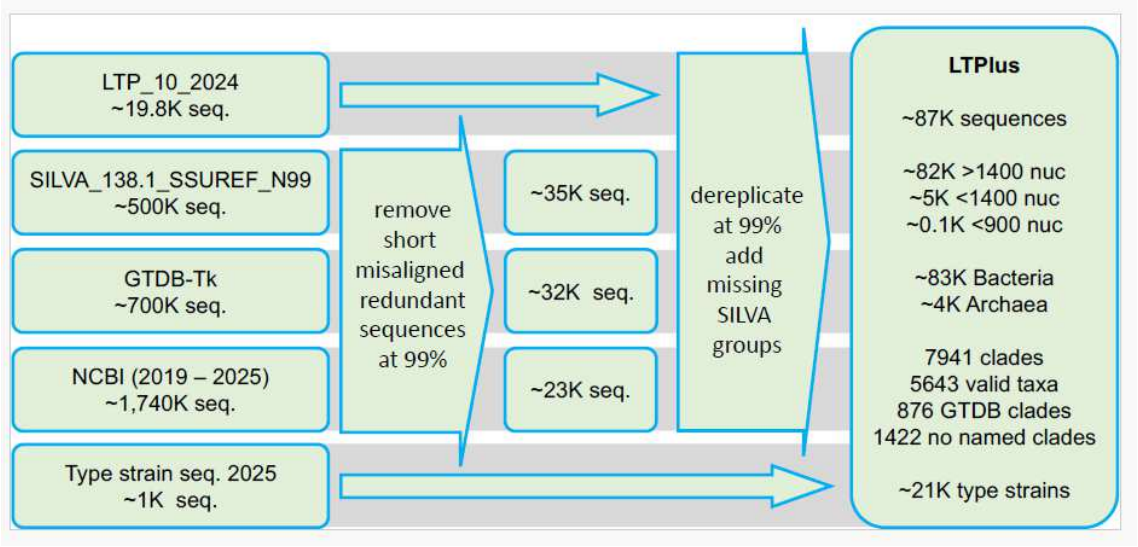

Figure 2: Sequence selection from public databases to generate the LTPlus containing non-redundant sequences with high quality standards.

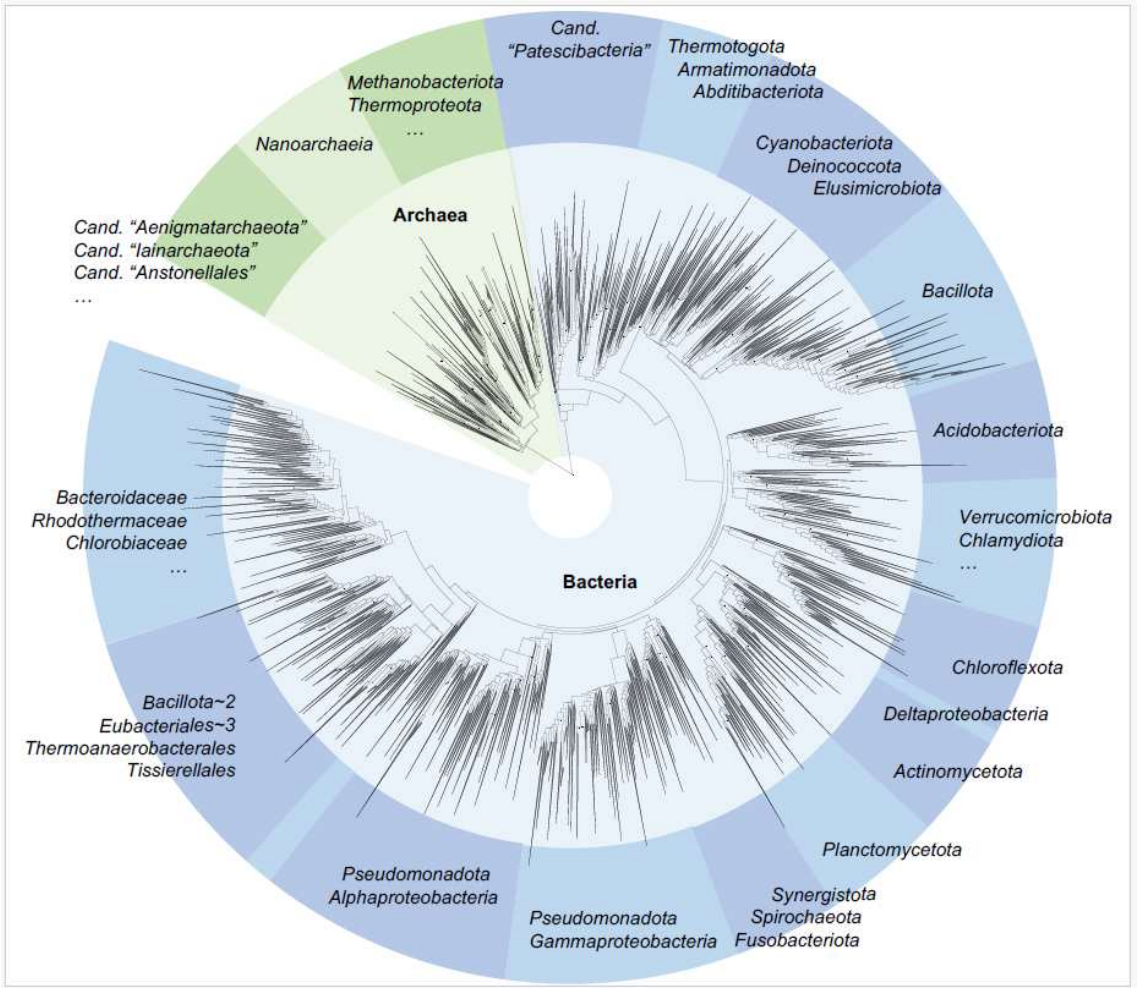

Figure 3: LTPlus reconstructed tree showing the topology distribution of the major lineages. The tree had been reconstructed using the ARB program, and the image has been improved using the itol tree display tool[25]

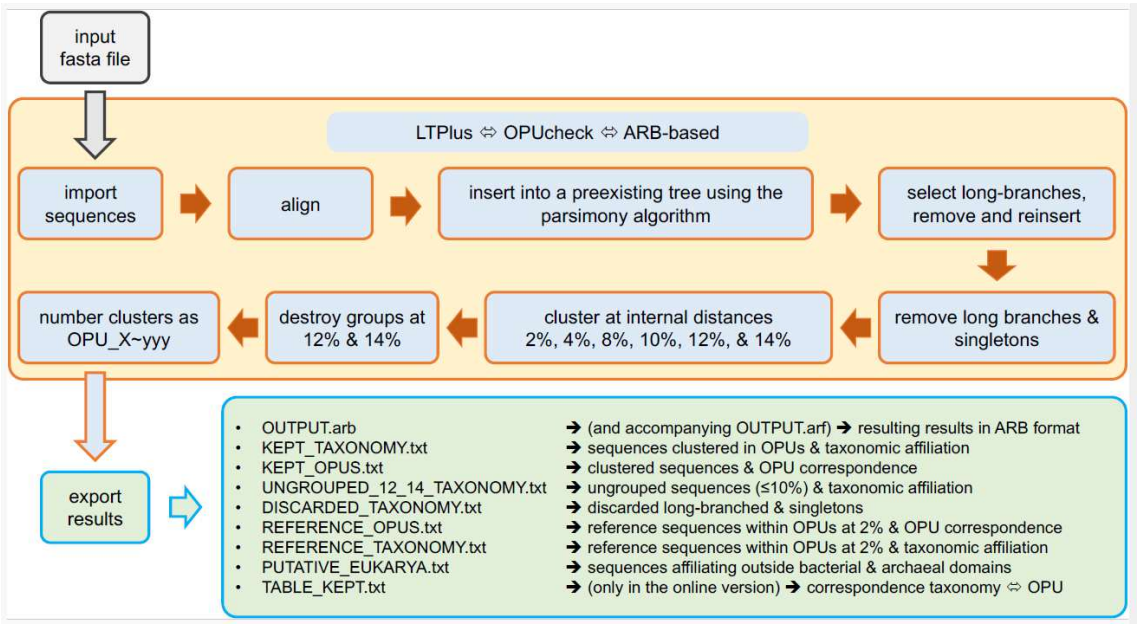

Figure 4: Procedure of OPUcheck tool implemented in the downloadable ARB package or as service in the online tool available at https://bioinfo.uib.es/ltp.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SUPPMAT.docx