Systematic and Applied Microbiology 44 (2021) 126218

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.com/locate/syapm

Release LTP_12_2020, featuring a new ARB alignment and improved 16S rRNA tree for prokaryotic type strains

Wolfgang Ludwig [a], Tomeu Viver [b], Ralf Westram [a,c], Juan Francisco Gago [b], Esteban Bustos-Caparros [b], Katrin Knittel [a], Rudolf Amann [a], Ramon Rossello-Mora [b,*]

[a] Department of Molecular Ecology, Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany
[b] Marine Microbiology Group, Department of Animal and Microbial Diversity, IMEDEA (CSIC-UIB), C/Miquel Marqués 21, 07190 Esporles, Spain
[c] Ribocon GmbH, Fahrenheitstraße. 1, D-28359 Bremen, Germany

ARTICLE INFO

Article history:
Received 16 April 2021
Revised 27 April 2021
Accepted 14 May 2021

Keywords:
16S rRNA gene
Living Tree Project
LTP
Phylogeny
Taxonomy

ABSTRACT

The new release of the All-Species Living Tree Project (LTP) represents an important step forward in the reconstruction of 16S rRNA gene phylogenies, since we not only provide an updated set of type strain sequences until December 2020, but also a series of improvements that increase the quality of the database. An improved universal alignment has been introduced that is implemented in the ARB format. In addition, all low-quality sequences present in the previous releases have been substituted by new entries with higher quality, many of them as a result of whole genome sequencing. Altogether, the improvements in the dataset and 16S rRNA sequence alignment allowed us to reconstruct robust phylogenies. The trees made available through this current LTP release feature the best topologies currently achievable. The given nomenclature and taxonomic hierarchy reflect all the changes available up to December 2020. The aim is to regularly update the validly published nomenclatural classification changes and new taxa proposals. The new release can be found at the following URL: https://imedea.uib-csic.es/mmg/ltp/.

© 2021 Elsevier GmbH. All rights reserved.

Introduction

Thirteen years with the Living tree project (LTP)

Since the 1980s, the senior member of the LTP team (WL) has been engaged on maintaining and improving a universal alignment of 16S rRNA gene sequences, implemented in the ARB program package [10]. The analyses based on the sequence comparison of 16S rRNAs have had over forty years of thorough application in prokaryotic taxonomy [22,34]. This is by far the most sequenced gene, with over 9 million entries in the public repositories (in accordance with SILVA database figures in December 2020; [18]. For decades, it has been the basis for classification of Bacteria and Archaea due to the robustness of tree topologies at various taxonomic levels ranging from genera to phyla. It was within the ARB context that the LTP and its database [30] were created as a tool to help, especially taxonomists, find the right sequence selection when reconstructing 16S rRNA-based phylogenies, and to easily establish the accurate affiliation of new taxa. This database has been regularly updated [12,28,30] by adding the sequences of newly classified species, as well as by improving the alignments, substituting low-quality sequences and through *de novo* reconstruction of trees. When the LTP was released for the first time in 2008 [30], the number of described species was just over 7,300 (as of December 31st 2007). Over the past 13 years, this number has grown arithmetically to the point where there are now ~17,000 almost full-length 16S rRNA gene sequences essentially representing all bacterial and archaeal type species (Fig. 1). Since then, the quality of many of the deposited sequences has also improved with the development of new sequencing technologies, together with international efforts to obtain the genome sequence of as many type strains as possible [11]. Validly described species represent only a minor fraction of the conservatively estimated 1 to 10 million prokaryotic species currently existing in the biosphere [31]. Nevertheless, the efforts invested in the compilation of the LTP have been rewarding. Firstly, the curation of the database allowed the recognition of those species for which the 16S rRNA gene sequence had never been determined. In addition, international collaboration involving many culture collections through the Sequencing Orphan Species (SOS) project [29] managed to (al-

* Corresponding author at: IMEDEA, C/Miquel Marqués 21, 07190 Esporles, Illes Balears, Spain.
E-mail address: rossello-mora@uib.es (R. Rossello-Mora).

https://doi.org/10.1016/j.syapm.2021.126218
0723-2020/© 2021 Elsevier GmbH. All rights reserved.

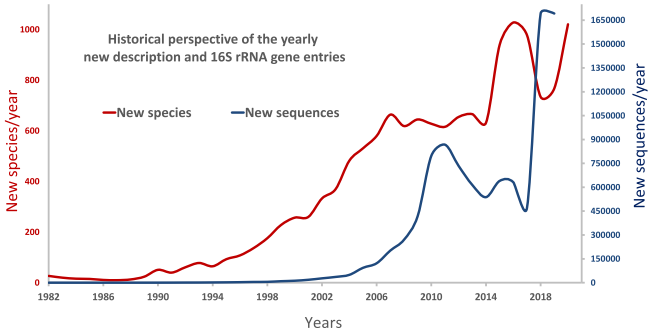

Fig. 1. Diagram showing the yearly increase in validly published new species (red line and left Y axis) and total 16S rRNA gene sequence entries (blue line and right Y axis). The values of new species descriptions have been taken from the LTP updates that have been directly generated from the notification and validation lists in the IJSEM since 1982. The 16S rRNA gene sequence entries were obtained from the SILVA releases [18] up to 2019 and, due to the delay in the last release, the values for years 2018 and 2019 represent 50% of the total biannual increase.

most) complete the catalog of 16S rRNA sequences of those species for which a viable deposit was available. Secondly, the highly curated LTP database, together with its accurate phylogenetic tree reconstructions allowed us, for the first time, to calculate the median and minimum taxonomic thresholds used by taxonomist for the classification of higher taxa. By applying these 16S rRNA sequence identity values to the sequences of as yet uncultured prokaryotes, the number of extant taxa could be estimated, and it suggested there were, for instance, more than 1,000 phyla [31]. In addition, the LTP project provided the 4th edition of The Prokaryotes [21] with harmonized phylogenetic reconstructions [12]. Due to this updated taxonomic information, the LTP has become a much-used reference database that has been of help in hundreds of new classifications (according to Google Scholar more than 1,200 papers have referred to LTP).

The 16S rRNA gene sequence is still a necessary marker for taxonomic purposes

The field of taxonomy has experienced pronounced changes during the past 13 years since the first release of the LTP. The most important changes are derived from the ease and low cost of sequencing that has allowed the generation of thousands of whole genome entries for cultivated prokaryotic strains and so-called metagenome-assembled genomes (MAGs) of as yet uncultured microorganisms. Besides the "old" 16S rRNA-based classification, a "new" genome-based classification has emerged [15] and these developments have resulted in a massive expansion of database-based taxonomies [23]. As a result, there are now several initiatives designed to re-organize the taxonomy of prokaryotes based on genome relatedness, such as the Genome Taxonomy Database (GTDB-Tk; [16]), the Microbial Genome Atlas (MiGA; [20], LINbase [27]), JSpeciesWS [19] or TrueBacID [32], among others. These initiatives provide important new insights into prokaryotic taxonomy, although their major conclusions do not always coincide. The main problem hindering the establishment of one stable taxonomic framework is that the different initiatives use different data sets to reconstruct phylogenetic trees. Initiatives like the GTDB-Tk, a tool that is widely used by the scientific community, uses 120 single-copy universal protein genes, and assigns taxonomic ranks based on the normalization of the branch lengths for the whole reconstructed tree. MiGA uses, as basic parameters, the average nucleotide identity (ANI) and the average amino acid identity (AAI), and similar approaches are used by JSpeciesWS and LINbase. In addition, there are controversial classifications that are based on the use of specially selected lineage sets of genes combined with the presence/absence of gene insertions and deletions (indels) (e.g. [5]), core gene sets combined with selected concatenates of small sets of conserved genes (e.g. [5,6]), or other practices that may be outside the bounds of the International Code of Nomenclature for Prokaryotes (ICNP; [14]) that could create confusion [24].

High-throughput genomics has forced molecular ecologists to face an important challenge in creating a new nomenclature code [13] based on DNA genome sequences as the type material (or *nomenclatural type* that is the element of the taxon with which the name is permanently associated; [14]. This alternative to the ICNP has benefits that outweigh the problems derived from not having a live organism growing as pure culture in the laboratory [9], and permits the classification of MAGs and single amplified genomes (SAGs) with the same genotypic standards as cultivation, thus expanding the taxonomic framework to the vast uncultivated majority [8]. However, we are convinced that high-quality MAGs and SAGs must contain the binned rRNA gene sequences [8] if these are used to classify new taxa. In addition, the 16S rRNA genes from MAGs, SAGs and genomes are of the utmost relevance for comparing and evaluating the great diversity that had been unveiled by the countless sequencing surveys using this gene [31]. The benefit of a 16S rRNA-based phylogenetic reconstruction using the ARB universal alignment is that it is based on a highly curated alignment that considers the secondary structure of the RNA and minimizes the effect of homoplasies and misplaced bases [10], thus, enhancing the robustness of the results. The reliability of the reconstructed trees using this gene may well surpass the genome-based phylogenies, and as these are not manually curated, homologies may be difficult to assess with large evolutionary distance, and distinct genes do not often show concordant topologies [25]. In addition, we believe that rRNA sequences will continue to be of the utmost importance in microbial ecology, since visualization, localization and quantification of Bacteria and Archaea are routinely achieved by phylogenetic probing of environmental samples with 16S rRNA-targeted oligonucleotides [2]. Fluorescence *in situ* hybridization will also add to a future largely genome-based taxonomy of as yet uncultivated prokaryotes, since it contributes to their phenotypic information, such as the cell shape, cell–cell interactions, and ecological relevance. This widely used visualization method relies on curated high-quality rRNA databases for the design of oligonucleotide probes that target monophyletic taxa with few or no outgroup hits [1].

The current project presents improvements related not only to the LTP database itself, to which we have added new sequences corresponding to the newly classified species and have improved the quality of existing entries whenever possible, but also through the incorporation of a new universal 16S rRNA sequence alignment implemented in the ARB program package [10]. These updates will result in more accurate phylogenetic reconstructions.

Results and discussion

Sequence selection substituting high-quality entries and adding newly described species

The last LTP_132 release containing 13,903 sequences was published in June 2017. The current new release was updated up to IJSEM volume 70, issue 11, which appeared in November 2020.

During the past 3.5 years, a total of 3,001 new species have been classified and, taking into account reclassifications, recognition of heterotypic synonymies and other modifications on the nomenclatural status of the published names, the current LTP now includes an additional ~3,200 sequences resulting in a total of 17,133 type strain sequences. The current taxonomic layout of the LTP accounts for 39 phyla, 99 classes, 234 orders, 582 families, 3,261 genera, 17,137 species and 431 subspecies (Supplementary Table S1). The number of genera and type species included in the LTP differs by 76 sequences from the total number validly published, since some genera lack the 16S rRNA gene sequence of the type strain of the type species, or because some species appear orphan of type species due to taxa reclassifications, as is well recorded on the LPSN website [17].

One of the major constraints in previous LTP releases was the low quality of many sequence entries that represented type strains. In many cases, these were partial 16S rRNA sequences of low quality with sequence ambiguities, homopolymers and gaps. Based on improved sequencing technologies, initiatives, such as the GEBA project [11], have generated thousands of high-quality drafted genomes of type strains, and consequently many complete high-quality rRNA gene sequences. In the course of the preparation of this release, a total of 1,775 low-quality 16S rRNA sequences were replaced (Supplementary Table S2) by higher quality sequences obtained from the same type material (Supplementary Table S3). Nomenclatural changes were also considered due to recent reclassifications (Supplementary Table S4). The average quality of sequences in accordance with an LTP-internal ranking (0 = best to 16 = worst; see below in the "New ARB Alignment" section) was thereby improved from a mean of 10.1 and a median of 9 in the former release to a mean of 8.9 and a median of 7 in the current release. In addition, approximately 3,702 entries were detected for which the strain designation and/or the indication that they represented type material was missing in the SILVA 138 release (Supplementary Table S5). The fields have been updated after manual supervision with the information obtained either for the NCBI entries where the strain designation was given in the "isolate" instead of "strain" field, or, if missing, the information was obtained from the original publications that were linked to the sequence release. In cases where the information was not obtainable, the original strain designation given by the authors and listed on the LPSN website (https://lpsn.dsmz.de/; [17] was added (Supplementary Table S5). In addition, some of the entries contained extra unnecessary texts that have been removed (Supplementary Table S6).

Changes in taxonomy

In recent years, several reclassifications have been proposed and accepted by the list editors of the IJSEM. All such changes have been manually checked and were recorded by the LTP team, such as the 770 species that changed names mostly due to reclassification at the genus level (Supplementary Table S7), or the 6,047 taxa that had been reclassified in one or several of the higher taxonomic ranks (Supplementary Table S8). In the current LTP list, all tax_ltp fields in which we spotted missing or incomplete information have also been recorded in order to offer the best corrected taxonomic path for all entries. Taxa accuracy had been cross checked with the nomenclature list available on the LPSN website [17] for years 2020–2021, and the specific and generic nomenclature changes have also been checked with the list kindly provided by the curators of the LPSN.

Among the many taxonomic changes, there is one controversial proposal to reclassify the phylum "*Tenericutes*" by creating one novel order, two novel families and five novel genera [7]. This proposal has been discussed by the respective subcommittee on the taxonomy of *Mollicutes* [3] and it has especially proposed the rejection of the reclassification of the members of the genus *Mycoplasma*. We decided to follow the recommendation of this subcommittee, and did not change the nomenclature of the proposed genera that would move some *Mycoplasma* species into newly generated genera. However, in order to record the changes in taxonomy that we did not want to highlight in the LTP release, a new field correct_name_ltp was created that records the new validly published names so the user can choose between both nomenclatures for *Mycoplasma*, as well as some *Lacticaseibacillus* and *Lapidilactobacillus*, among others, whose phylogeny shows better accordance with the names given in the fullname_lpt field rather than the considered correct names (e.g. Gupta_2018; Zheng_2018).

New ARB alignment

The previous LTP releases were based on the alignment provided by the SILVA project. However, alignments and trees are not static constructs, since they have always been further adjusted and improved according to the rapidly evolving sequencing techniques together with the fast-growing quantity of available sequence data. Any new sequence variant may contribute hitherto unknown information. The alignment of the current LTP database release comprises many more columns than that provided by SILVA, extending the 16S rRNA gene alignments from 42,283 homologous positions to 98,863, and the whole column set from 50,000 to 240,783 in order to fit all 5′ and 3′ flanking sequences occurring in entries that have not been curated by the depositors. The main reasons for this expansion are the helix extensions and intervening sequences (IVS) for which space had to be created. The insertions have often resulted from apparent sequencing artifacts, such as internal repeats, disposed symbols (bases) or improper assembly. The *Escherichia coli* sequence used as a standard [4] global consensus sequence guided the insertions of a few to several thousand columns that created a work bench for analyzing new raw data. Such additional columns were also used to place nucleotide symbols that probably represented sequencing errors.

Appropriate column filters can be applied to exclude such symbols from treeing procedures. For this purpose, in the current LTP release, filters that may help researchers to reconstruct their trees have been updated and created (Table 1). Thousands of columns also had to be added beyond the rRNA termini, given that many raw data are not trimmed to them. Certainly, the many columns that are often finally not used for analyses represent a burden when editing. However, editor tools are available to hide unwanted columns while working. All operations concerning alignment, quality checking and treeing were performed using the ARB software package. Over 30,000 sequences from the most recent decades were inspected by eye as a quality check and for alignment improvement. This has been a cyclical process of software-supported and manual interaction of data in order to check against local and global consensus sequences, while replacing apparently lower quality and partial data, and repeating the processes whenever new data had been inserted. The initial SILVA-based alignment was modified by placing the IVS and helix extensions properly. Further major rearrangements concerned variable helix regions. When evaluating potentially base-paired regions, the fit

Table 1
Filters included in the LTP_12_2020 for phylogenetic reconstruction determinations. The filters can be used depending on the level of resolution and the sequence divergence of the setup to be analyzed. The asterisk (*) indicates the number of remaining homologous positions in the alignment after applying the filters.

Filter designation	Valid columns*	Purpose of the filter
3_prime	32,273	Defines positions corresponding to the 50 3′ terminal nucleotides of the *E. coli* reference sequence
5_prime	293	Defines positions corresponding to the 50 5′ terminal nucleotides of the *E. coli* reference sequence
Termini	98,864	Defines positions corresponding to the *E. coli* reference sequence
Termini_5_3	66,298	Defines positions corresponding to the *E. coli* reference sequence, excluding the 50 5′ and 3′ terminal nucleotides, respectively
ECOLI	1,542	Number of bases of the reference sequence of *E. coli*
Gap95_q0_to_q5	94,814	Defines columns sharing gap characters in at least 95% of the LTP extended sequence selection with quality ranks 0 to 5 assigned
Rr20_q5_09jan21	1,474	Removes all positions with <20% conservation
Rr30_q5_09jan21	1,472	Removes all positions with <30% conservation
Rr40_q5_09jan21	1,423	Removes all positions with <40% conservation
Rr50_q5_09jan21	1,324	Removes all positions with <50% conservation
Rr60_q5_09jan21	1,206	Removes all positions with <60% conservation
Rr70_q5_09jan21	1,092	Removes all positions with <70% conservation
Rr80_q5_09jan21	958	Removes all positions with <80% conservation
Rr90_q5_09jan21	802	Removes all positions with <90% conservation

to a higher order consensus structure was given more emphasis than nucleotide identity, assuming that structure maintenance represents a driving force for evolution. Furthermore, in helical regions of different length, positioning of strong versus weak base pairing was used as a guide for placing nucleotide symbols in common columns. The base symbols in large helix extensions - either stable '*in vivo*' or removed during processing - were also put in separate columns dedicated to paired or non-paired positions as far as potential folding was possible and convincing. Potential terminal loop sequences were assigned to common columns. Given that similarities in helical regions with variable primary structure often do not correlate with, especially, sequence similarity, these were individually arranged according to size, similarity and loop structure by applying helix- and/or loop-specific filters that allowed consistent improvements of local alignments in particular.

Besides the alignment quality, data selection according to quality is highly important for phylogenetic analyses and taxonomic conclusions. As mentioned above, the sequences in our working databases were subjected over the decades to restrictive quality checking by applying ARB tools, as well as visual examination. Presumably erroneous or strange base symbols were identified as such in the context of 'good quality' reference sequences, and global and local consensus strings of phylogenetic neighbors. Given the alignment differences with SILVA and other databases, the documentation is not in the exact position, but rather assigned to helix and downstream regions and their numbers, respectively. A quality scoring system was established comprising 17 ranking levels from 0 (best) to 16 (worst) that was used for data selection. As described in detail below, for the 'good quality' core sequence set and core trees, only sequences belonging to ranking classes 0 to 5 were used (for both, the type strain sequence dataset given in the tree_ltp_core in the LTP release, and the set of supporting non-type strain sequences given in Supplementary Table S9). Parameters considered for establishing the quality values assigned to the individual database entries were: overall number of nucleotides (minimum 1,300 in the core region; i.e. all alignment columns except the terminal regions defined by the filters 5′-prime and 3′-prime, respectively; for level 0 to 14, any for 15, 16), ambiguities (0 for level 0, and 1 for 2, 2 for 3,. . .,10 for 11, 15 for 12, 20 for 13 and 14, any for 15 and 16), number of determined versus undetermined (missing sequence information) terminal and internal positions (i.e. undetermined internal: 0 for level 0 and 1 for 2, 2 for 3,. . .,10 for 11, 15 for 12, 20 for 13 and 14, any for 15 and 16; 5′ and 3′ terminal: missing 0 for 0, at least 1 nucleotide present for 2 to 15, no nucleotide for 16; database field: 0 for 0 and 1,1 for 2, 2 for 3,. . .,10 for 11, 15 for 12, 20 for 13, any for 14 to 16), and number of entries in the "quest_reg_ltp" database field (Supplementary Table S10; together with other new fields added). Defining the undetermined sequence (alignment) positions was performed separately for the terminal and the internal core regions. In the past, 5' and/or 3' terminal regions were often not covered by the sequencing approaches, or the sequence data provided often "faded out", so that only part of the expected sequence positions were documented, which were often scattered over the respective terminal regions. The latter finding often correlates with quality deficiencies in the core region. In a previous internal study (unpublished), it was shown that removing alignment regions corresponding to *E. coli* positions 1–50 and 1483–1532 (i.e. 50 nucleotides from both termini for comprehensive analyses) improved the tree reconstructions. Respective alignment filters (5_prime, 3_prime, and termini_5_3; Table 1) for the terminal as well as the core regions are provided with the LTP database. Admittedly, application of such restrictive quality standards comes with the risk of classifying real evolutionary changes as errors. However, one of the missions of the LTP is to provide the scientific community with a robust framework backbone and therefore we opted for excluding the impact of any questionable sequence positions. Users are free to perform their own analyses applying other criteria, but we recommend that it should be done by undertaking comparisons with different datasets and algorithms.

Unfortunately, despite the progress in sequencing techniques, as well as the accompanying improvement of data quality, the majority of type taxa sequences (~10,400 out of ~17,100) still had to be assigned to quality ranks of 6 or higher indicating the low quality of the sequences when the desirable qualities should always be <5. During the continuous database maintenance in recent years, lower quality data were replaced whenever possible. However, for some species and genera only lower quality sequences were available and the respective data had to remain included in the LTP. Consequences arising from this situation concerned phylogenetic analyses, as described in the following section. The visualization of the current rRNA-based tree

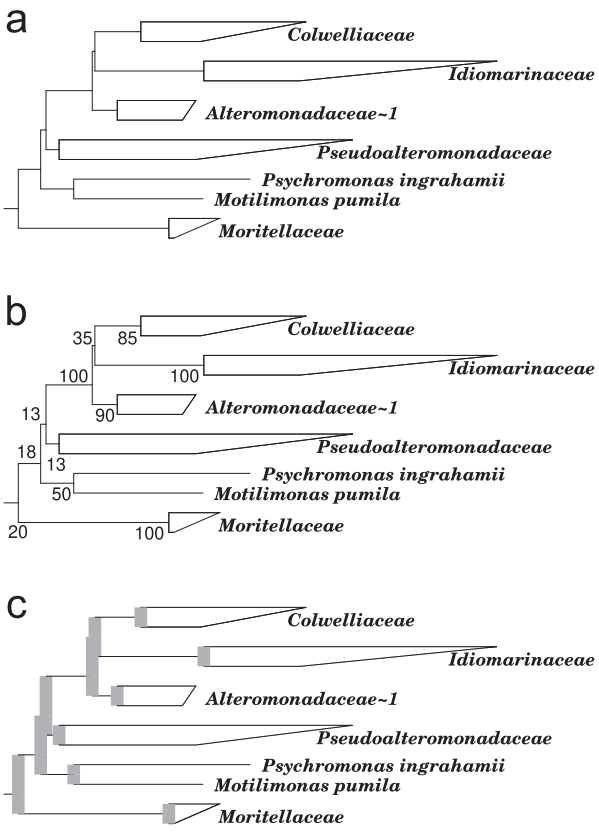

Fig. 2. Excerpt from the consensus tree_ltp_core. a: pure tree topology; b: the same tree with support values indicating the fraction (%) of individual trees generated during the multiple treeing approach described supporting the respective branching pattern; c: the same tree regions of 'unsharpness' around the branching. The (horizontal) dimensions (broadness) of grayed areas do not result from calculations but are arbitrarily defined with experience in order to provide an impression of how trees should be critically interpreted. Horizontal branch length and support values normally correlate approximately.

reconstruction is a central component of this LTP database release. The core tree ("tree_ltp_core") was only based on 6,659 high-quality sequences, assigned to quality rankings 0 to 5.

Tree reconstruction

Given the limited information content of rRNA sequences, as well as the differences and shortcomings of the commonly used treeing approaches and models of evolution, the significance of any tree reconstruction has to be seriously taken into consideration when interpreting it, especially for taxonomic purposes. Important, yet often undervalued information, can be drawn from internode distances. Whereas distance matrix procedures, such as neighbor joining, visualize the means of measured sequence differences, maximum parsimony and maximum likelihood methods provide estimated distances reflecting the significance of separation of the individual nodes by the length of the respective branches. Short internode distances indicate low significance (Fig. 2). Bootstrapping and jackknifing are commonly used tools to assign support numbers to the nodes. On the one hand, randomly deleting or selecting alignment columns only visualizes whether there are many characters common to node separating clades and not what is roughly expressed by branch lengths. On the other hand, by applying bootstrapping to the treeing methods, their model or parameters are not changed. In this project, a more directed approach proved to be an advantage.

The current LTP release presents a *de novo* reconstructed tree using the sequences of all available type strains of high quality together with a supporting set of 10,167 sequences (Supplementary Table S9) representing the widest range of lineages that constitute our current knowledge of the global diversity based on environmental data [18]. We considered this reconstruction as the core tree showing the most robust topology achievable with the current data set. It subsequently served to place all additional type strain sequences of lower quality into the best phylogenetic framework. For this reconstruction, three different basic methods were applied, which were neighbor joining, maximum parsimony and maximum likelihood, as implemented in the ARB software package. The underlying assumption is that nodes supported by more treeing methods are more likely to be correct compared to those only supported by a few. In the case of maximum likelihood, the calculations were additionally performed by applying two different models: GTRCAT and GTRGAMMA [26]. Furthermore, individual trees were reconstructed by all methods but included various alignment column selections defined by column filters calculated according to positional variability. These individual filters (Table 1) included 0%, 10%, 20%, ..., 90% conserved (identical characters) columns. They were combined with a termini filter (excluding the terminal regions corresponding to the 50 terminal positions of the *E. coli* standard). A third filter (gap95_q0_to_q5) was combined in each tree calculation that removed columns with 95% and higher gap symbols in order to prevent either the impact of potential errors positioned there (see above) or IVS not commonly present. Following this concept, 40 single trees were reconstructed based on the high-quality (ranks 0 to 5) core data set mentioned above. A consensus tree was generated based on these 40 single trees by applying the respective ARB software tool.

The current LTP release contains two trees: the **tree_ltp_all**, containing all sequences that appear in this release, and the **tree_ltp_core** that contains the best quality sequences used for the tree reconstruction, but for which the non-type strain sequences had been removed (as listed in Supplementary Table S9), together with the support values at the nodes indicating the percentage of the individual trees sharing the respective local topology. All type taxa sequences of lower quality were added stepwise to the core consensus tree using the ARB parsimony tool with presets in order to optimally place the respective sequences but exclude their impact on the topology of the core tree. These additions were performed according to the quality ranks (i.e. better-quality sequences first). The complete type taxa tree does not contain support values, given that the positioning of lower quality sequences has to be regarded as questionable anyway. When interpreting the trees, support values and internode lengths have to be taken into consideration. This also concerns the validity of taxonomic groups superimposed on the trees (see below).

Tree topology and taxonomy, incongruences, missing species and features to mention

Where supported by topology, the taxonomic groups were manually created and named. Group naming correlates with the currently valid nomenclature, as provided on the LPSN web page [17]. Care should be taken if the branch separating the respective group from the rest of the tree appears rather short. There are many cases where groups conform taxonomically but monophyly is not significantly supported. There are ~340 species that are clearly in need of being reclassified (Supplementary Table S11), since these affiliate very distantly from their type taxa, and often fall on distant branches that do not correspond to even the phylum where the nomenclatural types are located. Furthermore, a

remarkable number of taxa appear to be polyphyletic, appearing as dispersed multiple occurrences of groups sharing the same name. These polyphyletic genera and higher taxa have been listed in Supplementary Tables S12 and S13. The polyphyly of genera or higher taxa is indicated by numbers connected to the respective name: name ~ number, whereby the type taxon group remains unnumbered. On the other hand, many of the groups shown in the tree clearly comprise different taxa of the same category rank (i.e. genera within a family or species within a genus). These groups are indicated by displaying both taxon names in the tree using the " +" sign (i.e. name + name; such as *Sulfolobus* + *Stygilobus*), and up to three names are shown. Other cases, where more than three taxa of the same rank are included in the same group, are indicated by name + al (e.g. *Sediminibacterium* + al). The type taxon cluster name remains without a number, whereas other groups are numbered from 1 and so on (e.g. *Vibrio*; *Vibrio* ~ *1*; ... *Vibrio* ~ 9). Groups comprising hierarchical taxa are specified by: (lower taxon level) name_(higher taxon level) name (e.g. *Bacteroidales_Bacteroidia*). Combinations of these conventions also occur throughout the trees.

Database release and accompanying documents

The new release in ARB format can be downloaded at the following URL: https://imedea.uib-csic.es/mmg/ltp/, together with all the complementary documents. In the current 2021 release, besides the **arb** format database, we prepared a **csv** and **fasta** release, a **pdf** and **newick** document with the **tree_ltp_all**. On the website, there is an email entry list for users who wish to be updated and informed about news on the releases, have questions to be answered, or detect incongruences or failures that may need to be corrected in a subsequent release (e.g. wrong names, reclassifications not considered, etc.).

Recommendations for users

Users should be aware that the current ARB release includes a new improved universal alignment that has been manually curated after the analysis of the new sequences added from the high-quality supporting dataset, which included environmental sequences. Therefore, the SILVA database [18] still contains 16S rRNA gene sequences aligned with the former shorter alignment. Consequently, any new sequence that is not included in the LTP needs to be aligned in ARB prior to any tree reconstruction. As we aim to update the LTP database regularly on at least a bi-yearly basis, it is important that users of the LTP check for the species classified after each release. The current LTP contains sequences until December 2020, and since then to the date of submitting this manuscript approximately 147 new species have been described. Therefore, new species must be evaluated with caution if they fall within taxa that may already have newly classified species. It is recommended that the new sequences should be added to the LTP database first, then their phylogenetic position should be checked, and the literature reviewed to determine whether any new closely related taxa have been classified. Alternatively, a simple blast search in the sequence repositories, selecting only sequences from cultivated organisms, may also help to fine tune phylogenetic analyses.

As a workflow for using the LTP with the ARB program package [10], or other relevant software, it is recommended to: (i) add a new sequence to the LTP database, align it against the complete dataset and manually supervise the aligned bases; (ii) insert the sequence into a pre-existing tree (the best solution could be to use a copy of the tree_ltp_all) using the parsimony tool with the termini filters and Gap95_q0_to_q5; and (iii) evaluate the position of the newly inserted sequence. With this first approach, the user can decide the set of sequences that may be relevant for showing the phylogenetic position of the query. It might be necessary to manually supervise the aligned bases again versus this set of selected sequences.

In our experience, and to reconstruct the final tree to be used for the phylogenetic inference, we recommend: (i) reconstructing several trees with different datasets (one can expand and reduce the number of sequences in the analysis) and different filters (depending on how close or distant the query sequences are to the closest relatives) ranging from no filtering (termini and Gap95_q0_to_q5) or filtering using positional variability filters with different conservation (Table 1), and using different algorithms (our preference is always to include the maximum likelihood inference); and (ii) evaluating the different topologies obtained with the different reconstructions to show finally either the tree that reflects the majority of topologies or presents a consensus tree showing the branching orders as multifurcations that cannot be unambiguously resolved. Note that there will always be uncertain regions in the tree that cannot be resolved (Fig. 2). In addition, it should be recognized that the trees are dynamic structures that may change with the addition of new sequences, which may stabilize the branching patterns in the future.

Concluding remarks

Altogether, the trees provided in this LTP release provide a robust global phylogeny, despite the questionable positioning of the lower, especially lowest, quality sequences (included only in the tree_ltp_all). However, in the majority of cases either highly similar sequences of type strains with higher quality or selected supporting sequences have been included, which helped in stabilizing their assignment to the respective tree region. In general, most of the taxa that had been classified using 16S rRNA gene phylogenies as the backbone in order to show their uniqueness and genealogic position are phylogenetically congruent and appear in monophyletic lineages. Most of the large taxa generated prior to the use of 16S rRNA gene sequences for taxonomic purposes, showed a poly- or paraphyletic nature. This applied to genera such as *Bacillus* [6] or *Lactobacillus* [35]. Additional large polyphyletic genera are *Pseudomonas*,*Vibrio*,and *Sphingomonas*. Poly- and paraphyletic families included the *Desulfuromonadaceae*, *Myxococcaceae* and *Rhodospirillaceae*, to name just a few (see Supplementary Table S13). All of these taxa still need thorough analysis and reclassification, and it should be emphasized again that the current release contains updated classifications up to December 2020. There have been important reclassifications, such as for the classes *Deltaproteobacteria* and *Oligoflexia*, and the phylum *Thermodesulfobacteria* [33], which have been considered as controversial [24], that had still not been recorded in the LTP release, and as to the date of this submission (April 2021) these names had still not been validly published. Nevertheless, we aim to integrate the latest taxonomic changes, as well as those occurring in the near future, on at least a bi-annual basis.

Finally, the LTP team would like to encourage taxonomists to revise as many taxa as possible that are in need of reclassification. When starting such a global scientific effort our inspiration should be to create a phylogenetic classification with a stable nomenclatural framework. Taxonomists initiating such a significant endeavor might also want to perform a systematic comparison of 16S rRNA-based and whole genome trees that need to be conducted anyhow

in the near future, if only to provide a smooth transition from the old to the new taxonomical practices.

Acknowledgements

The authors would like to thank Elsevier for the continuous support to the LTP since 2008, which has granted the curator Tomeu Viver the resources for the maintenance of the database. In addition, we want to acknowledge the members of the LPSN team Markus Göker and Aidan Parte for their help with the curation of the newly classified taxa; and to the SILVA team for having hosted the LTP during the last 12 years on their webserver. This study was partially funded by the Spanish Ministry of Science, Innovation and Universities projects Salploma CLG2015_66686-C3-1-P, Micromates PGC2018-096956-B-C41 and Marbiom RTC-2017-6405-1, which were also supported with European Regional Development Fund (FEDER) funds. Funding for WL, RW and RA was provided by the Max Planck Society.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.syapm.2021.126218.

References

[1] Amann, R.I., Fuchs, B.M. (2008) Single-cell identification in microbial communities by improved fluorescence *in situ* hybridization techniques. Nat. Rev. Microbiol. 6, 339–348.
[2] Amann, R.I., Ludwig, W., Schleifer, K.-H. (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169.
[3] Balish, M., Bertaccini, A., Blanchard, A., Brown, D., Browing, G., Chalker, V., Frey, J., Gasparich, G., Hoelzle, L., Knight Jr., T., Knox, C., Kuo, C.-H., Manso-Silvan, L., May, M., Pollak, J.D., Ramirez, A.S., Spergser, J., Taylor-Robinson, D., Volokhov, D., Zhao, Y. (2019) *Mesomycoplasma* gen. nov., *Metamycoplasma* gen. nov., *Metamycoplasmataceae* fam. nov., *Mycoplasmoidaceae* fam. nov., *Mycoplasmoidales* ord. nov., *Mycoplasmoides* gen. nov., *Mycoplasmopsis* gen. nov. [Gupta, Sawnani, Adeolu, Alnajar and Oren 2018] and all proposed species comb. nov. placed therein. Int. J. Syst. Evol. Microbiol. 69, 3650–3653.
[4] Brosius, J., Dull, T.J., Sleeter, D.D., Noller, H.F. (1981) Gene organization and primary structure of a ribosomal RNA operon from *Escherichia coli*. J. Mol. Biol. 148, 107–127.
[5] Gupta, R.S., Lo, B., Son, J. (2018) Phylogenomics and comparative genomic studies robustly support division of the genus *Mycobacterium* into an emended genus *Mycobacterium* and four novel genera. Front. Microbiol. 9, 67.
[6] Gupta, R.S., Patel, S., Saini, N., Chen, S. (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 only to the members of the Subtilis and Cereus clades of species. Int. J. Syst. Evol. Microbiol. 70, 5753–5798.
[7] Gupta, R.S., Sawnani, S., Adeolu, M., Alnajar, S., Oren, A. (2018) Correction to: Phylogenetic framework for the phylum *Tenericutes* based on genome sequence data: proposal for the creation of a new order *Mycoplasmoidales* ord. nov., containing two new families *Mycoplasmoidaceae* fam. nov. and *Metamycoplasmataceae* fam. nov. harbouring *Eperythrozoon*, *Ureaplasma* and five novel genera. Ant. van Leeuwenhoek 111, 2485–2486.
[8] Konstantinidis, K.T., Rosselló-Móra, R., Amann, R. (2017) Uncultivated microbes in need of their own taxonomy. ISMEJ 11, 2399–2406.
[9] Konstantinidis, K.T., Rosselló-Móra, R., Amann, R. (2020) Advantages outweigh concerns about using genome sequence as type material for prokaryotic taxonomy. Environ. Microbiol. 22, 819–822.
[10] Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, A., Buchner, A., Lai, T., Steppi, S., Jacob, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüßmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K.H. (2004) ARB: a software environment for sequence data. Nucleic Acids Res. 32, 1363–1371.
[11] Mukherjee, S., Seshadri, R., Varghese, N.J., Eloe-Fadrosh, E.A., Meier-Kolthoff, J. P., Göker, M., Coates, R.C., Hadjithomas, M., Pavlopoulos, G.A., Paez-Espino, D., Yoshikuni, Y., Visel, A., Whitman, W.B., Garrity, G.M., Eisen, J.A., Hugenholtz, P., Pati, A., Ivanova, N.N., Woyke, T., Klenk, H.-P., Kyrpides, N.C. (2017) 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nat. Biotechnol. 35, 676–683.
[12] Munoz, R., Yarza, P., Rosselló-Móra, R. (2014) Harmonized phylogenetic trees for the prokaryotes. Pp 1-3. doi: 10.1007/978-3-642-30138-4_415. In: The Prokaryotes (Rosenberg, E., DeLong, E.F., Lory, S., Stackebrandt, E., Thompson, F.; Eds) Springer Berlin Heidelberg (ISNB: 978-3-642-30138-4).
[13] Murray, A.E., Freudenstein, J., Gribaldo, S., Hatzenpichler, R., Hedlund, B.P., Hugenholtz, P., Kämpfer, P., Konstantinidis, K.T., Lane, C.E., Papke, R.T., Parks, D. H., Reysenbach, A.-L., Rosselló-Móra, R., Stott, M.B., Sutcliffe, I.C., Thrash, J.C., Venter, S.N., Whitman, W.B., et al. (2020) Roadmap for naming uncultivated Archaea and Bacteria. Nat. Microbiol. 5, 987–994.
[14] Parker, C.T., Tindall, B.J., Garrity, G.M. (2019) International code of nomenclature of prokaryotes. Int. J. Syst. Evol. Microbiol. 69, S1–S111.
[15] Parks, D., Chuvochina, M., Chaumeil, P.A., Rinke, C., Mussig, A.J., Hugenholtz, P. (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. Nat. Biotechnol. 38, 1079–1088.
[16] Parks, D., Chuvochina, M., Waite, D.V., Rinke, C., Skarshewsky, A., Chaumeil, P. A., Hugenholtz, P. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36, 996–1004.
[17] Parte, A.C., Sardà Carbasse, J., Meier-Kolthoff, J.P., Reimer, L.C., Göker, M. (2020) List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. Int. J. Syst. Evol. Microbiol. 70, 5607–5612.
[18] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41 (D1), D590–D596.
[19] Richter, M., Rosselló-Móra, R., Glöckner, F.O., Peplies, J. (2016) JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. Bioinformatics 32, 929–931.
[20] Rodriguez-R, L.M., Gunturu, S., Harvey, W.T., Rosselló-Móra, R., Tiedje, J.M., Cole, J.R., Konstantinidis, K.-T. (2018) The microbial genomes atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nucleic Acids Res. 46 (W1), W282–W288.
[21] Rosenberg, E., De Long, E.F., Stackebrandt, E., Lory, S., Thompson, F. The Prokaryotes. Springer-Verlag, Berlin-Heidelberg.
[22] Rosselló-Móra, R., Amann, R. (2015) Past and future species definitions for Bacteria and Archaea. Syst. Appl. Microbiol. 38, 209–216.
[23] Rosselló-Móra, R. (2012) Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. Environ. Microbiol. 14, 318–334.
[24] Sanford, R.A., Lloyd, K.G., Konstantinidis, K.T., Löffler, F.E. (2021) Microbial taxonomy run amok. Trends Microbiol. 29, 394–404.
[25] Soria-Carrasco, V., Valens-Vadell, M., Peña, A., Antón, P., Amann, R., Castresana, J., Rosselló-Móra, R. (2007) Phylogenetic position of *Salinibacter ruber* based on concatenated protein alignments. Syst. Appl. Microbiol. 30, 171–179.
[26] Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30, 1312-1313.
[27] Tian, L., Huang, C., Mazloom, R., Heath, L.S., Vinatzer, B.A. (2020) LINbase: a web server for genome-based identification of prokaryotes as members of crowdsourced taxa. Nucleic Acids Res. 48 (W1), W529–W537.
[28] Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K.-H., Glöckner, F.O., Rosselló-Móra, R. (2010) Update of the All-Species Living-Tree Project based on 16S and 23S rRNA sequence analyses. Syst. Appl. Microbiol. 33, 291–299.
[29] Yarza, P., Spröer, C., Swiderski, J., Mrotzek, N., Spring, S., Tindall, B.J., Gronow, S., Pukall, R., Klenk, H.-P., Lang, E., Verbarg, S., Crouch, A., Lilburn, T., Beck, B., Unosson, C., Cardew, S., Moore, E.R.B., Gomila, M., Nakagawa, Y., Janssens, D., De Vos, P., Peiren, J., Suttels, T., Clermont, D., Bizet, C., Sakamoto, M., Iida, T., Kudo, T., Kosako, Y., Oshida, Y., Ohkuma, M., Arahal, D.R., Spieck, E., Pommerening Roeser, A., Figge, M., Park, D., Buchanan, P., Cifuentes, A., Munoz, R., Euzéby, J.P., Schleifer, K.-H., Ludwig, W., Amann, R., Glöckner, F.O., Rosselló-Móra, R. (2013) Sequencing orphan species initiative (SOS): filling the gaps in the 16S rRNA gene sequence database for all species with validly published names. Syst. Appl. Microbiol. 36, 69–73.
[30] Yarza, P., Richter, M., Peplies, J., Euzéby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F.O., Rosselló-Móra, R. (2008) The All-Species Living Tree Project: a regularly updated 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. 31, 241–250.
[31] Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.-H., Whitman, W., Euzéby, J., Amann, R., Rosselló-Móra, R. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Revs. Microbiol. 12, 635–645.
[32] Yoon, S.H., Ha, S.M., Kwon, S., Lim, J., Kim, Y., Seo, H., et al. (2017) Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int. J. Syst. Evol. Microbiol. 67, 1613–1617.
[33] Waite, D.W., Chuvochina, M., Pelikan, C., Parks, D.H., Yilmaz, P., Wagner, M., Loy, A., Naganuma, T., Nakai, R., Whitman, W.B., Hahn, M.W., Kuever, J., Hugenholtz, P. (2020) Proposal to reclassify the proteobacterial classes *Deltaproteobacteria* and *Oligoflexia*, and the phylum *Thermodesulfobacteria*

into four phyla reflecting major functional capabilities. Int. J. Syst. Evol. Microbiol. 70, 5972–6016.
[34] Woese, C.R. (1987) Bacterial evolution. Microbiol. Rev. 51, 221–271.
[35] Zheng, J., Wittouck, S., Salvetti, E., Franz, C.M.A.P., Harris, H.M.B., Mattarelli, P., O'Toole, P.W., Pot, B., Vandamme, P., Walter, J., Watanabe, K., Wuyts, S., Felis, G.E., Gänzle, M.E., Leeber, S. (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.