Journal of Biotechnology 261 (2017) 169–176

Contents lists available at ScienceDirect

Journal of Biotechnology

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

Review

25 years of serving the community with ribosomal RNA gene reference databases and tools

Frank Oliver Glöckner[a,b,*], Pelin Yilmaz[b], Christian Quast[b], Jan Gerken[a], Alan Beccati[a], Andreea Ciuprina[a], Gerrit Bruns[a], Pablo Yarza[c], Jörg Peplies[c], Ralf Westram[c], Wolfgang Ludwig[d]

[a] Department of Life Sciences and Chemistry, Jacobs University gGmbH, Bremen, Germany
[b] Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
[c] Ribocon GmbH, D-28359 Bremen, Germany
[d] Department for Microbiology, Technical University Munich, D-85354 Freising, Germany

ARTICLE INFO

Keywords:
Ribosomal RNA
Alignments
Taxonomy
Phylogeny
Biodiversity
Databases

ABSTRACT

SILVA (lat. forest) is a comprehensive web resource, providing services around up to date, high-quality datasets of aligned ribosomal RNA gene (rDNA) sequences from the Bacteria, Archaea, and Eukaryota domains. SILVA dates back to the year 1991 when Dr. Wolfgang Ludwig from the Technical University Munich started the integrated software workbench ARB (lat. tree) to support high-quality phylogenetic inference and taxonomy based on the SSU and LSU rDNA marker genes. At that time, the ARB project maintained both, the sequence reference datasets and the software package for data analysis. In 2005, with the massive increase of DNA sequence data, the maintenance of the software system ARB and the corresponding rRNA databases SILVA was split between Munich and the Microbial Genomics and Bioinformatics Research Group in Bremen. ARB has been continuously developed to include new features and improve the usability of the workbench. Thousands of users worldwide appreciate the seamless integration of common analysis tools under a central graphical user interface, in combination with its versatility.

The first SILVA release was deployed in February 2007 based on the EMBL-EBI/ENA release 89. Since then, full SILVA releases offering the database content in various flavours are published at least annually, complemented by intermediate web-releases where only the SILVA web dataset is updated. SILVA is the only rDNA database project worldwide where special emphasis is given to the consistent naming of clades of uncultivated (environmental) sequences, where no validly described cultivated representatives are available. Also exclusive for SILVA is the maintenance of both comprehensive aligned 16S/18S rDNA and 23S/28S rDNA sequence datasets. Furthermore, the SILVA alignments and trees were designed to include Eukaryota, another unique feature among rDNA databases. With the termination of the European Ribosomal RNA Database Project in 2007, the SILVA database has become the authoritative rDNA database project for Europe. The application spectrum of ARB and SILVA ranges from biodiversity analysis, medical diagnostics, to biotechnology and quality control for academia and industry.

1. Introduction

Pioneered by Fox et al. (1977) more than 40 years ago and propagated by Giovannoni et al. (1988), Olsen et al. (1986), Pace et al. (1985) and Ward et al. (1990), the use of ribosomal RNA (rRNA) molecule has become the “gold-standard” for nucleic acid-based investigations of microbial diversity, their taxonomic assignment and phylogenetic reconstructions (Amann et al., 1995; Fuhrman et al., 2015; Pace, 1997). The increasing awareness to catalogue and protect biodiversity on Earth, in combination with the advent of next generation sequencing technologies has further fuelled the interest in using the rRNA gene as a ‘barcode’ of life. The resulting millions of small and large subunit (SSU and LSU) rRNA gene sequences in the public archives require specialised tools and databases for alignment, analysis, phylogenetic inference, and classification. 25 years ago, in anticipation of the impending deluge of rDNA data, the development of the ARB software workbench and the curation of its rDNA databases was initiated (Ludwig et al., 2004). ARB offers a broad spectrum of interacting software tools built around a central database and complemented with a time-tested graphical user interface. From the

* Corresponding author at: Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany.
E-mail address: fog@mpi-bremen.de (F.O. Glöckner).

http://dx.doi.org/10.1016/j.jbiotec.2017.06.1198
Received 10 February 2017; Received in revised form 17 June 2017; Accepted 21 June 2017
Available online 23 June 2017

beginning, the ARB project provided phylogenetically structured integrative knowledge databases for small and large subunit rDNAs for Bacteria, Archaea, and Eukaryota. Until 2004, these datasets were maintained and deployed by manually collecting sequences from the International Nucleotide Sequence Database Collaboration (INSDC) (Leinonen et al., 2011), the European Ribosomal RNA Database (Wuyts et al., 2001) as well as the Ribosomal Database Project (Cole et al., 2014). Using the ARB workbench, all sequences were aligned, quality checked and used for comprehensive phylogenetic tree reconstruction.

In January 2004 Wolfgang Ludwig released the last SSU dataset comprising 59,609 sequences. Further releases were hampered by the rapid increase of sequence data, which could no longer be manually inspected and processed. In 2005, the decision was made to split the workload by separating the maintenance of the ARB software package from the production and dissemination of the rDNA datasets. This laid the foundation of the SILVA ribosomal RNA database project in the Microbial Genomics and Bioinformatics Group in Bremen. Within two years, a semi-automatic software pipeline was designed to mimic the manual curation process and amended by automated quality assessments. To compete with the growing amount of rDNA data that need to be aligned, the SILVA INcremental Aligner (SINA) was implemented and integrated (Pruesse et al., 2012). SINA is able to align millions of sequences within hours using a SEED alignment. In February 2007, SILVA released the first SSU (353,366 sequences) and LSU (46,979 sequences) datasets under the version number 89 and since then the SILVA release numbers follow the numbering of the EMBL-EBI/ENA releases.

The current SILVA database release 128 (September 2016) contains 5,616,941 SSU and 735,238 LSU rDNA sequences. All sequences are checked for anomalies and carry a rich set of sequence-associated contextual information, multiple taxonomic classifications (obtained from EMBL-EBI/ENA (INSDC) (Cochrane et al., 2012), RDP (Cole et al., 2014), Greengenes (DeSantis et al., 2006b), LTP (Yarza et al., 2008)) as well as the latest validly described nomenclature. SILVA maintains manually curated and non-public reference alignments of 75,000 16S/18S and 23S/28S ribosomal RNA genes (the SEED) in order to re-align all sequences for each SILVA release. With every full release, also a curated phylogenetic guide tree is provided that contains the latest taxonomy and nomenclature based on multiple references. The complete history of all releases is available on the SILVA website under ‘Documentation’ and ‘Archive’.

2. The ARB workbench

Powerful interoperable bioinformatics tools are inevitable for creating comprehensive multiple alignments and the reconstruction of phylogenetic trees, as well as the design of probes and primers for the in situ analysis of microbial communities. The two major objectives that were formulated at the beginning of the ARB project and were followed until today are: (1) the maintenance of a structured integrative secondary (knowledge) database of high-quality sequences, combining processed primary structures and any type of additional data assigned to the individual sequence entries, and (2) a comprehensive selection of directly interacting software tools, as well as a central database controlled via a common graphical user interface. Initially, the ARB package was designed for analysis of rDNA data only. Later, it was extended by developing and including features for managing protein sequences.

2.1. The ARB main window

The ARB main window provides the central workbench for accessing the various software tools and functions that are interacting with a local ARB sequence database. Users can select a tree representing the complete database or subsets thereof. To easily dive into the sequence space, the selected tree can be displayed in radial or dendrogram form. Any primary and metadata can be visualised at the terminal nodes.

2.2. The central database

The central component of ARB is a highly compressed hierarchical database. During operation it is loaded into the main memory (RAM) of the computer, ensuring rapid access and operation. The sequences representing genes (DNA) or gene products (rRNA or proteins) are stored in individual database fields.

2.3. Sequence editor

A powerful sequence editor facilitates user access to primary structure visualisation, arrangement, and modification (nucleotide or amino acid sequences). A special feature of the editor is the instant secondary structure check while rDNA alignments are visualised. Symbols indicating expected base pairings (or their absence) in the secondary structure of the native gene product are shown below each nucleotide symbol and constantly refreshed while editing the alignment. A three-domain consensus secondary structure mask, based on commonly accepted secondary structure models (Cannone et al., 2002), provides a guide for this tool.

2.4. Phylogenetic reconstruction

ARB hosts distance matrix, maximum likelihood, and maximum parsimony software tools for nucleotide and amino acid sequence based tree reconstruction. They directly cooperate with the respective ARB components and database elements such as alignment and filters.

ARB-parsimony has been specifically developed to handle several thousands of sequences (more than 600,000 in the current small subunit (SSU Ref NR 99) rDNA SILVA dataset (Quast et al., 2013)). New sequences are successively added to an existing tree according to the parsimony criterion. A special feature of ARB-parsimony allows adding sequences to an existing tree without altering the initial tree. This enables the user to include partial, low quality, or preliminarily aligned sequences, without disturbing the topology of an optimised tree constructed with high-quality data.

2.5. Probe design and evaluation

Gene or taxon-specific probes or primers (sequence signatures) are central for many molecular biological research and analysis projects. Examples are the delineation and identification of microorganisms in complex environmental samples (Amann et al., 1995) and the amplicon based analysis of microbial communities using next generation sequencing technologies. The ARB ‘Probe Design’ and ‘Probe Match’ tools use a suffix tree based search engine to identify short (10–100 k-mers) diagnostic sequence stretches, which are evaluated against the background of all sequences in the dataset.

2.6. Acceptance of ARB

The ARB workbench is used worldwide in academia and industry with an estimated user community of more than 10,000 users. The corresponding paper (Ludwig et al., 2004) has been cited 5431 times (Google Scholar, last assessed June 2017). The download statistics show an average combined download rate of 930 downloads for the full-text and the PDF file per year.

2.7. Future developments

Future developments of ARB are focused on the taxonomic curation of marker gene data sets. Various features are added to improve the topology comparison of phylogenetic trees and to integrate topology-based group detection and search functions. Additionally, the ARB-internal tool, which allows adding new sequences to existing phylogenetic trees (ARB Parsimony), is revised in order to build extensive trees

of improved quality. In the future, the emphasis of software development will be on performance aspects such as multi-core CPU support for selected functions.

The ARB package (Kumar et al., 2005, 2006; Ludwig and Schleifer, 2005; Ludwig et al., 2004) is freely available at http://www.arb-home.de.

2.8. Similar software tools

MEGA (Molecular Evolutionary Genetics Analysis) was developed to provide a biologist-centric, integrated suite of tools for statistical analyses of DNA and protein sequences from an evolutionary standpoint. It has grown to include tools for sequence alignment, phylogenetic tree reconstruction and visualisation, testing an array of evolutionary hypotheses, estimating sequence divergences, and web-based acquisition of sequence data. Expert systems to generate natural language descriptions of the analysis methods have been added (Tamura et al., 2011). Recently MEGA has been optimised for 64-bit computing systems to be able to analyse larger datasets (Kumar et al., 2016). MEGA is freely available for academic use at www.megasoftware.net.

PAUP* (Phylogenetic Analysis Using Parsimony) is a program for inferring phylogenetic trees and has been released as a provisional commercial version by Sinauer Associates of Sunderland, Massachusetts. It includes an integrated user interface providing access to parsimony, distance matrix and maximum likelihood methods and many indices and statistical tests. Since September 2015 Sinauer Associates is no longer distributing PAUP* but time-expiring test versions are still available. Further information about PAUP* is available at http://paup.sc.fsu.edu.

3. SILVA – high quality ribosomal RNA gene datasets

Over the last years, almost 10 million rDNA sequence data have been accumulated and released by the assembled and annotated EMBL-EBI/ENA datasets as part of the International Nucleotide Sequence Database Collaboration (INSDC) (Leinonen et al., 2011). For optimal usability, these sequences need to be checked for quality, their taxonomy and annotations must be updated to reflect current understanding. All data must be prepared in a coherent, easily accessible manner. These tasks are beyond the scope of the INSDC databases and therefore performed by knowledge (gene-specific) databases like SILVA. During its initial conception, SILVA was designed to cover the Bacteria, Archaea, and Eukaryota domains for both SSU and LSU sequences. This is a unique feature among the existing rDNA databases. All SILVA sequences are amended with a rich set of contextual information. This includes taxonomic classifications from LTP, RDP, Greengenes and INSDC, type strain information, as well as the latest valid nomenclature. All sequences are checked for quality with an intuitive representation of the quality information on the website. The corresponding datasets can be downloaded as ARB files (Ludwig et al., 2004) as well as FASTA and comma-separated value (CSV) format. Finally, they can be searched and browsed via the SILVA website. Combining the SILVA datasets with the ARB software suite provides researchers with a fine-tuned workbench for in-depth sequence analysis and phylogenetic reconstructions. The FASTA files can be used to build in-house software pipelines e.g. for the analysis of next-generation sequence data.

3.1. Sequence data retrieval

All sequence data in SILVA are retrieved from the EMBL-EBI/ENA assembled and annotated database. A combination of keywords including permutations of 16S/18S; 23S/28S; SSU; LSU; ribosomal and RNA is used to retrieve a comprehensive subset of all available small and large subunit ribosomal RNA sequences. The detection of rRNA gene candidate sequences is done using Hidden Markov model based rRNA gene prediction. The models and parameters from the RNAmmer software package (Lagesen et al., 2007) are used to scan all sequences in EMBL-EBI/ENA for the presence of rRNA gene sequences.

3.2. Alignment

All retrieved sequences are aligned using the SILVA dynamic incremental profile sequence aligner (SINA) (Pruesse et al., 2012). The aligner uses the suffix tree concept of ARB to search for up to 40 closely related sequences in the SILVA SEED-alignment. The reference sequences from the SEED are transferred into a partial order graph as used in (Lee et al., 2002), while preserving the positional identity from the reference alignment. The sequence under investigation is then aligned to this graph using a variant of the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). The SSU and LSU SEED-alignments consist of a set of 75,000 manually curated reference sequences, which are extended with every full release. 50,000 and 150,000 alignment positions for SSU and LSU, respectively, are available to host the majority of insertions across all three domains of life.

3.3. Quality control/quality assurance

The retrieved or predicted rDNA sequences are aligned to determine which parts extend beyond the boundaries of the global SILVA SSU and LSU gene alignments. The number of ambiguous bases and bases comprising long (> 4 bases) homopolymers are confined to the region within the respective rDNA. Vector contaminations are restricted to outside the rDNA boundaries of a sequence. The overall “sequence quality” value gives the averaged number to which the thresholds for each criterion were expended in percent.

Sequences are rejected based on the quality values or based on the alignment quality to achieve specificity of the SILVA datasets, thereby correcting over-prediction by the RNAmmer models as well as removing sequences wrongly annotated as rDNA. The quality thresholds for the different datasets can be found in Table 1.

Sequence anomalies are detected and flagged by using a custom version of the Pintail software (Ashelford et al., 2005). Sequences are not removed from the dataset based on this metric, due to the difficulty in unambiguously differentiating between PCR amplification artefacts and unusual yet natural evolutionary events. Only SSU sequences are checked for anomalies.

3.4. Third party contextual data

Several fields containing additional contextual data are added to the SILVA datasets. Basic fields include organism name, author, title, publication ID, collection, submission and modification dates as well as latitude/longitude, depth, habitat and taxonomic classifications by various other databases. An up-to-date description of all fields accompanies each release. Deep integration, as well as bidirectional cross-linking to BacDive (Söhngen et al., 2014), StrainInfo (Dawyndt et al., 2005) and EMBL-EBI/ENA (Leinonen et al., 2011), enables the users to seamlessly move from one database to the other.

Table 1
Alignment quality thresholds used to exclude sequences from the different SILVA datasets.

	Minimum Sequence Length	Alignment identity	Alignment score (quality)	Base pair score
SSU Parc	300 bases	50%	40	30
SSU Ref	1200 bases (Bacteria/ Eukaryota) 900 bases (Archaea)	70%	50	30
LSU Parc	300 bases	40%	30	30
LSU Ref	1900 bases	60%	30	30

3.5. Taxonomy

The bacterial and archaeal classification in SILVA is based on Bergey's Taxonomic Outlines (Boone et al., 2001; Brenner et al., 2005; Krieg et al., 2010; Vos et al., 2009). While being the most widely accepted taxonomic framework for Bacteria and Archaea, these Outlines are not updated regularly enough to keep up with the description of new taxa. Hence, any taxa that are not covered by the Outlines are adapted from a wide variety of resources, including the List of Prokaryotic Names with Standing in Nomenclature" (LPSN) (Euzeby, 1997), NCBI Taxonomy, other rRNA databases, and literature searches. New taxa are placed in the hierarchy primarily based on the suggestions of authors describing these taxa, but also taking into account where they are placed in the SILVA guide tree, hence resolving discrepancies with the aim of making the taxonomy consistent with phylogeny.

Bacterial and archaeal species and taxa are dynamic entities, and species and taxa can be merged, split or "deleted". Such changes are adapted from the LPSN resource. In addition to validly published names, LPSN is also used to track down names without standing in nomenclature (not validly published taxa) and Candidatus taxa.

In addition to this traditional taxonomic backbone, extensive effort is spent for every release to represent prominent clades known only from environmental sequences. The majority of these clades and groups are annotated in the guide tree based on literature surveys and occasionally based on personal communications; therefore, not all of these clades are available in publications. Additionally, we annotate phylogenetically coherent groups of environmental clone sequences above the family rank, based on the clone name of the earliest submitted sequence.

The numbers of bacterial and archaeal phyla are currently under a dramatic expansion (Hug et al., 2016; Rinke et al., 2013; Seitz et al., 2016; Zaremba-Niedzwiedzka et al., 2017). While most studies employ a reasonable phylogenetic reconstruction strategy to propose new phyla, some maybe a result of the enormous pressure on scientists to create the deepest taxonomic rank possible, rendering their work more dramatic. This situation is also aggravated by the limitations of 16S rRNA gene to resolve deeper evolutionary relationships amongst certain groups. As such, we employ a conservative measure of not including every single new phylum in the SILVA taxonomic framework, but only those that are vetted by other rRNA databases, as well databases employing multi-marker reconstruction approaches such as Genome Taxonomy DataBase (GTDB; http://gtdb.ecogenomic.org/about).

3.6. Data products

The SILVA datasets are made available as releases, rather than being updated continuously. This enhances the comparability of the studies employing these datasets. Each release is numbered according to the EMBL-EBI/ENA release from which the sequence data were extracted. All releases are available for download via the SILVA archive. Currently, one full and one web release are provided per year. The SILVA releases are structured into two datasets for each gene (SSU/LSU): SILVA Parc and SILVA Ref (NR). The Parc datasets comprise the entire SILVA data for the respective gene, whereas the Ref datasets represent a subset of the Parc, comprising only high-quality, nearly full-length sequences. The SSU Parc dataset is only accessible through the website and by the tools hosted on the website. Complete downloads are not provided due to the size of this dataset.

3.6.1. Description of the rDNA data products

Ref: SSU Ref comprises only of sequences with a minimum length of 900 and 1200 bases for Archaea and Bacteria/Eukaryota, respectively. The minimum alignment score is set to 50 and the minimum identity score is set to 70. Detailed criteria for SSU and LSU Parc and Ref datasets can be found in Table 1.

SSU Ref NR: The non-redundant (NR) version of the SSU Ref dataset targets users interested in a reduced, but still representative, collection of SSU rRNA gene sequences. The Ref NR dataset is created by clustering all Ref sequences at 99% sequence identity using UCLUST (Edgar, 2010). Only the longest sequence of each cluster is kept. Type strains and cultivated species, as well as multiple operons, are preserved in all cases to serve as a taxonomic anchor. The SSU Ref NR dataset is around 35% smaller than the Ref dataset and has a more even phylogenetic distribution of sequences. SSU Ref NR is recommended as the standard SILVA reference dataset for rRNA gene-based classification, phylogenetic reconstruction and probe design.

LTP: The "All-Species" Living Tree Project (LTP) is a collaborative initiative of the ARB, LPSN, and SILVA projects, coordinated by the journal Systematic and Applied Microbiology (Elsevier publisher). It provides highly curated ribosomal 16S and 23S RNA gene datasets of all type strains with validly published names (Yarza et al., 2008). The current LTP release 128/123 contains 12,955 SSU entries and 1614 LSU entries. This rather small but taxonomically "comprehensive" dataset is frequently used for taxonomic and classification purposes and as a test dataset for developers.

3.7. Dynamic tree viewer

The SILVA Tree Viewer is a web application designed to visualise and to interact with large phylogenetic trees without the need to download any software tool or data files. It enables direct access to the SSU Ref NR 99 and LSU Ref phylogenetic trees provided by the SILVA database. The viewer provides tree navigation, search and browse tools, and an interactive feedback system to interactively guide data curation.

3.8. SILVA website

The SILVA website provides database access, several online tools (TestProbe, TestPrime, sequence based search, online aligner & classifier (SINA)) and regularly updated documentation pages. The latter provides tutorials for all SILVA tools and functionalities, FAQs and details for each database release. Partner projects and collaborations such as LTP are shown as well.

The SILVA website and databases are available at https://www.arb-silva.de.

3.9. Usage of SILVA

The two main SILVA papers have been cited 5450 times according to Google Scholar (last assessed June 2017). According to Scopus (last assessed January 2017), the top 5 subject areas of the papers citing SILVA are 'agricultural and biological sciences', 'immunology and microbiology', 'biochemistry, genetics and molecular biology' and 'medicine and environmental sciences', which once more reflects the broad field of application of rDNA technologies. The SILVA web page handles around 85,000 users (164,000 sessions, 747,000 page views) from 186 countries per year (monitored by Google Analytics, accessed June 2017). The majority is coming from Europe (38%), USA (33%) and Asia (24%). In the last four years, on average, more than 80,000 jobs per year have been processed by the SILVA website. The datasets have been downloaded around 30,000 times per year.

The following high-impact projects and tools rely on the SILVA database and act at the same time as multiplicators.

3.9.1. Metagenomics

EBI Metagenomics (Mitchell et al., 2016) and Metagenomics-RAST Server (Meyer et al., 2008) use SILVA as their default reference dataset for the classification of ribosomal RNA gene reads in metagenomes.

3.9.2. Taxonomy

The UniEuk project (http://unieuk.org/) funded by the Gordon and Betty Moore Foundation aims to provide a universal taxonomic

framework and integrated reference gene databases for Eukaryotic biology, ecology, and evolution. The results, being a consolidation of major experts' advice, will help in further improving the Eukaryotic taxonomic framework in SILVA.

3.9.3. Database

RNAcentral is a database of non-coding RNA (ncRNA) sequences that aggregates data from specialised ncRNA resources and provides a single entry point for accessing ncRNA sequences of all ncRNA types from all organisms. SILVA is one of the resources used by this database (The RNAcentral Consortium, 2015).

3.9.4. Tools

Mothur is a comprehensive software package that allows users to analyse community sequence data. Mothur can be used to trim, screen and align sequences, calculate distances, assign sequences to operational taxonomic units and describe the α- and β-diversity based on 16S rDNA sequences. Mothur uses SILVA as its default reference dataset (Schloss et al., 2009).

QIIME is an open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data. This includes demultiplexing and quality filtering, OTU picking, taxonomic assignment, and phylogenetic reconstruction, diversity analyses and visualisations. QIIME uses SILVA as one of its reference datasets (Caporaso et al., 2010).

MEGAN allows laptop analysis of large metagenomic data sets. In a pre-processing step, the set of DNA sequences is compared against datasets of known sequences using BLAST or another comparison tool. The software allows large data sets to be dissected without the need for assembly or targeting of specific phylogenetic markers. It provides graphical and statistical output for comparing different data sets (Huson et al., 2007). Megan uses SILVA as one of its reference datasets (Mitra et al., 2011).

3.9.5. Non-academic usage

With respect to the usage of SILVA in commercial environments, a number of licenses have been filed with the company Ribocon GmbH, which takes care about the commercial usage of SILVA.

An overview of the application spectrum of SILVA in molecular diversity analysis is given in Fig. 1.

3.10. Similar resources

Worldwide, only two similar databases exist, both located in the USA: Greengenes and the Ribosomal Database Project (RDP).

The Greengenes rDNA database (DeSantis et al., 2006b) was initially hosted by the Lawrence Berkeley National Laboratory and later on by the company Second Genome. Greengenes provides 16S rDNA sequences from Bacteria and Archaea but lacks any LSU or Eukaryotic sequences. The resource seems to be discontinued, the last update dates back to 2013 (http://greengenes.secondgenome.com/downloads).

The Ribosomal RNA Database project (Cole et al., 2014), hosted by the Michigan State University, provides 16S rDNA sequences for Bacteria and Archaea, as well as 28S sequences for Fungi. The resource is updated at least once a year but lacks 18S, 23S as well as any 28S rDNAs except for Fungi.

With respect to the creation and maintenance of the underlying alignment of the two databases, it should be noted that Greengenes follows a similar strategy as SILVA by incrementally aligning new sequences with NAST (DeSantis et al., 2006a) using a SEED alignment. Compared to SILVA, the alignment contains only 7682 positions (SILVA has 50,000), which exclude the incorporation of Eukaryota. This number of alignment positions is also low for Bacteria and Archaea, which can lead to suboptimal positioning of insertions, or duplicated bases.

RDP uses the INFERNAL tool to align all rDNA sequences (Nawrocki and Eddy, 2013). INFERNAL builds consensus RNA secondary structure profiles called covariance models (CMs) and uses them to create new sequence- and structure-based multiple sequence alignments. The bacterial and archaeal aligners in RDP were trained using secondary structure information from the Comparative RNA Web Site (Cannone et al., 2002; Cole et al., 2014). Similar to SILVA, the overall number of positions in the alignment is 50,000. Due to the limited amount of INFERNAL models, variable regions are often not aligned. Lower case letters in the output indicate such regions. If new models are included, the length and the structure of the RDP alignment changes, forcing users to realign their sequences. The quality of the alignments in SILVA, Greengenes and RDP has been evaluated by Patrick D. Schloss and the results show that SILVA outcompetes both Greengenes and RDP (Schloss, 2009). A recent comparison of the taxonomies between SILVA, RDP, Greengenes, INSDC and OTT (Open Tree of life Taxonomy) attests that SILVA has a richer taxonomy than RDP and Greengenes when compared to INSDC and OTT (Balvočiūtė and Huson, 2017).

3.11. Latest and future developments

3.11.1. Candidate taxonomic units (CTUs) for uncultivated clades

The number of environmental rRNA gene sequences has surpassed the number of sequences from cultivated microorganisms by far. Hence, reconciliation of the established taxonomy and a classification of the uncultured fraction have become critical. The Bacteriological Code does not cover the classification of uncultured Archaea and Bacteria, and only a "Candidatus" status can be given to those uncultured organisms, leaving no possibility to generate a unified nomenclature for both cultured and uncultured microorganisms. The CTU method offers a simple procedure to devise phylogeny-aware taxonomies by overlaying a phylogenetic tree with OTU (Operational Taxonomic Unit) information (Yarza et al., 2014). A number of SILVA users have expressed their enthusiasm for seeing the CTU concept implemented for at least the so-called Candidate divisions. In this light, we are planning a re-classification of phylum level candidate divisions using the CTU thresholds with one of the next SILVA releases. This will introduce a more rational taxonomic resolution to these divisions, which appear very often in environmental sequencing surveys.

3.11.2. Calculation of phylogenetic trees

Since the start of the SILVA project, the guide trees have been based on phylogenetic trees that were available in the latest ARB database release by Wolfgang Ludwig in 2004. With every release of SILVA, new sequences have been added to these trees using ARB's 'quick add to tree' feature which uses the parsimony approach to place new sequences in an existing tree without changing its topology. Over time, more than ten times the number of initial sequences has been added to the SSU guide tree. In rare cases, sequences have been placed in the wrong vicinity, which subsequently attracted closely related sequences. Sequences belonging to the same phylogenetic group were thus split across multiple locations in the tree, giving the sequences different taxonomic assignments. For classification tools this poses a substantial problem, as unknown reads cannot unambiguously be classified to the genus; a higher level classification is created instead.

To solve the problem of non-monophyletic phylogenetic groups and to increase the overall quality of the SILVA phylogenetic guide trees, de novo trees have to be calculated. The challenge when calculating these trees is the large number of sequences they comprise, with more than six hundred thousand in the case of the SSU tree. It is planned to recalculate both guide trees from scratch, using RAxML (Stamatakis et al., 2005), for one of the upcoming releases.

After the new trees have been calculated, phylogenetic groups have to be defined automatically and taxonomic information needs to be transferred from the previous trees to the newly created ones. This on

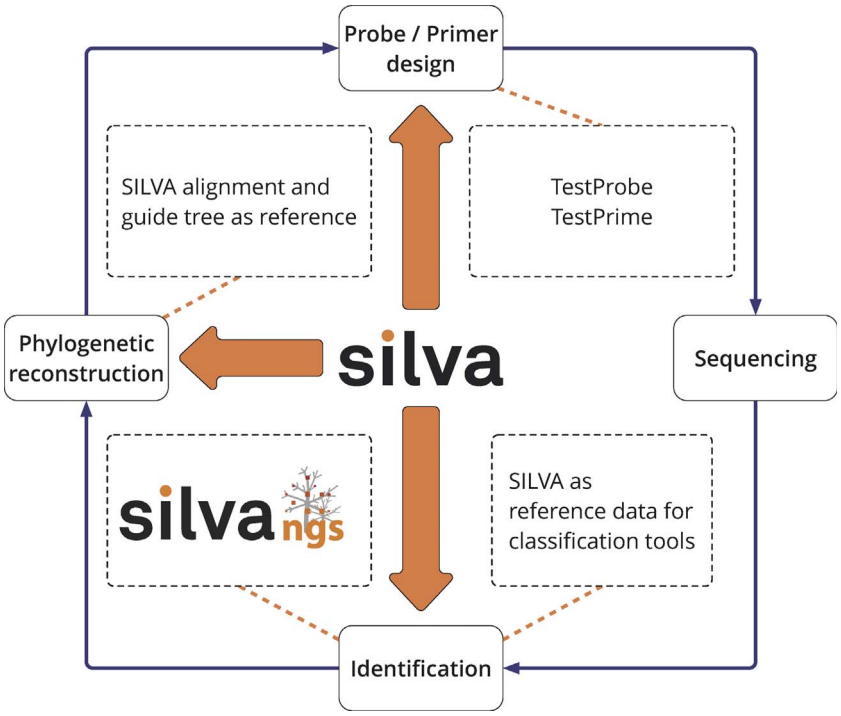

Fig. 1. Overview of the application of SILVA in microbial ecology, identification and taxonomy.

its own presents a challenge for which a novel solution must be implemented, due to the number of groups annotated in each tree. This affects more than twelve thousand groups in case of the SSU tree and more than three thousand for the LSU tree, respectively.

It is planned to make the tree calculation, including the detection and annotation of groups, available as a tool hosted on the SILVA website after it has been implemented and successfully used to calculate the new phylogenetic guide trees.

3.11.3. Redesign of the SILVA website

When the SILVA website first went online in early 2007 it had a fresh and coherent design. Since then, new features have been added to the website without ever adapting the website to current web standards. Within the next years, the SILVA website will be redesigned using state of the art web technologies with focus on responsiveness (scaling to different form factors), accessibility, and more responsive (faster user feedback) interface for the tools hosted on the SILVA website. Additionally, a REST service will be implemented to provide programmatic access to the data and tools of SILVA website.

We would like to note that the ongoing and future developments were inspired by SILVA users worldwide but could only be realised with the support of the German Network for Bioinformatics Infrastructure de.NBI. The long-term perspective of SILVA is currently bound to the availability of project funding. With Germany (de.NBI) having joined the European ELIXIR infrastructure in 2016, we hope that this will also provide a more stable perspective for the further maintenance and development of the SILVA databases and services.

4. SILVAngs

SILVAngs is a web-based service that satisfies the user need for a fast and accurate classification system of rDNA amplicon reads from high-throughput next generation sequencing (NGS) technologies. Compared to commonly used stand-alone solutions, like Mothur or QIIME, SILVAngs represents a centralized, online platform targeting (a) comparability of results among studies by centralized and standardized data analysis, (b) ease of use through an intuitive web interface, and (c) to save the users from complex installation procedures and high computational demands.

SILVAngs accepts short- and long read sequence data in Multi-FASTA format and performs quality control, alignment, and taxonomic classification of the ribosomal RNA gene sequences based on the curated SILVA reference taxonomy. All steps (upload, progress monitoring, visualisation of results, and download of data) can be controlled by the SILVAngs web-interface. Several samples can be combined in a project for comparative analysis.

In the backbone, the pipeline runs quality control, replicate identification, clustering (OTU definition), classification modules and the preparation of rarefaction curves on a compute cluster. All processes and results are handled by a database management system, instead of sequential file processing. For each project, an overview report (HTML and PDF) as well as a comprehensive set of results for download (ZIP archive) are automatically generated. These include detailed statistics, taxonomic breakdowns (Fig. 2) and sequence exports for subsequent in-depth data mining. The system is access controlled with the possibility to share projects between users.

In the framework of the German Network for Bioinformatics Infrastructure project (de.NBI), additional computational resources were made available through the network. This now allows running large scale projects (up to 100 million reads) on user request.

SILVAngs can be accessed at https://www.arb-silva.de/ngs.

5. User support & training

The SILVA team provides support via the e-mail address contact@arb-silva.de. A ticket system has been implemented as part of the de.NBI quality management. Based on our internal code of conduct, the users can expect an answer within 24 h. Additionally, an ARB-SILVA user group exists at Yahoo with currently 1200 subscribed members.

Training workshops are conducted on an annual basis in Bremen targeting ARB/SILVA, as well as NGS analysis of amplicons. Additionally, on-site workshops are provided to academia and industry by Ribocon GmbH Bremen, Germany.

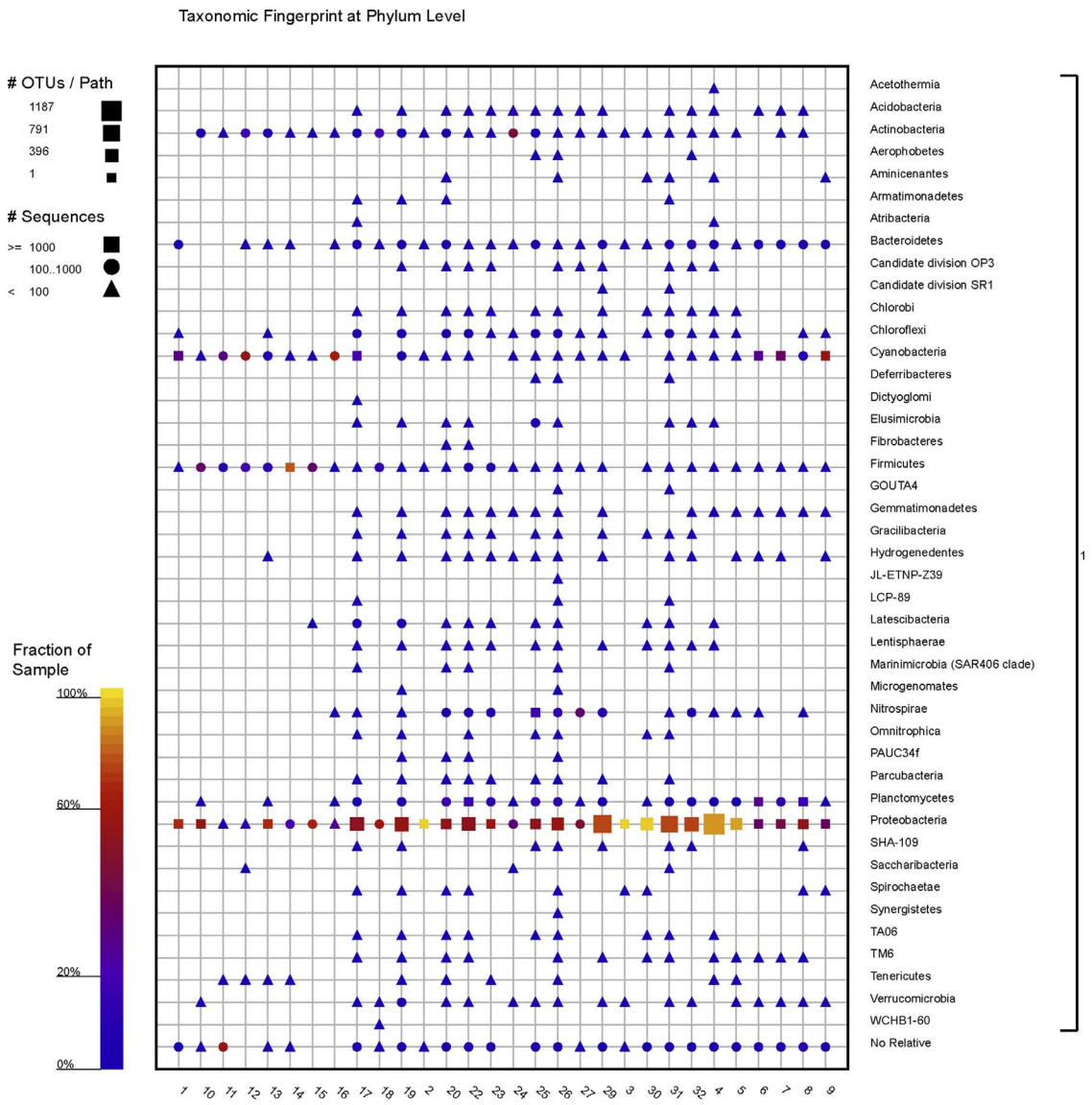

Fig. 2. Taxonomic fingerprint on phylum level indicating the number of sequences as well as the OTU 'diversity' per group.

Acknowledgements

We would like to thank all ARB and SILVA users for their support and feedback on the ARB software suite and SILVA databases and services. Your immediate critical evaluation is a great motivation for us to improve our databases and tools. We would also like to thank the LTP, RDP, StrainInfo, probeBase, BacDive, and LPSN teams for their support and fruitful discussions. Special thanks go to the whole de.NBI network, where foreigners turned into friends inspired to serve the scientific community at its best. We are grateful for funding from the Max Planck Society, the German Research Foundation grant GL 553/5-1, the Moore Foundation grant 5257, as well as the Federal Ministry of Education and Research grant 031A539A.

References

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.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J., Weightman, A.J., 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl. Environ. Microbiol. 71, 7724–7736.

Balvočiūtė, M., Huson, D.H., 2017. SILVA, RDP, greengenes, NCBI and OTT—how do these taxonomies compare? BMC Genom. 18, 114.

Boone, D.R., Castenholz, R.W., Garrity, G.M., Staley, J.T., 2001. The Archaea and the Deeply Branching and Phototrophic Bacteria. Springer, New York.

Brenner, D.J., Krieg, N.R., Garrity, G.M., Staley, J.T., 2005. The Proteobacteria. Springer, New York.

Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Müller, K.M., Pande, N., Shang, Z., Yu, N., Gutell, R.R., 2002. The comparative RNA Web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinform. 3, 2.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Tumbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R., 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335–336.

Cochrane, G., Cook, C., Birney, E., 2012. The future of DNA sequence archiving. GigaScience 1, 2.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., Tiedje, J.M., 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acid Res. 42, D633–D642.

Dawyndt, P., Vancanneyt, M., De Meyer, H., Swings, J., 2005. Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources. IEEE Trans. Knowl. Data Eng. 17, 1111–1126.

DeSantis, T.Z., Hugenholtz, P., Keller, K., Brodie, E.L., Larsen, N., Piceno, Y.M., Phan, R., Andersen, G.L., 2006a. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acid Res. 34, W394–W399.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L., 2006b. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72, 5069–5072.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461.
Euzeby, J.P., 1997. List of bacterial names with standing in nomenclature: a folder available on the Internet. Int. J. Syst. Bacteriol. 47, 590–592.
Fox, G.E., Pechman, K.R., Woese, C.R., 1977. Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. Int. J. Bacteriol. 27, 44–57.
Fuhrman, J.A., Cram, J.A., Needham, D.M., 2015. Marine microbial community dynamics and their ecological interpretation. Nat. Rev. Microbiol. 13, 133–146.
Giovannoni, S.J., DeLong, E.F., Olsen, G.J., Pace, N.R., 1988. Phylogenetic groupspecific oligodeoxynucleotide probes for identification of single microbial cells. J. Bacteriol. 170, 720–726.
Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D.A., Finstad, K.M., Amundson, R., Thomas, B.C., Banfield, J.F., 2016. A new view of the tree of life. Nat. Microbiol. 1, 16048.
Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C., 2007. MEGAN analysis of metagenomic data. Genome Res. 17, 377–386.
Krieg, N.R., Staley, J.T., Brown, D.R., Hedlund, B.P., Paster, B.J., Ward, N.L., Ludwig, W., Whitman, W.B., 2010. The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes. Springer, New York.
Kumar, Y., Westram, R., Behrens, S., Fuchs, B., Glöckner, F.O., Amann, R., Meier, H., Ludwig, W., 2005. Graphical representation of ribosomal RNA probe accessibility data using ARB software package. BMC Bioinform. 6, 61.
Kumar, Y., Westram, R., Kipfer, P., Meier, H., Ludwig, W., 2006. Evaluation of sequence alignments and oligonucleotide probes with respect to three-dimensional structure of ribosomal RNA using ARB software package. Bmc Bioinform. 7.
Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.
Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., Ussery, D.W., 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acid Res. 35, 3100–3108.
Lee, C., Grasso, C., Sharlow, M.F., 2002. Multiple sequence alignment using partial order graphs. Bioinformatics 18, 452–464.
Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tarraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V., Cochrane, G., 2011. The European nucleotide archive. Nucleic Acid Res. 39, D28–D31.
Ludwig, W., Schleifer, K.H., 2005. Molecular phylogeny of bacteria based on comparative sequence analysis of conserved genes. In: Sapp, J. (Ed.), Microbial Phylogeny and Evolution, Concepts and Controversies. Oxford university press, New York, pp. 70–98.
Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar Buchner, A., Lai, T., Steppi, S., Jobb, G., Forster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., Konig, A., Liss, T., Lussmann, 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 Acid Res. 32, 1363–1371.
Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., Edwards, R.A., 2008. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinform. 9, 386.
Mitchell, A., Bucchini, F., Cochrane, G., Denise, H., Hoopen P. t. Fraser, M., Pesseat, S., Potter, S., Scheremetjew, M., Sterk, P., Finn, R.D., 2016. EBI metagenomics in 2016—an expanding and evolving resource for the analysis and archiving of metagenomic data. Nucleic Acid Res. 44, D595–D603.
Mitra, S., Stärk, M., Huson, D.H., 2011. Analysis of 16S rRNA environmental sequences using MEGAN. BMC Genom. 12, S17.
Nawrocki, E.P., Eddy, S.R., 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29, 2933–2935.
Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to search for similarities in amino acid sequence of 2 proteins. J. Mol. Biol. 48, 443.
Olsen, G.J., Lane, D.J., Giovannoni, S.J., Pace, N.R., Stahl, D.A., 1986. Microbial ecology and evolution: a ribosomal RNA approach. Annu. Rev. Microbiol. 40, 337–365.
Pace, N.R., Stahl, D.A., Olsen, G.J., Lane, D.J., 1985. Analyzing natural microbial populations by rRNA sequences. ASM News 51, 4–12.
Pace, N.R., 1997. A molecular view of microbial diversity and the biosphere. Science 276, 734–740.
Pruesse, E., Peplies, J., Glöckner, F.O., 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28, 1823–1829.
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 Acid Res. 41, D590–D596.
Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.-T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P., Woyke, T., 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431–437.
Söhngen, C., Bunk, B., Podstawka, A., Gleim, D., Overmann, J., 2014. BacDive-the bacterial diversity metadatabase. Nucleic Acids Res. 42, D592–D599.
Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., Weber, C.F., 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541.
Schloss, P.D., 2009. A high-throughput DNA sequence aligner for microbial ecology studies. PLoS One 4, e8230.
Seitz, K.W., Lazar, C.S., Hinrichs, K.-U., Teske, A.P., Baker, B.J., 2016. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. ISME J. 10, 1696–1705.
Stamatakis, A., Ludwig, T., Meier, H., 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21, 456–463.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.
The RNAcentral Consortium, 2015. RNAcentral: a comprehensive database of non-coding RNA sequences. Nucleic Acids Res. 43 (D1), D123–D129 online.
Vos, P.D., Garrity, G.M., Jones, D., Krieg, N.R., Ludwig, W., Rainey, F.A., Schleifer, K.H., Whitman, W.B., 2009. The Firmicutes. Springer, New York.
Ward, D.M., Weller, R., Bateson, M.M., 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature 345, 63–65.
Wuyts, J., De Rijk, P., Van de Peer, Y., Winkelmans, T., De Wachter, R., 2001. The European large subunit ribosomal RNA database. Nucleic Acid Res. 29, 175–177.
Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.H., Ludwig, W., Glöckner, F.O., Rossello-Mora, R., 2008. The all-species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. 31, 241–250.
Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.H., Whitman, W.B., Euzeby, J., Amann, R., Rossello-Mora, R., 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Rev. Microbiol. 12, 635–645.
Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K.W., Anantharaman, K., Starnawski, P., Kjeldsen, K.U., Stott, M.B., Nunoura, T., Banfield, J.F., Schramm, A., Baker, B.J., Spang, A., Ettema, T.J.G., 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541, 353–358.