Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences

Pablo Yarza[1,2,3], Pelin Yilmaz[2], Elmar Pruesse[2], Frank Oliver Glöckner[2,4], Wolfgang Ludwig[5], Karl-Heinz Schleifer[5], William B. Whitman[6], Jean Euzéby[7], Rudolf Amann[2] and Ramon Rosselló-Móra[1]

Abstract | Publicly available sequence databases of the small subunit ribosomal RNA gene, also known as 16S rRNA in bacteria and archaea, are growing rapidly, and the number of entries currently exceeds 4 million. However, a unified classification and nomenclature framework for all bacteria and archaea does not yet exist. In this Analysis article, we propose rational taxonomic boundaries for high taxa of bacteria and archaea on the basis of 16S rRNA gene sequence identities and suggest a rationale for the circumscription of uncultured taxa that is compatible with the taxonomy of cultured bacteria and archaea. Our analyses show that only nearly complete 16S rRNA sequences give accurate measures of taxonomic diversity. In addition, our analyses suggest that most of the 16S rRNA sequences of the high taxa will be discovered in environmental surveys by the end of the current decade.

[1]Marine Microbiology Group, Department of Ecology and Marine Resources, Mediterranean Institute for Advanced Studies (Spanish National Research Council (CSIC)-University of the Balearic Islands (UIB)), E-07190 Esporles, Balearic Islands, Spain.
[2]Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany.
[3]Ribocon GmbH, Fahrenheitstrasse 1, D-28359 Bremen, Germany.
[4]Jacobs University Bremen, Campus Ring 1, D-28759 Bremen, Germany.
[5]Lehrstuhl für Mikrobiologie, Technische Universität München, D-85350 Freising, Germany.
[6]Department of Microbiology, University of Georgia, 527 Biological Sciences Building, Athens, Georgia 30605–2605, USA.
[7]Société de Bactériologie Systématique et Vétérinaire (SBSV) and École Nationale Vétérinaire de Toulouse (ENVT), F-31076 Toulouse cedex 03, France.
Correspondence to P.Y., R.R.M.
e-mails: pyarza@ribocon.com; ramon@imedea.uib-csic.es
doi:10.1038/nrmicro3330

Taxonomy is the scientific discipline that involves the characterization, classification and nomenclature of biological entities and is essential to understand the full extent of diversity in the biosphere. This discipline is also fundamental for precise communication between scientists and for the successful transfer of knowledge to the general public, with implications for environmental, legal and research policies[1]. It is remarkable that, although nearly 1.3 million eukaryotic species have been described so far, this number might represent only 20% of the richness that exists[2]. For the Bacteria and Archaea, the percentage of the existing richness that has been described so far is much lower, as we discuss below. Complex communities of bacteria and archaea are present in almost all environments, including microbiomes that are closely associated with humans, animals and plants. Next-generation sequencing of PCR-amplified SSU (small subunit) ribosomal RNA (also known as 16S rRNA) genes from these communities yields several hundred thousand new sequences annually (BOX 1) and has revealed a vast, previously hidden diversity that requires identification and taxonomic classification[3,4]. The classification of bacteria and archaea is currently based on genetic and phenotypic information and is restricted to cultured strains. Obtaining pure cultures is often time-consuming and difficult and is especially challenging for microorganisms that have complex metabolic requirements. Thus, only ~11,000 bacterial and archaeal species have been classified so far. At the current rate of ~600 new descriptions per year, it has been estimated that it would take >1,000 years to classify all of the remaining species[5,6].

As the number of environmental 16S rRNA gene sequences has greatly surpassed the number of cultured microorganisms, reconciliation of the established taxonomy with this diversity and classification of the uncultured microorganisms are crucial. Moreover, the full extent of their diversity is difficult to conceptualize, owing to the lack of objective criteria (such as numerical thresholds) for taxonomic circumscriptions of uncultured microorganisms, which are identified only by sequence data. Thus, estimates for the total number of bacterial and archaeal species vary widely, from 3×10^4 (REF. 2) to ~10^{12} (REF. 7). Recognition of taxonomic thresholds of cultured bacteria and archaea, and the application of these thresholds to sequences from uncultured microorganisms, will provide more objective estimates of the diversity of these organisms on Earth.

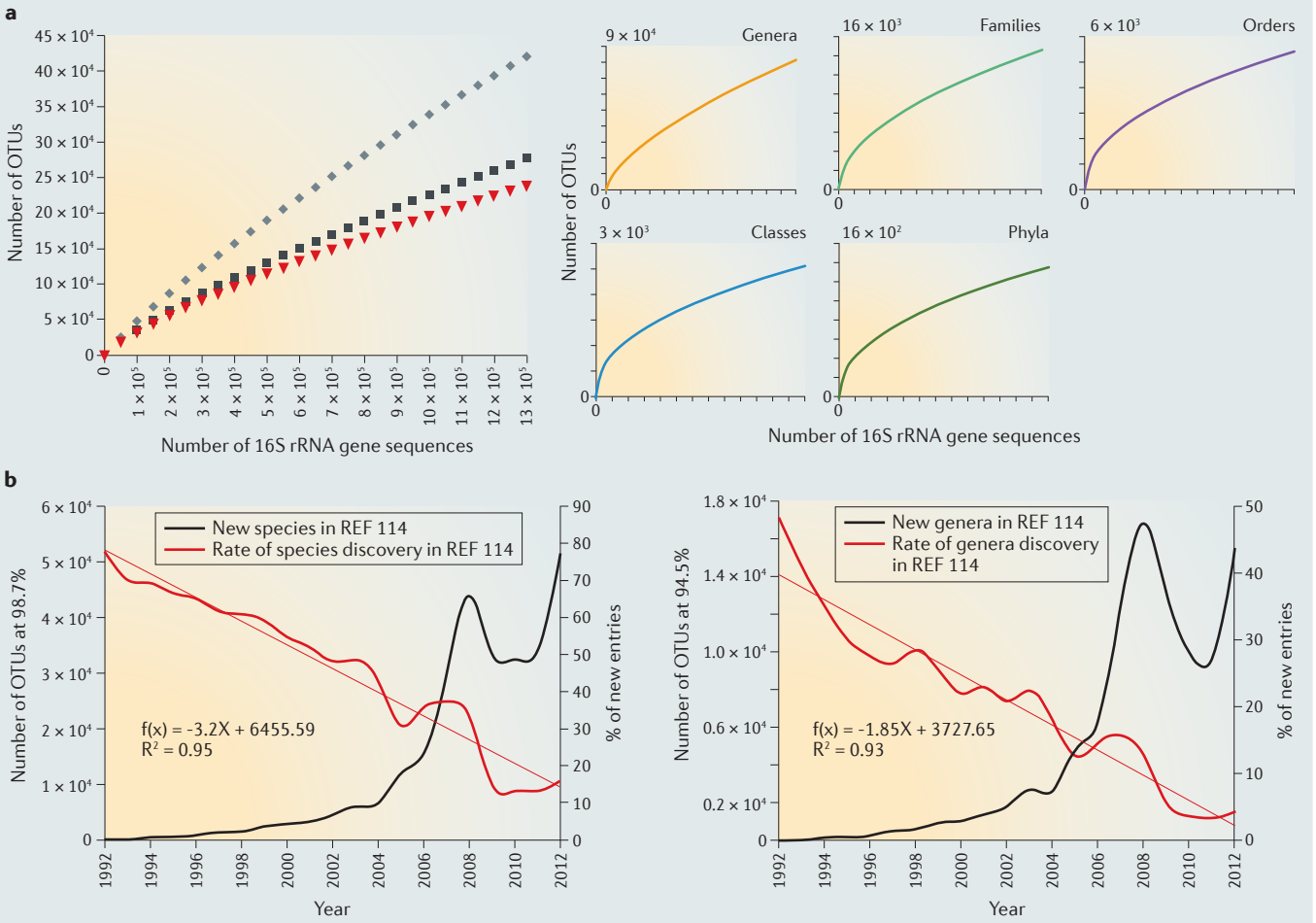
Although there are official rules for the nomenclature of bacteria and archaea, the entities that are known as taxa and their hierarchical classifications are artificial constructs[6] and so are somewhat subjective. Only

Box 1 | **Trends on taxa discovery rates**

The abundance of bacterial and archaeal taxa in the SILVA REF 114 database was estimated by means of OTUs (operational taxonomic units), which were calculated at the distinct taxonomic thresholds (see the figure, part **a**). In all cases, unsaturated rarefaction curves were obtained. In addition, 75% of the putative species and 64% of the genera consisted of single sequence units (Supplementary information S2 (box)).

A close examination of the database showed that the number of newly detected species (at 98.7% sequence identity) was about 4×10^4 taxa and the number of newly detected genera (at 94.5% sequence identity; see the figure, part **b**) was about 1.3×10^4 taxa per year over the past six years, despite the exponential increase of sequence deposits into public DNA repositories of the International Nucleotide Sequence Database Collaboration (INSDC; see Further information) (see the figure, part **b**). Intriguingly, when tracking the percentages of new taxa that are detected with regard to the sequencing effort, a clear linear decrease is observed ($R^2 = 0.95$ for species and $R^2 = 0.93$ for genus). Thus, many of the newly deposited sequences match already existing putative species and genera. If this tendency continues and the same sequencing effort persists, we could reach a very low efficiency of taxon detection within the current decade — that is, the rate of detection of new genera and of new species may be close to zero, by the end of 2015 and 2017, respectively. According to the number of species that were detected in 2012 (that is, ~21×10^4) (TABLE 2) and the estimated rate of species discovery (~4×10^4 per year for 5 years), we anticipate that the number of bacterial and archaeal species would increase to around ~4×10^5 species by 2017. Similarly, we conclude that the number of genera that are detectable by this method will be the total number that are observed by 2012 (~6.1×10^4) plus the numbers that are estimated in 2013, 2014 and 2015 (3.9×10^4), or around ~1×10^5 genera by 2015.

In part **a** of the figure, the taxonomic thresholds are as follows: genera 94.5% (yellow), families 86.5% (green), orders 82.0% (purple), classes 78.5% (blue) and phyla 75.0% (dark green). For species, thresholds of 98.7% (red), 99.0% (dark grey) and 99.5% (light grey) were used.



Diversity
A term used to describe the effective number of taxa (that is, the richness) of a particular rank and their respective abundances (that is, the evenness).

the rank of species is circumscribed by a combination of well-accepted criteria, which include: DNA–DNA hybridization (DDH), with a threshold around 70%; average sequence identities of shared genes (ANI), with a threshold of around 94–96%[8]; and 16S rRNA gene sequence identities, with a threshold of around 98.7%[9]. These genetic criteria should always be accompanied by a discriminant phenotypic property. To a lesser extent, genera are also recognized by their phylogenetic separation from other such groups and the possession of 16S rRNA gene sequence identities of >95%[10]. There are no robust rules for the circumscription of ranks above genus, although these high taxonomic ranks describe a large proportion of the phenotypic and ecological diversity[11–13]. Sorting species into higher taxa is one of the most fundamental problems in biology[14]. Moreover, the use of molecular criteria to classify bacterial and archaeal species offers the possibility of applying these numerical

Table 1 | **Taxonomic thresholds of bacteria and archaea***

	Genus	Family	Order	Class	Phylum
Number of taxa	568	201	85	39	23
Median sequence identity	96.4% (96.2, 96.55)	92.25% (91.65, 92.9)	89.2% (88.25, 90.1)	86.35% (84.7, 87.95)	83.68% (81.6, 85.93)
Minimum sequence identity	94.8% (94.55, 95.05)	87.65% (86.8, 88.4)	83.55% (82.25, 84.8)	80.38% (78.55, 82.5)	77.43% (74.95, 79.9)
Threshold sequence identity	94.5%	86.5%	82.0%	78.5%	75.0%

*Results based on the Living Tree Project (LTP) 102 data set. Values given are the Hodges–Lehmann estimator (also known as the 'pseudo-median') and its 95% confidence interval (in parentheses) of all of the taxa median and minimum sequence identities for the 16S ribosomal RNA genes. Values were calculated using the Wilcoxon signed rank test ('wilcox.test'), which was implemented in the R package 'stats'[53].

thresholds to the classification of the vast extent of diversity that is yet to be cultured. The pre-classification of environmental sequences in a global and unified system may have many advantages, such as avoiding cumbersome taxonomic revisions[14] and producing a stable taxonomic framework. Thus, there is an urgent need to establish uniform criteria for high taxonomic ranks, not least to enable a phylogenetic perspective.

Although the use of a single gene has many disadvantages, as discussed in detail below, the 16S rRNA gene is currently the only widely used taxonomic marker that is sufficiently informative and that has been compiled in comprehensive and quality-controlled databases along with reliable taxonomic information[3,15]. Thus, in our opinion, the defining criteria for high taxonomic ranks should be based on this gene, if only for pragmatic reasons[16–18]. In summary, our goal is to develop a stable classification system for high taxonomic ranks, based on the 16S rRNA gene, that is applicable to uncultured microorganisms and is compatible with the procedures that are currently used for cultured bacteria and archaea.

In this Analysis article, based on an analysis of current 16S rRNA gene sequence databases and the catalogue of taxa with standing in nomenclature, we provide suggestions for the taxonomic classification of environmental sequences. This includes rational taxonomic boundaries for high taxa of bacteria and archaea (that is, taxa at the genus level and above), an updated census of the bacterial and archaeal taxa on Earth and a rationale for stable hierarchical classification of high taxa that could apply to both cultured and uncultured microorganisms.

Taxonomic thresholds for bacteria and archaea

The ribosomes are functionally constant ubiquitous molecules that were first pioneered as universal molecular chronometers by Woese in the 1970s[19]. The primary structures of the two major rRNA subunits, 16S and 23S, comprise a particular combination of conserved, variable and hypervariable regions that evolve at different rates and enable the resolution of both very ancient lineages (for example, domains) and more modern lineages (for example, genera)[20–22]. Moreover, their secondary structures include a total of ~50 helices for the 16S rRNA and ~100 helices for the 23S rRNA, which results in base pairing of about 67% of the residues[17]. These highly conserved secondary structural features are of great functional importance and can be used to ensure positional homology in multiple sequence alignments and phylogenetic analyses[20]. Over the past few decades, the 16S rRNA gene has become the most sequenced taxonomic marker and is the cornerstone for the current systematic classification of bacteria and archaea[17].

The taxonomic thresholds for taxa that are higher than species were calculated using the curated and regularly updated 16S rRNA gene sequence database of type strains of species with validly published names that is found at the Living Tree Project[15] (LTP; see Further information). The LTP is compatible with the rRNA databases of the SILVA project[3] (see Further information) and is compliant with the universal and optimized alignments that have been generated using the ARB software package[23] since the early 1990s[20]. Moreover, the LTP only considers 16S rRNA gene sequences of modest to good quality (that is, sequences that are longer than 1,450 nucleotides, with the average proportion of ambiguous positions <1%). For this analysis, we used release LTPs102, which comprised sequences of type strains that had validly published names up to February 2010 (REF. 24). This data set contained 8,602 bacterial and archaeal species that were classified into the following high taxonomic ranks: 1,779 genera, 285 families, 115 orders, 52 classes, 29 phyla and two domains. To optimize the calculation, each of the different categories was initially 'sieved' by removing all taxa that had less than three members, as well as outliers (generally misclassified taxa) and taxa that have pending classifications (Supplementary information S1 (table)). A global identity matrix was then used to determine the median and minimum identity values for each taxon and to calculate the global descriptors for each of the high taxonomic categories. We calculated the threshold as the level of identity below which two units would belong to different taxa in the same rank (TABLE 1). A threshold for species category could not be calculated using this method, as only one member — the type strain — of each species was considered.

In summary, a sequence identity of 94.5% or lower for two 16S rRNA genes is strong evidence for distinct genera, 86.5% or lower is strong evidence for distinct families, 82.0% or lower is strong evidence for distinct orders, 78.5% or lower is strong evidence for distinct classes and 75.0% or lower is strong evidence for distinct phyla (TABLE 1). These values are mostly in accordance with previously reported boundaries, which were deduced using similar approaches but with less representative data sets[15,16]. As the ranks suborder and subclass overlapped with the categories family and order, respectively (Supplementary information S2 (box)), they were not

SSU
(Small subunit). The small subunit of the ribosome, which is 16S ribosomal RNA for the Bacteria and the Archaea and 18S rRNA for the Eukarya.

Bacterial and archaeal species
A monophyletic group of organisms with a high degree of coherence in their genetic and phenotypic traits, which differentiate it from its close relatives.

High taxonomic ranks
The taxonomic categories of genus and above.

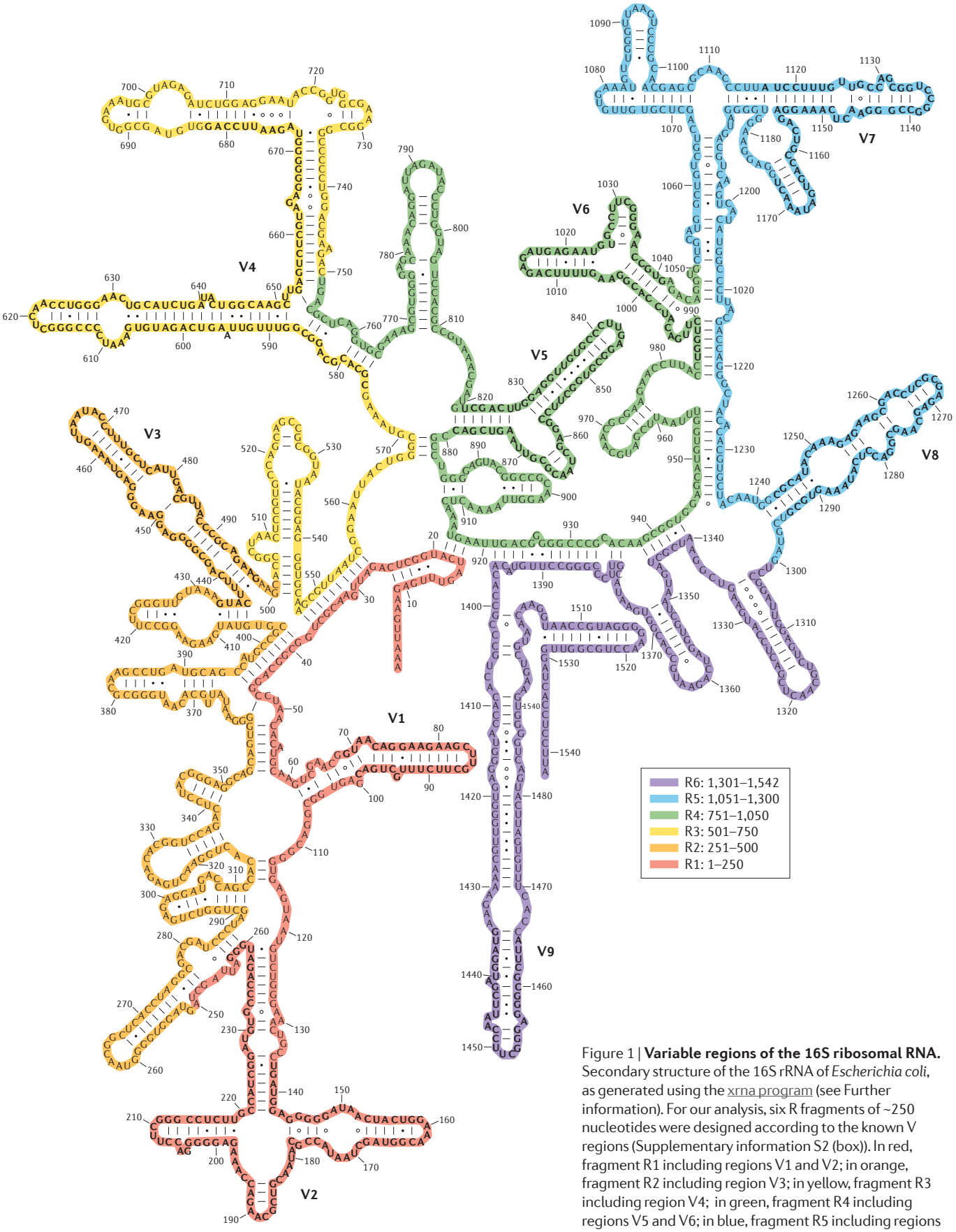


Figure 1 | **Variable regions of the 16S ribosomal RNA.** Secondary structure of the 16S rRNA of *Escherichia coli*, as generated using the xrna program (see Further information). For our analysis, six R fragments of ~250 nucleotides were designed according to the known V regions (Supplementary information S2 (box)). In red, fragment R1 including regions V1 and V2; in orange, fragment R2 including region V3; in yellow, fragment R3 including region V4; in green, fragment R4 including regions V5 and V6; in blue, fragment R5 including regions V7 and V8; and in purple, fragment R6 including region V9.

further considered. Importantly, these thresholds are minimum values, and the sequence identities of different taxa within each rank may be higher than the threshold if it is supported by other types of evidence; for example, the 94.5% threshold for genera does not preclude the formation of genera that have sequence identities of 96% if it is supported by other phenotypic, genetic or environmental data.

We consider the tight confidence intervals at the lowest ranks to be a result of the thorough sampling of many groups and the historical consensus among taxonomists when circumscribing new genera and families (Supplementary information S2 (box)). By contrast, the larger confidence intervals for ranks that are above family may be due to several artefacts and confounding factors. For example: if only a few species or sequences are available for a taxon, the scarce input may lead to poor conclusions with weak statistical outcomes (such as the phylum Acidobacteria, which contained only eight type strain sequences); early classifications were carried out using obsolete techniques, which grouped together species that have been shown by modern methods to belong to different taxa, such as the family Bacillaceae and genus *Mycoplana*[25,26]; limitations of the phylogenetic resolution of the 16S rRNA marker, including lack of informative content[18] and variations in its evolutionary tempo[27], can lead to anomalous sequence identities; and, in certain taxa that have less phylogenetic coherence, their members may be distributed between a well-defined monophyletic core and several borderline taxa, which are often responsible for most of the divergence.

Optimum sequence length for taxonomic assignments. As only 23% of the 16S rRNA sequences that are in public gene repositories are longer than 900 bp, we analysed the accuracy of predicting taxonomic thresholds and richness estimates from partial sequences. As the degree of conservation of the 16S rRNA gene varies between different regions[22], we expected the threshold to depend on the particular regions that were sequenced. To test this hypothesis, the full-length sequences in LTP release 108 were subdivided into segments of 250 bp to mimic much of the recently collected data[21,28,29] (FIG. 1; Supplementary information S2 (box)). For each segment, and for combinations of segments, the taxonomic thresholds were calculated in the same way as for the full-length 16S rRNA gene sequences. All gene segments except R1 produced thresholds and error values that were comparable to the whole gene sequences (Supplementary information S2 (box)). However, clustering analyses of individual segments and combinations of segments greatly underestimated the taxa richness of the LTP data set, especially at the high ranks. This effect was strongest with R1, which includes the V1 and V2 regions and is highly variable (FIG. 2a). Analysis of sequences >250 bp partially ameliorated the underestimation for the lower taxa, and including the first 750 nucleotides (that is, from R1 to R3) was sufficient to discriminate up to 90% of the total richness that is retained within the ranks of family, genus and species (FIG. 2b). To capture the high taxonomic ranks, full-length 16S rRNA gene sequences

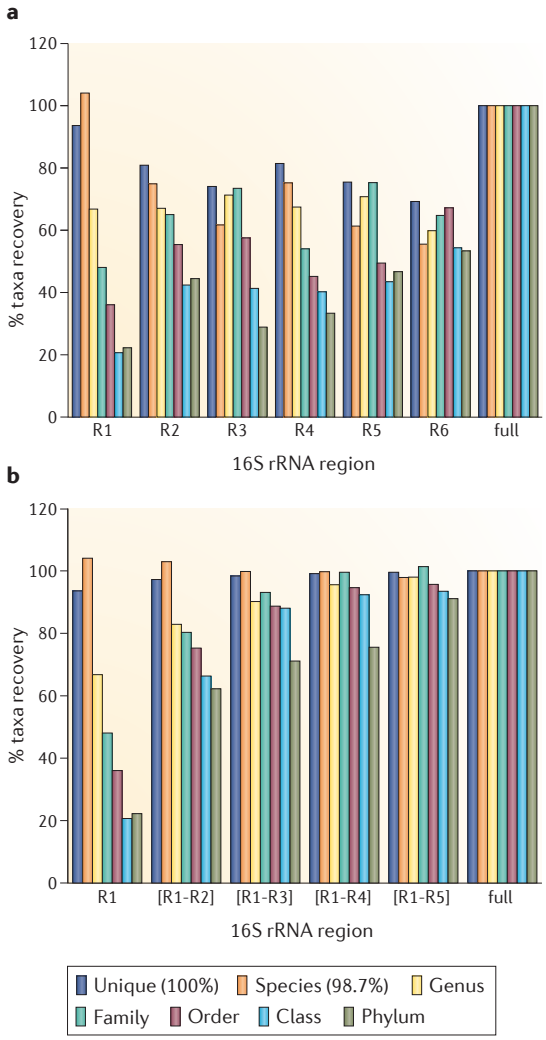


Figure 2 | **Predicting taxa richness using partial 16S ribosomal RNA sequences.** The 16S rDNA data set that was used was obtained from the Living Tree Project (LTP) release 108. Taxa recovery was approached using OTUs (operational taxonomic units) that were calculated using CD-HIT version 4.5 (REF. 54). In all cases, the general taxonomic thresholds that were used for clustering are: 98.7% (species), 94.5% (genus), 86.5% (family), 82.0% (order), 78.5% (class) and 75.0% (phylum). **a** | Six R fragments of ~250 nucleotides were designed according to the V regions (Supplementary information S2 (box)); the results that were obtained for the full-length sequence are included for comparison. **b** | Based on the R fragments, four additional segments were created, all of which start at the 5′ end (that is, all including the V1 and V2 regions) but with different sizes: segment R1 contains the 250 nucleotides at the 5′ end, R1–R2 contains the 5′ 500 nucleotides, R1–R3 contains the 5′ 750 nucleotides, R1–R4 contains the 5′ 1050 nucleotides, R1–R5 contains the 5′ 1300 nucleotides, and 'full' corresponds to the full *Escherichia coli* 16S rRNA gene (which is 1,542 nucleotides). Analysis of the taxa recovery rate indicates a great underestimation of taxa richness when partial sequences are used. Although the situation tends to ameliorate as longer segments are considered, near full-length 16S rRNA genes sequences are required for accurate richness estimations and accurate classifications of high taxa.

Table 2 | **Numbers of taxa with standing in nomenclature and putative classifiable taxa present in release 114 of the SILVA database***

	LPSN 2012		SILVA REF 114 incremental‡		SILVA REF 114 single-step‡	SILVA REF 114 ambiguous§
	Bacteria	Archaea	Bacteria	Archaea	Bacteria and archaea	Bacteria and archaea
Sequences			1,265,442	41,228	1,306,670	31,469
Species	9,624	391	246,841	13,159	241,254	28,983
Genera	1,916	113	90,068	4,383	80,939	19,607
Families	290	27	16,719	1,239	14,369	4,803
Orders	122	15	6,230	682	5,366	1,148
Classes	79	9	2,969	431	2,573	408
Phyla	27	2	1,481	287	1,356	84

*Numbers obtained from the List of Prokaryotic Names with Standing in Nomenclature (LPSN; see Further information), excluding: homotypic synonyms, new combinations, nomina nova, later heterotypic synonyms, illegitimate names and subspecific epithets, and from SILVA database, according to EMBL release 114 (up to December 2012). The complete domain Eukarya and all sequences that have alignments that start after *Escherichia coli* 16S rRNA gene position 252 and/or end before *E. coli* position 349 were excluded from the analysis. ‡Clustering analysis was carried out using CD-HIT version 4.5 (REF. 54) using two distinct approaches. In one approach (that is, the incremental approach), multiple data sets were generated according to the year of sequence submission to the International Nucleotide Sequence Database Collaboration (INSDC; see Further information) database and then incrementally clustered into a final non-redundant data set; in the second approach (that is, the single-step approach), the complete sequence data set was clustered at once. In both cases, the general taxonomic thresholds were used for clustering, which were: 98.7% (species), 94.5% (genus), 86.5% (family), 82.0% (order), 78.5% (class) and 75.0% (phylum). §All sequences that had a SILVA pintail score of less than 75% were reported as ambiguous taxa when their alignment identity surpassed the taxonomic thresholds (extended information in Supplementary information S2 (box)).

were necessary when sequence identities were <84.0%, which is the family threshold that was reported by the R1–R3 fragment (Supplementary information S2 (box)). For fragments that were smaller than R1–R3, the information content was too low to provide acceptable results, even for genera estimations. Although the most variable region, R1, seems to reflect the species recovery that was achieved using the whole gene (FIG. 2a), this short region will not provide sufficient resolution for the taxonomic ranks that are higher than species. In conclusion, the results indicate that near full-length SSU rRNA gene sequences are required for accurate richness estimations and accurate classifications of high taxa.

Detected versus described taxa

The thresholds that were calculated above were used to estimate total taxon abundances for the curated SILVA REF 114 database[3] (see Further information; corresponding to EMBL release 114 of December 2012), which compiled ~1.3 million 16S rRNA sequences of >900 bp in length. Rarefaction curves of taxa recovery did not show saturation (BOX 1). For the high taxa, the analyses provide evidence for a current census of at least 6.1×10^4 genera, 9.6×10^3 families, 4.2×10^3 orders, 2.2×10^3 classes and 1.3×10^3 phyla of bacteria and archaea. These conservative estimates were reached by subtracting the numbers of ambiguous taxa from the numbers obtained using the single step approach (TABLE 2) and exceed the numbers of currently described taxa by orders of magnitude. Major problems that could bias these estimates include anomalous sequences that result from amplification and/or sequencing errors as well as the production of sequence chimaeras[30]. However, even in the worst-case scenario for the number of chimaeras, the number of phyla that are detected would only be reduced from ~1,350 to ~1,100 (Supplementary information S2 (box)). The SILVA PARC database was not explored owing to the large uncertainties that are caused by the presence of many low-quality sequences.

The number of high taxa that are yet to be described may seem surprising; however, for many microbiologists, the current definitions of these categories are somewhat arbitrary and follow "negative phylogenetic criteria"[12], or are "highly subjective in many cases"[31]. Taxonomists might have been too cautious and tended to classify new taxa into existing units (for example, placing the genus *Sphaerochaeta* in the family Spirochaetaceae[32]) rather than proposing new ones, owing to the absence of accepted standards to set taxa boundaries. Consequently, most of the 29 extant phyla currently contain only a single class, most of the classes contain only a single order and most orders contain only a single family. The results reinforce the fact that the implementation of the thresholds that are proposed in this Analysis article will produce taxa circumscriptions and classifications of comparable diversity and phylogenetic depth among all bacterial and archaeal lineages. Moreover, these thresholds could be applied to both cultured and uncultured putative taxa.

Global species richness estimate. Previous exhaustive studies on determining the species thresholds indicate that a plausible species boundary would be between 98% and 99% 16S rRNA gene sequence identity at reasonable probabilities[9,33]. As an attempt to measure species richness in the SILVA REF 114 data set, three different thresholds of 98.7%[9], 99.0% and 99.5% were considered. Both lower thresholds yielded similar results, whereas at 99.5%, the number of total OTUs (operational taxonomic units) that were detected nearly doubled (BOX 1; Supplementary information S2 (box)). However, given the noise that is introduced by sequencing errors[33], the intragenomic variability of the rRNA operon[24] and the strong support for the threshold at 98.7%[9], the conservativeness of this cut-off (that is, the number of OTUs that are produced at 98.7% sequence identity using the LTP database was 78% of the total number of species that are in the same data set) was preferred for subsequent

OTUs
(Operational taxonomic units). Groups of sequences that are meaningfully separated from other sequences by hierarchical clustering techniques (independent of phylogenetic inferences) and using strict sequence identity thresholds.

Box 2 | **Guidelines for CTUs**

- Every sequence that is under consideration should be assigned to at least two CTUs (candidate taxonomic units): one candidate genus and one candidate phylum.
- Each OTU (operational taxonomic unit), solely as it has been created using a taxonomic threshold (for example, 86.5% for family), is a potential CTU.
- If an OPU (operational phylogenetic unit) contains at least 70% of the members of a given OTU, that OPU becomes a CTU.
- For an OPU that represents less than 70% of the members of an OTU: if it is possible to merge it with another OPU (or OPUs) of the same 'rank' and the resultant 'merged' OPU has better stability and its sequence identity boundary is close to the taxonomic threshold, that new OPU becomes a CTU. Otherwise, as long as it is supported by monophyly and has a minimum sequence identity that is appropriate to its rank, then this OPU becomes a CTU. If the OPU cannot be assigned to any other rank except genus and phylum, its sequence members should be assigned to 'family-unknown', 'order-unknown' or 'class-unknown' as appropriate.
- If two CTUs of distinct rank share the same sequence data set, the name of the higher rank has to be used to give the name of the group in a tree figure.
- CTUs that contain taxa with standing in nomenclature according to the Bacteriological Code must follow the rules of priority of the code.
- Nomenclature of CTUs follows the format X.AB-C (for example, Spirochaetes. Family3-1), where X, A and B are mandatory elements and C is optional. X is the denomination of a candidate rank — for example, Spirochaetes — and it is an alphanumeric string. A is the name of the taxonomic rank, that is, species, genus, family, order, class or phylum. B is a positive integer (not necessarily correlative) of 1–5 digits. C is an optional element that can be used to differentiate several CTUs that are derived from the same OTU (that is, non-monophyletic OTUs).

analyses. Approximately 2.1×10^5 species-level OTUs were detected in the SILVA REF 114 database, excluding counts by anomalous sequences (for example, chimaeras) (BOX 1; TABLE 2). However, the discovery of new species-level OTUs is expected to rapidly decrease. Despite large increases in sequencing, the number of new species that have been discovered has been relatively constant at ~$4 \times 10^4 \pm 0.8 \times 10^4$ for the past five years (BOX 1). By the end of 2017, we expect the rate of discovery of new species to decrease to zero (BOX 1) and the total number of species to be around 4×10^5. This value is much lower than an earlier species estimation of 2×10^6 (REF. 34), which was just for oceanic waters. Although our estimate may double or even increase fourfold or fivefold, we speculate that it is unlikely that the final number of species of bacteria and archaea will exceed 1×10^7.

Caveats. These global predictions of taxa richness must be taken with care. First, it remains uncertain whether rarefaction curves will saturate if the taxa recovery rates reach zero within the next few years (BOX 1). Even if that is the case, an unknown proportion of less abundant organisms might remain undetected, which would result in an underestimation of the true taxa richness. Biases might also be introduced by the fact that most of the sequences were obtained by PCR amplification from environmental DNA and, perhaps, the sequences might reflect the biases of that methodology[35]. In this regard, amplification-free metagenome reads generally contain only short sequences and cannot replace sequences from PCR. However, the metagenomic information that is available does not indicate that there are large discrepancies with the PCR-derived databases[36]. There is also a certain redundancy of sampled environments in diversity studies; for example, surface ocean samples and the human microbiome have been examined in many studies, whereas remote and extreme habitats have been sampled much less frequently and will almost certainly contain as-yet-undiscovered bacteria and archaea. Therefore, our results probably best reflect the diversity of the most common terrestrial and aquatic habitats and suggest that these habitats will be exhaustively described within the next five years (BOX 1).

Candidate taxonomic units

Based on the general taxonomic thresholds that are outlined above, we propose a new biodiversity unit called the CTU (candidate taxonomic unit), which is compatible with the hierarchy that was established in the Bacteriological Code (see Further information). We define a CTU as a biological entity that is delineated by a monophyletic set of sequences with a sequence identity that stays within, or very close to, the taxonomic threshold that is proposed for a given rank. The purpose of the CTU is to harmonize the classification of cultured and uncultured microorganisms. CTUs can be applied to any gene or set of genes (for example, concatenates) for which the taxonomic value has been proven and thresholds have been calculated (such as the 16S rRNA gene that is discussed in this Analysis article). The CTU is conceived to be a combination of the OTU (for example, see REF. 37) and OPU (operational phylogenetic unit; for example, see REF. 38), which are currently used by microbial ecologists and systematists.

The recognition of CTUs starts with a preliminary classification based on OTUs (which is calculated using standard clustering algorithms with the different taxonomic thresholds that are identified in TABLE 1), the results of which guide the search for local meaningful phylogenetic clades in a reliable tree topology. This OTU-based procedure guarantees the analysis of many more meaningful clades than would be examined in a 'naked' topology, which hence maximizes the number of CTUs that are discovered (see BOX 2 for further methodological considerations). The ability to specify a taxonomic rank for particular clades is a major advance in understanding tree topologies and goes beyond the classic phylogenetic delineation, as the rank enables objective comparisons of the distinct lineages and also provides substantial information about the coherence of genetic, phenotypic and ecological traits (for example, see REFs 31,39,40).

CTUs can be readily applied to re-evaluate the taxonomy of described organisms, as in the case of the phylum Spirochaetes (FIG. 3). All high taxa in this phylum were described on the basis of phenotypic traits[41–44]. By the end of 2013, the 112 described species had been organized into 16 genera, four families, one order and one class. Many of the 112 species corresponded to uncultured organisms, and just 82 type strains were represented by good-quality 16S rRNA sequences in the LTP data set (release 111). The 72.3% sequence identity that was found between *Leptospira licerasiae* (European Nucleotide Archive (ENA) accession number EF612284)

CTU
(Candidate taxonomic unit). A biological entity that is delineated by a monophyletic set of sequences with a sequence identity that stays within, or very close to, the taxonomic threshold that is proposed for a given rank.

OPU
(Operational phylogenetic unit). A group of sequences that appear as a monophyletic clade that is meaningfully separated from the remaining sequences in a genealogical reconstruction.

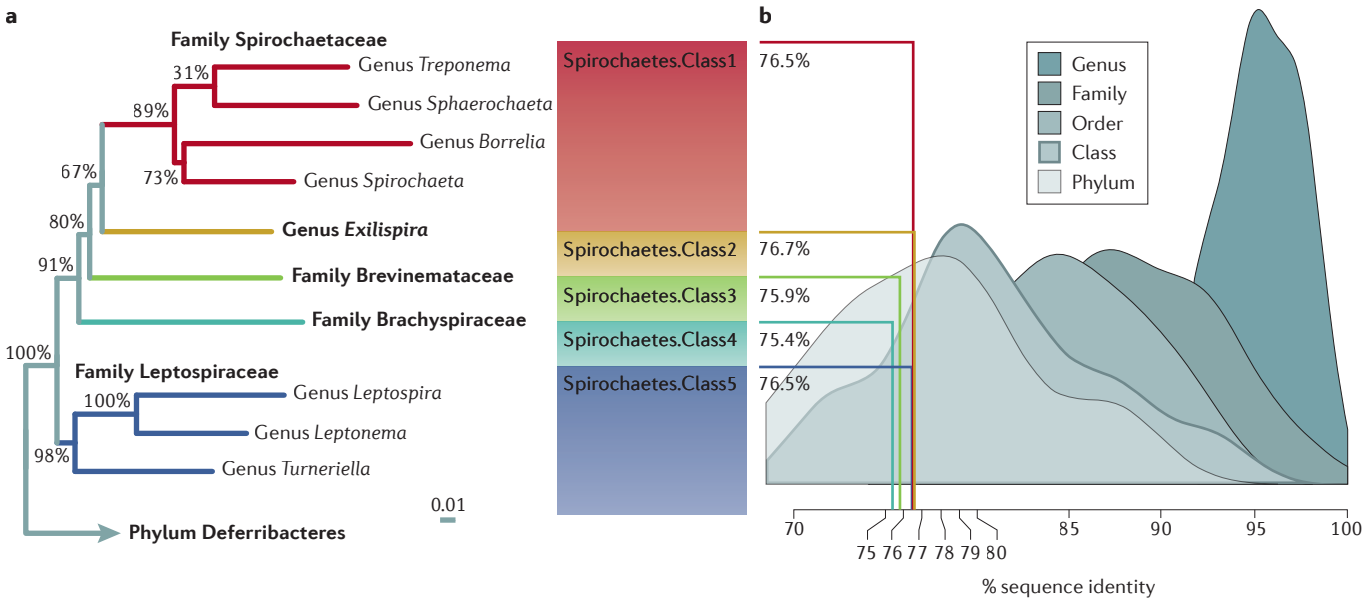


Figure 3 | **Classification of phylum Spirochaetes into candidate taxonomic units.** **a** | The phylum Spirochaetes is currently classified within one class (Spirochaetes), one order (Spirochaetales) and four families (Spirochaetaceae, Brevinemataceae, Brachyspiraceae and Leptospiraceae). A 16S ribosomal RNA phylogenetic reconstruction of the phylum Spirochaetes, based on 82 type strains obtained from the Living Tree Project (LTP) release 111 and using the RAxML algorithm[55], shows a clear phylogenetic separation of five lineages within the phylum. Percentage bootstrap values are indicated next to the tree nodes. Scale bar indicates estimated sequence divergence. **b** | Distribution of the minimum sequence identity that is found in each bacterial and archaeal taxon in LTP release 102 (Supplementary information S1 (table) and Supplementary information S2 (box)). The average sequence identity among the five groups is shown as percentages and is linked to the density plot on the right-hand side to further confirm that the distances are within the values that are commonly displayed between taxa belonging to the rank of class. According to the strong monophyletic support and the low inter-clade sequence identities below the taxonomic threshold of 78.5% (TABLE 1), our analysis predicts five classes within Spirochaetes rather than one. The ten branches displayed in part **a** correspond to the following European Nucleotide Archive (ENA) accession numbers: *Treponema maltophyla*, X87140; *Sphaerochaeta globosa*, AF357916; *Borrelia americana*, EU081285; *Spirochaeta litoralis*, FR733665; *Exilispira thermophila*, AB364473; *Brevinema andersonii*, GU993264; *Brachyspira aalborgi*, Z22781; *Leptospira interrogans*, Z12817; *Leptonema illini*, AY714984; *Turneriella parva*, AY293856.

and *Treponema saccharophilum* (ENA accession number M71238) and the low sequence identity of 74.9% that was observed within the family Spirochaetaceae (owing to the large differences between the genera *Borrelia*, *Spirochaeta* and *Treponema*) suggest that they should be classified into more than one class (that is, Spirochaetes) and one order (that is, Spirochaetales). In addition, the lowest sequence identity in the current family Leptospiraceae is 79.2% (that is, close to the class level; see TABLE 1) owing to the high divergence among the genera *Leptospira*, *Leptonema* and *Turneriella*. Overall, these results indicate that, although there is high phylogenetic support for the group Spirochaetes, the internal structure of the phylum should be revised in order to more accurately reflect its diversity. A previous attempt to reclassify this phylum on the basis of molecular signatures from multiprotein data sets also showed that its main lineages should have high taxonomic ranks[45]. However, the lack of criteria to establish ranks on the basis of protein sequences prevented a complete reclassification, and the four existing families were only elevated to orders (that is, the next level)[45]. The CTU analysis that we carried out for this Analysis article provides evidence for the circumscription of five independent classes within the phylum Spirochaetes: Spirochaetes.Class1 (which comprises the genera *Borrelia*, *Treponema*, *Sphaerochaeta* and *Spirochaeta*), Spirochaetes.Class2 (which comprises the genus *Exilispira*), Spirochaetes.Class3 (which comprises the genus *Brevinema*), Spirochaetes.Class4 (which comprises the genus *Brachyspira*) and Spirochaetes.Class5 (which comprises the genera *Leptospira*, *Leptonema* and *Turneriella*) (FIG. 3).

Classification of environmental diversity. CTUs also enable the systematic classification of uncultured bacteria and archaea that are present in the 16S rRNA gene databases. To test the method with sufficient intensity, 9,426 sequences that were distributed into 15 candidate divisions and well-known environmental clades were selected from the SILVA REF 108 database. This selection encompassed the phyla Elusimicrobia, Caldiserica, Armatimonadetes, Korarchaeota, Nanoarchaeota, the SAR11 clade and the candidate divisions that are commonly known as TM7, OD1, OP3, OP9, OP11, WS3, WS6 and SR1. A detailed phylogenetic classification is available for each group in Supplementary information S3 (figure) and Supplementary information S4 (table). As the monophyly is a more important premise than the

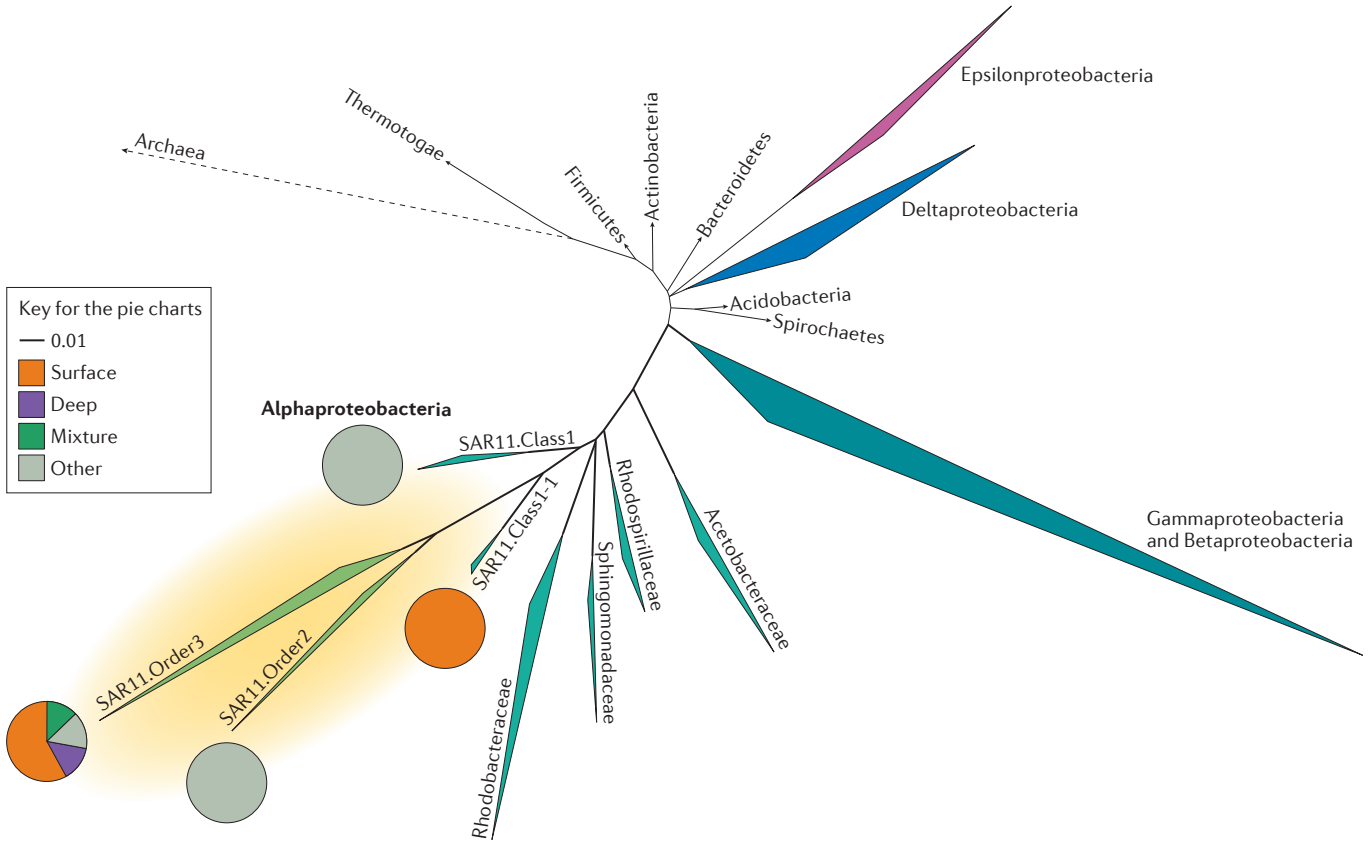


Figure 4 | **Phylogenetic reconstruction of the SAR11 group within the Proteobacteria based on 16S ribosomal RNA gene sequences.** The reconstruction was based on Living Tree Project (LTP) release 111. Several phyla and the major alphaproteobacterial families are shown for reference. The CTU (candidate taxonomic unit) analysis that was carried out on the SAR11 group, which was suggested to form a coherent taxon within the class Alphaproteobacteria, revealed the existence of three classes, SAR11.Class1, SAR11.Class1-1 and SAR11.Class2. One class, SAR11.Class2, is composed of two candidate orders (SAR11.Order2 and SAR11.Order3; yellow). The pie charts indicate the percentage of genera that comprise only surface-water sequences (orange), only deep-water sequences (purple), both surface-water and deep-water sequences (green) and other samples (grey). The ecological coherence with respect to the site of sequence retrieval is inversely correlated with the rank of the taxon. In addition to the large phylogenetic depth of Alphaproteobacteria, the positions of classes Epsilonproteobacteria and Deltaproteobacteria in this tree reconstruction strongly suggest that the phylum Proteobacteria is not monophyletic. Scale bar represents estimated sequence divergence.

fulfillment of strict thresholds, CTUs did not always comprise the same sequence data sets as the original OTUs. In this regard, accurate affiliations can only be inferred by means of phylogenetic methods using evolutionary models and not simply using hierarchical clustering techniques (that is, OTU modelling). However, it is remarkable that the total number of delineated CTUs (3,553 CTUs, distributed into 2,053 genera, 767 families, 411 orders, 240 classes and 82 phyla) was nearly identical to the total number of calculated OTUs (3,562 in total; Supplementary information S2 (box)). Therefore, the total number of candidate taxa can be accurately estimated by OTU modelling of 16S rRNA using the general taxonomic thresholds (TABLE 2; Supplementary information S2 (box)). It is also important to remember that cultivation and further description (such as genotype and phenotype) of microorganisms that are affiliated with those genealogical groups will consolidate their classification.

The so-called environmental clades were originally defined on the basis of the occurrence of sequences that are associated with certain environments, at a time when only a few species or sequences were described (for example, OD1 (REF. 46)). However, as new sequences were added to the database, the ecological coherence of the groups frequently blurred. In addition, the circumscription of environmental clades has generally not considered aspects that are related to size, phylogenetic depth, internal taxonomic classification or a standard nomenclature format; for example, the alphaproteobacterial clade SAR11 should, on the basis of the methodology that is suggested here, be regarded as comprising multiple classes (FIG. 4). Using our approach, SAR11 would be classified into 105 candidate genera, five candidate families and four candidate orders, which would be distributed into three candidate classes. In ecological studies, this clade is often contrasted with the Roseobacter clade affiliated (RCA) group, which is a much narrower marine lineage of Alphaproteobacteria, and which, according to our method, would be a family within Rhodobacterales[47].

In our experience, if the CTU rationale is applied to environmental clades, the habitat specificity becomes more homogeneous and is inversely correlated with the level of the taxonomic rank, as pointed out by Philippot *et al.*[11]; for example, the surface and deep groups of SAR11 were segregated into separate CTUs at the genus level, which provides an example that different niches are occupied by different phylotypes (FIG. 4). From 105 total candidate genera that were found in SAR11, 52 consisted only of sequences that were retrieved from surface waters, and 12 consisted only of sequences that were retrieved from deep-sea samples. To a lesser extent, habitat specificity can also be observed at higher CTU ranks; for example, the small Sar11.Class1-1 so far contains only surface-water sequences (FIG. 4). Likewise, the candidate division OP10 (now known as Armatimonadetes) showed habitat specificity at the phylum level. Armatimonadetes. Phylum3 sequences were mostly obtained from microbial-mat or hot-spring samples, whereas Armatimonadetes. Phylum6 originated from terrestrial habitats.

Phylogenetic coherence of classes and phyla

Although Alphaproteobacteria, Betaproteobacteria and Gammaproteobacteria form a well-supported monophyletic cluster (FIG. 4), the 16S rRNA gene sequence identities between members of the three classes are lower than the class thresholds that are proposed above (TABLE 1); for example, among the different species of the genera *Gallibacterium* (from Gammaproteobacteria) and *Ehrlichia* (from Alphaproteobacteria), sequence identities are about 75%. When SAR11 (which is a multiclass data set) is included in the calculations, the sequence identities decrease to 74.5%. Considering its phylogenetic depth (FIG. 4), the 'class' Alphaproteobacteria constitutes a well-separated phylum in the global topology of the Bacteria and Archaea. Together with the remarkable tendency of Deltaproteobacteria and Epsilonproteobacteria to segregate from the Gammaproteobacteria and Betaproteobacteria, we suggest that the phylum rank of Proteobacteria should be reconsidered (FIG. 4).

Similarly, the candidate divisions TM7, OD1, OP3, OP11 and the phyla Elusimicrobia, Caldiserica and Armatimonadetes (Supplementary information S3 (figure)) seemed to be monophyletic assemblages of several candidate phyla, which is reminiscent of the superphyla Planctomycetes–Verrucomicrobia–Chlamydiae (PVC)[48] or Thaumarchaeota–Aigarchaeota–Crenarchaeota–Korarchaeota (TACK)[49]. For example, OD1 was described as a new candidate division using fewer than 100 16S rRNA gene sequences[46]. At that time, these data were sufficient to demonstrate the stable segregation of a new clade from the parent OP11 candidate division[50]. Using an updated OD1 data set (that is, according to SILVA classification[51] using more than 700 nearly full-length entries), evidence is obtained for up to 102 class CTUs and 28 phylum CTUs within this clade (Supplementary information S2 (box)). Likewise, without clear rules for phylum circumscription, the authors who described and published the species *Elusimicrobium minutum* suggested that the whole Termite Group 1 (TG1)[50] could be equated to a single phylum, Elusimicrobia[52]. However, according to our analysis, Elusimicrobia is a stable monophyletic clade that unites up to 16 class CTUs and seven phylum CTUs; the type species *E. minutum* is classified as a member of the following candidate high taxa: Elusimicrobia. Phylum1, Elusimicrobia.Class1, Elusimicrobia.Order2, Elusimicrobia.Family5 and Elusimicrobia.Genus8 (Supplementary information S3 (figure)).

These discrepancies again emphasize the need to reconcile classification and nomenclature at all levels. They illustrate the fact that nomenclatural changes, particularly on large uncultured or environmental clades, will have major implications for sequence databases, which are committed to guarantee nomenclatural consistency for classification purposes[51].

Conclusions

The comparative analysis of 16S rRNA gene sequences enables the establishment of taxonomic thresholds that are useful not only for the classification of cultured microorganisms but also for the classification of the many environmental sequences. In this Analysis article, we have shown that a reliable hierarchy of high taxa requires 16S rRNA gene sequences that are longer than 1,300 nucleotides, which is not achieved in current high-throughput 16S rRNA surveys. The calculated thresholds enabled us to circumscribe an enormous set of taxa that are yet to be formally classified; for example, the ~2,000 genera that have previously been described make up ~2% of the estimated ~1×10^5 genera that are likely to exist. However, we predict that, with the current sequencing efforts, most taxa will be detected before the end of this decade, and so the Sisyphean task of describing the extent of the diversity of bacterial and archaeal high taxa now seems to be achievable. In this regard, CTU circumscriptions and the implementation of a logical (that is, human-readable) and structured (that is, machine-readable) taxonomic nomenclature will help to reconcile the perceptions of microbial ecologists and taxonomists about the diversity of bacteria and archaea. By providing explicit and well-documented guidelines for this effort, we think that our analysis should facilitate the implementation of the many changes in the current taxonomy that are necessary to develop a common taxonomic classification of high taxa of bacteria and archaea based on 16 rRNA gene sequences.

1. Godfray, H. C. J. Challenges for taxonomy. *Nature* **417**, 17–19 (2002).
2. Mora, C., Tittensor, D. P., Adl, S., Simpson, S. G. B. & Worm, B. How many species are on Earth and in the ocean. *PLoS Biol.* **9**, e1001127 (2011).
3. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
This paper reports the SILVA project, which is a comprehensive web resource (see Further information) for up-to-date, quality-controlled databases of aligned rRNA gene sequences from the Bacteria, the Archaea and the Eukarya.
4. Mole, B. Microbiome research goes without a home. *Nature* **500**, 16–17 (2013).
5. Amann, R. I., Ludwig, W. & Schleifer, K. H. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev.* **59**, 143–169 (1995).
6. Rosselló-Móra, R. Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. *Environ. Microbiol.* **14**, 318–334 (2012).
7. Dykhuizen, D. E. Santa Rosalia revisited: why are there so many species of bacteria? *Antonie Van Leeuwenhoek* **73**, 25–33 (1998).
8. Richter, M. & Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl Acad. Sci. USA* **106**, 19126–19131 (2009).

9. Stackebrandt, E. & Ebers, J. Taxonomic parameters revisited: tarnished gold standards. *Microbiol. Today* **8**, 6–9 (2006).
10. Tindall, B. J., Rosselló-Móra, R., Busse, H.-J., Ludwig, W. & Kämpfer, P. Notes on the characterization of prokaryote strains for taxonomic purposes. *Int. J. Syst. Evol. Microbiol.* **60**, 249–266 (2010).
11. Philippot, L. *et al.* The ecological coherence of high bacterial taxonomic ranks. *Nature Rev. Microbiol.* **8**, 523–529 (2010).
This study demonstrates that high bacterial taxa (that is, genus and above) are ecologically meaningful and their coherence is inversely correlated to their taxonomic rank. These observations provide a new perspective for the study of bacterial taxonomy, evolution and ecology.
12. Gribaldo, S. & Brochier-Armanet, C. Time for order in microbial systematics. *Trends Microbiol.* **20**, 209–210 (2012).
13. Garrity, G. M. & Oren, A. Response to Gribaldo and Brochier-Armanet: time for order in microbial systematics. *Trends Microbiol.* **20**, 353–354 (2012).
In this paper, the International Committee on Systematics of Prokaryotes (ICSP) supports a call for order in microbial systematics to address the lack of criteria to circumscribe high taxa, which represents a major problem in microbiology today.
14. Ereshefsky, M. Some problems with the Linnaean hierarchy. *Phylos. Sci.* **61**, 186–205 (1994).
15. Yarza, P. *et al.* The All-Species Living Tree Project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst. Appl. Microbiol.* **31**, 241–250 (2008).
This paper reports the All-Species Living Tree Project (LTP), which is an initiative of *Systematic and Applied Microbiology* for the creation and maintenance of highly curated 16S rRNA and 23S rRNA gene sequence databases, alignments and phylogenetic trees for all the type strains of bacteria and archaea.
16. Cole, J. R., Konstantinidis, K., Farris, R. J. & Tiedje, J. M. in *Environmental molecular microbiology* (eds Liu, W.-T. & Jansson, J. K.) 1–19 (Caister Academic Press, 2010).
17. Ludwig, W., Klenk, H.-P. in *Bergey's Manual of Systematic Bacteriology* 2nd edn (eds Boone, D. R., Castenholz, R. W. & Garrity, G. M.) 49–65 (Springer, 2001).
18. Ludwig, W. in *Molecular Phylogeny of Microorganisms* (eds Oren, A. & Papke, R. T.) 65–83 (Caister Academic Press, 2010).
This chapter reports the classification of high ranks of the Bacteria and the Archaea, which is currently based on comparative analyses of rRNA and is supported by other markers and multigene approaches. The high information content and great availability in databases mostly justify the usage of rRNA gene sequences in taxonomy.
19. Fox, G. E., Pechman, K. R. & Woese, C. R. Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. *Int. J. Syst. Bacteriol.* **27**, 44–57 (1977).
20. Ludwig, W. & Schleifer, K. H. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiol. Rev.* **15**, 155–173 (1994).
21. Van de Peer, Y., Chapelle, S. & Wachter, R. D. A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Res.* **24**, 3381–3391 (1996).
22. Fuchs, B. M. *et al.* Flow cytometric analysis of the *in situ* accessibility of *Escherichia coli* 16S rRNA for fluorescently labeled oligonucleotide probes. *Appl. Environ. Microbiol.* **64**, 4973–4982 (1998).
23. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).
24. Yarza, P. *et al.* Update of the all-species living tree project based on 16S and 23S rRNA sequence analyses. *Syst. Appl. Microbiol.* **33**, 291–299 (2010).
25. Shida, O., Takagi, H., Kadowaki, K. & Komagata, K. Proposal for two new genera, *Brevibacillus* gen. nov. and *Aneurinibacillus* gen. nov. *Int. J. Syst. Bacteriol.* **46**, 939–946 (1996).
26. Kang, S.-J. *et al. Brevundimonas naejangsanensis* sp. nov., a proteolytic bacterium isolated from soil, and reclassification of *Mycoplana bullata* into the genus *Brevundimonas* as *Brevundimonas bullata* comb. nov. *Int. J. Syst. Evol. Microbiol.* **59**, 3155–3160 (2009).
27. Keswani, J. & Whitman, W. B. Relationship of 16S rRNA sequence similarity to DNA hybridization in prokaryotes. *Int. J. Syst. Evol. Microbiol.* **51**, 667–678 (2001).
28. Chakravorty, S., Helb, D., Burday, M., Connell, N. & Alland, D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J. Microbiol. Methods* **69**, 330–339 (2007).
29. Mizrahi-Man, O., Davenport, E. R. & Gilad, Y. Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *PLoS ONE* **8**, e53608 (2013).
30. Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J. & Weightman, A. J. 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 (2005).
31. Saddler, G. S. International Committee on Systematics of Prokaryotes. Xth International (IUMS) Congress of Bacteriology and Applied Microbiology. Minutes of the meetings, 28 and 30 July 2002, Paris, France. *Int. J. Syst. Evol. Microbiol.* **55**, 533–537 (2005).
32. Ritalahti, K. M. *et al. Sphaerochaeta globosa* gen. nov., sp. nov. and *Sphaerochaeta pleomorpha* sp. nov., free-living, spherical spirochaetes. *Int. J. Syst. Evol. Microbiol.* **62**, 210–216 (2012).
33. Meier-Kolthoff, J. P., Göker, M., Spröer, C. & Klenk, H. P. When should a DDH experiment be mandatory in microbial taxonomy? *Arch. Microbiol.* **195**, 413–418 (2013).
34. Curtis, T. P., Sloan, W. T., & Scannell, J. W. Estimating prokaryotic diversity and its limits. *Proc. Natl Acad. Sci. USA* **99**, 10494–10499 (2002).
35. Salman, V., Amann, R., Shub, D. A. & Schulz-Vogt, H. N. Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria. *Proc. Natl Acad. Sci. USA* **109**, 4203–4208 (2012).
36. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1 (2013).
37. Wang, T. *et al.* Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *ISME J.* **6**, 320–329 (2012).
38. Guazzaroni, M.-E. *et al.* Metaproteogenomic insights beyond bacterial response to naphthalene exposure and bio-stimulation. *ISME J.* **7**, 122–136 (2013).
39. Stackebrandt, E., Rainey, F. A. & Ward-Rainey, N. L. Proposal for a new hierarchic classification system, Actinobacteria classis nov. *Int. J. Syst. Bacteriol.* **47**, 479–491 (1997).
40. Lee, K. C. Y. *et al.* Phylogenetic delineation of the novel phylum Armatimonadetes (former candidate division OP10) and definition of two novel candidate divisions. *Appl. Environ. Microbiol.* **79**, 2484–2487 (2013).
41. Cavalier-Smith, T. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int. J. Syst. Evol. Microbiol.* **52**, 7–76 (2002).
42. Buchanan, R. E. Studies in the nomenclature and classification of the bacteria: II. The primary subdivisions of the Schizomycetes. *J. Bacteriol.* **2**, 155–164 (1917).
43. Hovind-Hougen, K. Leptospiraceae, a new family to include *Leptospira Noguchi* 1917 and *Leptonema* gen. nov. *Int. J. Syst. Bacteriol.* **29**, 245–251 (1979).
44. Swellengrebel, N. H. Sur la cytologie comparée des spirochètes et des spirilles. *Ann. Inst. Pasteur.* **21**, 562–586 (in French) (1907).
45. Gupta, R. S., Mahmood, S. & Adeolu, M. A phylogenomic and molecular signature based approach for characterization of the phylum Spirochaetes and its major clades: proposal for a taxonomic revision of the phylum. *Front. Microbiol.* **4**, 217 (2013).
46. Harris, J. K., Kelley, S. T. & Pace, N. R. New perspective on uncultured bacterial phylogenetic division OP11. *Appl. Environ. Microbiol.* **70**, 845–849 (2004).
47. Giebel, H.-A. *et al.* Distribution of *Roseobacter* RCA and SAR11 lineages in the North Sea and characteristics of an abundant RCA isolate. *ISME J.* **5**, 8–19 (2011).
48. Wagner, M. & Horn, M. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr. Opin. Biotechnol.* **17**, 241–249 (2006).
49. Guy, L. & Ettema, T. J. G. The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**, 580–587 (2011).
50. Hugenholtz, P., Goebel, B. M. & Pace, N. R. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**, 4765–4774 (1998).
51. Yilmaz, P. *et al.* The SILVA and 'All-species Living Tree Project (LTP)' taxonomic frameworks. *Nucl. Acids Res.* **42**, D643–D648 (2013).
52. Geissinger, O., Herlemann, D. P. R., Mörschel, E., Maier, U. G. & Brune, A. The ultramicrobacterium '*Elusimicrobium minutum*' gen. nov., sp. nov., the first cultivated representative of the Termite Group 1 phylum. *Appl. Environ. Microbiol.* **75**, 2831–2840 (2009).
53. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2014) [online], www.r-project.org/
54. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
55. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).

Acknowledgements
This work has been co-funded by the Max Planck Society and the European Union (EU) project SYMBIOMICS (grant number 264774). R.R.M. acknowledges the scientific support given by the Spanish Ministry of Economy with the projects CE-CSD2007-0005 and CGL2012-39627-C03-03, which are both also supported with European Regional Development Fund (FEDER) funds, and the preparatory phase of Microbial Resource Research Infrastructure (MIRRI) funded by the EU (grant number 312251). W.B.W. acknowledges support of the Dimensions in Biodiversity program at the US National Science Foundation (NSF). P.Y. acknowledges support of the EU's Seventh Framework Program funds BioVeL, grant no. 283359.

Competing interests statement
The authors declare no competing interests.

DATABASES
European Nucleotide Archive (ENA):
http://www.ebi.ac.uk/ena/
X87140 | AF357916 | EU081285 | FR733665 | AB364473 | GU993264 | Z22781 | Z12817 | AY714984 | AY293856

FURTHER INFORMATION
Bacteriological Code (maintained by the International Committee on Systematics of Prokaryotes): http://icsp.org/
International Nucleotide Sequence Database Collaboration (INSDC): http://www.insdc.org/
List of Prokaryotic Names with Standing in Nomenclature (LPSN): http://www.bacterio.net
Living Tree Project: http://www.arb-silva.de/projects/living-tree
SILVA project: http://www.arb-silva.de
xrna Program: http://rna.ucsc.edu/rnacenter/xrna/

SUPPLEMENTARY INFORMATION
See online article: S1 (table) | S2 (box) | S3 (figure) | S4 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF