A rank-normalized archaeal taxonomy based on genome phylogeny resolves widespread incomplete and uneven classifications

Christian Rinke[1], Maria Chuvochina[1], Aaron J. Mussig[1], Pierre-Alain Chaumeil[1], David W. Waite[2], William B Whitman[3], Donovan H. Parks[1], Philip Hugenholtz[1]

[1] Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia

[2] School of Biological Sciences, The University of Auckland, New Zealand

[3] Department of Microbiology, University of Georgia Athens, GA, USA

Abstract

The increasing wealth of genomic data from cultured and uncultured microorganisms provides the opportunity to develop a systematic taxonomy based on evolutionary relationships. Here we propose a standardized archaeal taxonomy, as part of the Genome Taxonomy Database (GTDB), derived from a 122 concatenated protein phylogeny that resolves polyphyletic groups and normalizes ranks based on relative evolutionary divergence. The resulting archaeal taxonomy is stable under a range of phylogenetic variables, including marker genes, inference methods, and tree rooting scenarios. Taxonomic curation follows the rules of the International Code of Nomenclature of Prokaryotes (ICNP) while taking into account proposals to formally recognise the rank of phylum and to use genome sequences as type material. The taxonomy is based on 2,392 quality screened archaeal genomes, the great majority of which (93.3%) required one or more changes to their existing taxonomy, mostly as a result of incomplete classification. In total, 16 archaeal phyla are described, including reclassification of three major monophyletic units from the Euryarchaeota and one phylum resulting from uniting the TACK superphylum into a single phylum. The taxonomy is publicly available at the GTDB website (https://gtdb.ecogenomic.org).

Main

Carl Woese's discovery of the Archaea in 1977, originally termed Archaebacteria (Woese and Fox, 1977) gave rise to the recognition of a new domain of life and fundamentally changed our view of cellular evolution on Earth. In the following decades, an increasing number of Archaea were described, initially from extreme

environments but subsequently also from soils, oceans, freshwater, and animal guts, highlighting the global importance of this domain (Gribaldo and Brochier-Armanet, 2006). Since their recognition, Archaea have been classified primarily via genotype, i.e. small subunit (SSU) rRNA gene sequences, and hence, compared to Bacteria, they suffer less from historical misclassifications based on phenotypic properties (Zuo et al., 2015). Using the SSU rRNA gene, Woese initially described two major lines of archaeal descent, the Euryarchaeota and the Crenarchaeota (Woese et al., 1990), and in the following years all newly discovered archaeal lineages were added to these two main groups. Non-extremophile Archaea were generally classified as Euryarchaeota (Spang et al., 2017), which led to a considerable expansion of this lineage. Eventually two new archaeal phyla were proposed based on phylogenetic novelty of their SSU rRNA sequences; the Korarchaeota (Barns et al., 1996) recovered from hot springs in Yellowstone National Park, and the nanosized, symbiotic Nanoarchaeota (Huber et al., 2002) co-cultured from a submarine hot vent. By the late 2000s, archaeal classification had begun to leverage genome sequences and the first genome sequence of a crenarchaeote, *Cenarchaeum symbiosum*, was used to argue that mesophilic archaea are different from hyperthermophilic Crenarchaeota and should be considered as a separate phylum for which the name *Thaumarchaeota* was proposed (Brochier-Armanet et al., 2008). Subsequently the field experienced a burst in availability of genomic data due to the substantial acceleration of culture-independent genome recovery driven by improvements in high throughput sequencing (Adam et al., 2017; Spang et al., 2017). This resulted in the description and naming of several new archaeal lineages, including the candidate phyla Aigarchaeota (Nunoura et al., 2011), Geoarchaeota (Kozubal et al., 2013) and Bathyarchaeota (Meng et al., 2014), previously reported as unclassified Crenarchaeota based on SSU rRNA data. Some of the proposed phyla were met with criticism, such as the Geoarchaeota, which was considered to be a member of the order Thermoproteales rather than a novel phylum (Guy et al., 2014). All former crenarchaeal lineages (Adam et al., 2017) were then recombined into the TACK superphylum, originally comprising the Thaumarchaeota, Aigarchaeota, the remaining Crenarchaeota, and Korarchaeota (Guy and Ettema, 2011), and more recently the Verstraetearchaeota (Vanwonterghem et al., 2016).

New archaeal lineages were also described outside the Euryarchaeota and TACK belonging to two superphyla, DPANN and Asgard (Rinke et al., 2013; Zaremba-Niedzwiedzka et al., 2017). DPANN was originally proposed based on five phyla; Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota (Rinke et al., 2013; Zaremba-Niedzwiedzka et al., 2017)), but now also includes the Micrarchaeota, Woesearchaeota, Pacearchaeota, Altiarchaeota and Huberarchaeota (Adam et al., 2017; Baker et al., 2010; Castelle et al., 2015; Probst et al., 2018). The Asgard archaea are notable for their inferred sister relationship to the eukaryotes and were originally proposed to based on the phyla Lokiarchaeota, Thorarchaeota, Odinarchaeota and Heimdallarchaeota ((Rinke et al., 2013; Zaremba-Niedzwiedzka et al., 2017), followed by the Helarchaeota (Seitz et al., 2019). The net result of these cumulative activities is that archaeal classification at higher ranks is currently very uneven. The Euryarchaeota absorbed novel lineages and grew into a phylogenetic behemoth, whereas the Crenarchaeota were split into multiple subordinate taxa including several shallow lineages, e.g. Geoarchaeota, which despite their phylogenetic disproportion have been given the same rank of phylum. Attempts to rectify this taxonomic bias included a proposal to reclassify TACK as a single phylum termed Proteoarchaeota (Petitjean et al., 2014) and to introduce a new taxonomic

rank above the class level which would generate several superclasses within the Euryarchaeota (Adam et al., 2017; Petitjean et al., 2015).

Most of this activity, i.e. the proposing and naming of new phyla, has occurred outside of the International Code of Nomenclature of Prokaryotes (ICNP; (Parker et al., 2019) because the ICNP does not yet extend to uncultured microorganisms nor recognise the ranks of phylum and superphylum. However, proposals have been made to include the rank of phylum (Oren et al., 2015) and to allow gene sequences to serve as type material (Whitman, 2016), which have been increasingly adopted informally by the research community (Chuvochina et al., 2019). Currently, uncultured prokaryotes can be provisionally named using *Candidatus* status (Murray and Stackebrandt, 1995). However, these names have no formal standing in nomenclature, do not have priority, and often contradict the current ICNP rules or are otherwise problematic (Oren, 2017). The waters have been further muddied by the proposal of names for higher taxa without designation of lower ranks and type material, which leaves them without nomenclatural anchors and their circumscription subject to dispute (Chuvochina et al., 2019). In addition to these higher level classification issues, the current archaeal taxonomy suffers from the same phylogenetic inconsistencies observed in the Bacteria, such as polyphyletic taxa (e.g. class Methanomicrobia; see below), but to a lesser degree than the Bacteria due to the early integration of phylogeny and the relatively small size of the archaeal dataset. More problematic is the widespread incomplete classification of environmental archaeal sequences in the NCBI taxonomy, which are often only assigned to a candidate phylum with no subordinate rank names. This high degree of incomplete classification is likely due to a natural hesitancy to create novel genera and intermediate taxa for groups lacking isolated representatives.

For the domain Bacteria all of these long standing taxonomic issues and inconsistencies were recently addressed by proposing a standardized taxonomy referred to as the Genome Taxonomy Database (GTDB; gtdb.ecogenomic.org). The GTDB normalizes rank assignments using relative evolutionary divergence (RED) in a genome phylogeny, followed by an extensive automated and manual taxonomy curation process. This approach resulted in changes to the taxonomy of 58% of the nearly 95,000 analyzed bacterial genomes (Parks et al., 2018) and has recently been extended to a complete classification from domain to species (Parks et al., 2019). Here we present the GTDB taxonomy for the domain Archaea (release R04-RS89), comprising 2,392 quality screened genomes from cultivated and uncultured organisms. This taxonomic release circumscribes 16 phyla, including 3 phyla from major monophyletic units of the Euryarchaeota, and one phylum resulting from the amalgamation of the TACK superphylum. The archaeal GTDB taxonomy is publicly available at the Genome Taxonomy Database website (https://gtdb.ecogenomic.org/).

Results

Reference genome tree and initial decoration with the NCBI taxonomy

The archaeal Genome Taxonomy Database (GTDB; 04-RS89) currently comprises 2,392 quality-filtered genomes obtained from RefSeq/GenBank release 89 (*see Methods*, (Haft et al., 2018). Genomes were clustered into species units based on average nucleotide identity (ANI; *see Methods*), resulting in 1,248 species (Parks et al., 2019). A representative sequence of each species, i.e. the genome of the type strain or the highest quality genome of an uncultivated species, was then used for phylogenomic analyses (**Table S1**). The protein sequences for up to 122 conserved single-copy ubiquitous archaeal genes were recovered from each genome (Parks et al., 2017); **Table S2, Fig. S1**), aligned, concatenated into a supermatrix, and trimmed to 5124 columns (*see Methods*). The archaeal GTDB genome tree (ar122.r89) was inferred using the C10 protein mixture model with the posterior mean site frequency (PMSF) approximation (Wang et al., 2018) implemented in IQ-TREE (Nguyen et al., 2015). The PMSF model is a faster approximation of finite mixture models, which capture the heterogeneity in the amino acid substitution process between sites, and can effectively resolve long-branch attraction artefacts (Wang et al., 2018). This approach was subsequently compared to trees calculated from different alignments and with alternative inference methods (*see below*). The archaeal genome tree was initially decorated with taxon names obtained from the NCBI taxonomy (Federhen, 2012) standardized to seven canonical ranks as previously described (Parks et al., 2018). Strikingly, many archaeal genomes had no canonical rank information beyond their phylum affiliation (31.0%), which is partly offset by the extensive use of non-canonical names (*i.e.* names with no rank) in the NCBI taxonomy (**Fig. S2a**). This was particularly the case for DPANN phyla, which almost entirely lack information in the family to class ranks (**Fig. S2b**).

Removal of polyphyletic groups and rank normalization

More than a fifth of NCBI-defined taxa above the rank of species (21.4%; 54 of 252) could not be reproducibly resolved as monophyletic or operationally monophyletic in the ar122.r89 tree (defined as having an F measure $\geq$0.95; Parks et al., 2018; **Table S3**). These include the phyla Nanoarchaeota, Aenigmarchaeota and Woesearchaeota, which were intermingled with each other and with unclassified archaeal genomes (**Fig. S3**). Another prominent example is the class Methanomicrobia, which comprised three orders in NCBI, two of which (Methanosarcinales and Methanocellales) were clearly separated from the type order Methanomicrobiales (**Fig. S4.**). To resolve these polyphyletic cases, the lineage containing the nomenclature type retained the name. Where possible, all other groups were renamed following the International Code of Nomenclature of Prokaryotes **(**ICNP) and recent proposals to modify the Code (*see Methods*). For example, the Methanomicrobia were resolved in GTDB by reserving the name for the lineage containing the type genus of the type order of the class, i.e. *Methanomicrobium*, and by reclassifying the two remaining orders into their own classes; Methanosarcinia class. nov. and Methanocellia class. nov. (**Fig. S4; Table S13**). When nomenclature types were not available, the existing names were retained as placeholders with alphabetical suffixes indicating polyphyly. For example, the genus *Thermococcus* is polyphyletic because it also comprises

species of the genus *Pyrococcus.* This polyphyly was resolved by retaining the name for the monophyletic group containing the type species, *Thermococcus celer,* and by assigning alphabetical suffixes to three basal groups comprising other *Thermococcus* genomes (**Fig. S5**). Note that *Thermococcus chitonophagus* was transferred to *Pyrococcus* according to the GTDB reclassification due to its proximity to the type species of this genus (**Fig. S5**).

Taxonomic ranks were normalized using the relative evolutionary divergence (RED) approach. This method linearly interpolates the inferred phylogenetic distances between the last common ancestor (set to RED=0) and all extant taxa (RED=1) providing an approximation of relative time of divergence (Parks et al., 2018). Rank distributions in the NCBI-decorated ar122.r89 tree were extremely broad, highlighting severely under- and over-classified outlier taxa (**Fig. 1a)**. Rank distributions were normalized (Parks et al., 2018) by systematically reclassifying outliers either by re-assignment to a new rank with associated nomenclatural changes for latin names, or by moving names to new interior nodes in the tree (**Fig. 1b)**. This resulted in large movements of the median RED values of the higher ranks (order and above) to produce a normalized distribution (**Fig. 1b**). In the end, 56.4% of all archaeal NCBI taxa names had to be changed in the GTDB taxonomy, with the largest percentage of changes occurring at the phylum level (76.6%) (**Fig. 1c**). Examples of changes at lower ranks due to RED normalization include the genus *Methanobrevibacter* which was divided into five genus-level groups; *Methanobrevibacter* which includes the type species and four genera with alphabetical suffixes (*Methanobrevibacter_A, Methanobrevibacter_B*, *etc*; **Fig. 1a; Fig S6**). GTDB names were assigned to nodes with high bootstrap support (bs 98.5% ± 5.0%) to ensure taxonomic stability with a small number of exceptions (bs <90%) to preserve existing classifications (**Table S4**). Overall, 93.3% of the 2,239 archaeal genomes present in NCBI release 89 had one or more changes in their taxonomic assignments (**Fig. 1c**).

Robustness of proposed archaeal taxonomy

We next tested the robustness of the monophyly and rank normalization of the proposed taxonomy in relation to a number of standard phylogenetic variables: marker genes, inference method, compositional bias, and rooting of the tree. Comparing tree similarities indicated that marker choice had a stronger influence on the tree topology than inference methods and substitution models (**Fig. S7; Fig. S8)**. However, for all subsequent comparisons we focused on the robustness of the taxonomy, not the overall consistency of tree topologies, as only a subset of interior nodes (69.9%) in the ar122.r89 tree were used for taxon classification (**Table S5**).

Markers

As expected, individual protein phylogenies of the 122 markers had lower phylogenetic resolution than the supermatrix ar122.89 tree, particularly at the higher ranks of class and phylum (**Fig S9a, b**). However, 78.5% of the GTDB taxa with ≥2 representatives above the rank of species were still recovered as monophyletic groups in ≥50% of single protein trees, and on average taxa were resolved as monophyletic in 74.1% of the single protein trees (**Fig 9c, d**). We also compared the GTDB taxonomy to SSU rRNA gene trees, due to their historical importance in defining archaeal taxa. However, this was complicated by the absence of this gene

(>900 nt after quality-trimming) in almost half of the species representatives (578/1248; 46.3%). The majority (94.1%) of species representatives lacking SSU rRNA sequences were draft MAG assemblies (**Table S6**), which often lack this gene due to the difficulties of correctly assembling and binning rRNA repeats in metagenomic datasets (Hugenholtz et al., 2016; Parks et al., 2017). Over 84% of the GTDB taxa with ≥2 SSU rRNA representatives were operationally monophyletic in the SSU rRNA tree, with loss of monophyly most pronounced in the higher ranks (**Fig. S10**) as seen for the single protein phylogenies. For alternate concatenated protein marker datasets, nearly all (≥97.7%) GTDB taxa above the rank of genus with ≥2 representatives were recovered as monophyletic groups in trees calculated from either 16 (rp1; **Table S7)** (Hug et al., 2016) or 23 ribosomal proteins (rp2; **Table S8**; **Fig. 4a; Fig.S11**) (Rinke et al., 2013). The small percentage of taxa not resolved in rp1 and/or rp2 trees (**Fig S11; Table S9**) were well supported in the ar122.89 tree (average bootstrap support 90.3±8.8% ;**Fig. S12**). However, these taxa were resolved as operationally monophyletic in a lower proportion of individual protein trees than other GTDB taxa (33.7±14.7% vs 77.1±18.3% respectively). The RED distributions of the GTDB taxa were comparable in the ar122.89, rp1 and rp2 trees, whereas the SSU rRNA tree had a substantially broader distribution (**Fig. 2a,b, Fig. S13**), likely reflecting both the undersampling of the topology and lower resolution of single marker genes relative to concatenated marker sets.

Inference methods and models

We overlaid the IQ-TREE-based taxonomy onto trees inferred with different models (**Table S10**) and phylogenetic tools, including FastTree and ExaML (maximum likelihood), PhyloBayes (Bayesian), and ASTRAL (supertree). All methods used the 122 marker set with the exception of the ASTRAL supertree, which was also applied to a 253 marker set, subsampled from the PhyloPhlAn dataset (Segata et al., 2013). Overall, the GTDB taxonomy was remarkably consistent with comparable RED distributions for taxa at each investigated rank (**Fig. 2a,c; Fig. S13**). All GTDB taxa were recovered as monophyletic or operationally monophyletic (**Fig S11**) for all supermatrix methods, regardless of the underlying inference algorithm or model. The ASTRAL supertrees, inferred by the most divergent approach tested, recovered 96% of GTDB taxa with ≥2 representatives as monophyletic or operationally monophyletic (**Fig. S11**). The only major inconsistency affected the Euryarchaeota, with 48 taxa being placed outside of this phylum in both supertrees (**Fig S14e, Fig S15**).

Compositional bias

To test for possible compositional bias we employed tools for character trimming and for clustering of high confidence positional homology (*see Methods*) which have been shown to alleviate long branch attraction artifacts and to increase tree accuracy for alignments of distantly-related sequences (Ali et al., 2019; Criscuolo and Gribaldo, 2010). Decorated ML trees calculated from the trimmed or clustered alignments were in strong taxonomic agreement with the ar122.89 tree, showing comparable RED values and recovering >98% of GTDB taxa (with ≥2 representatives) as monophyletic or operationally monophyletic (**Fig. 2d; Fig. S13; Fig S11**). Less than 3.9% of genomes had conflicting taxonomic assignments at any rank above species (**Table S11**), with the largest difference being observed in the class Methanosarcinia (18 taxa) and the order

Desulfurococcales (46 taxa) for the character trimmed alignment (**Fig S14d**). The clustered alignment resulted in fewer differences, with a total of seven conflicting taxa (**Fig S14d**).

Rooting effects

Rank normalization is sensitive to the placement of the tree root, which defines the last common ancestor (set to RED=0) (Parks et al., 2019)(Parks et al., 2018)(Parks et al., 2019), and can therefore potentially influence the resulting taxonomy. Since the rooting of both the archaeal and bacterial domains remains contested, GTDB uses an operational approach whereby the median of multiple plausible rootings is considered (Parks et al., 2018). We assessed the effect of fixed rooting of the archaeal domain on the taxonomy by testing two recently proposed archaeal root placements; the first, within the Euryarchaeota (as currently defined by NCBI) (Raymann et al., 2015) and the second, between the DPANN superphylum and all other Archaea (Williams et al., 2017). Overall, RED values were stable across the tested rooting scenarios and the intervals defining taxonomic ranks in GTDB were largely preserved (**Fig. 3**). As expected, a fixed root caused the taxa within the rooted lineages to be drawn slightly closer to the root, although still within the RED assigned rank interval (**Fig S16; Fig. 3**).

Proposal of new and revised taxa based on the GTDB taxonomy.

After resolving polyphyletic groups and normalizing ranks, the archaeal GTDB taxonomy (release R04-RS89) comprises 16 phyla, 36 classes, 96 orders, 238 families, 534 genera and 1248 species (**Table S12**). This entailed the proposal of 13 new taxa and 33 new candidatus taxa above the rank of genus including five novel species combinations and three novel candidatus species combinations (**Tables S13 to S16**). We also used 25 Latin names without standing in nomenclature as placeholders in the GTDB taxonomy (**Table S17)** to preserve literature continuity. The extensive rearrangement of phyla to normalize phylogenetic depth resulted in both division and amalgamation of release 89 NCBI phyla (**Fig. 4**). For example, the phylum Euryarchaeota was divided into six separate phyla in the GTDB taxonomy due to its anomalous depth (**Fig. 1a; Fig 4**). The name Euryarchaeota has been retained as it is deeply ingrained in the literature (Cavicchioli, 2011) and assigned to the GTDB phylum containing the majority of taxa first described as euryarchaeotes, namely the classes Thermococci, Methanobacteria and Methanococci (**Fig. 4**; originally classified as orders by (Woese et al., 1990)). However, we note that the name Euryarchaeota would be illegitimate if the rank of phylum is introduced into the Code as a type was not designated in the original proposal (Woese et al., 1990), which is one of the requirements for validation of a name. Thus, in the future, this phylum could be named Methanobacteriota as recommended by Whitman (Whitman et al., 2018). This new name would also avoid confusion because the circumscription of the new phylum is significantly different from that of the original Euryarchaeota. Most of the remaining NCBI euryarchaeotes were assigned to two new phyla with names derived from the oldest validly published class in each group using the recently proposed standardised phylum suffix -ota (Whitman et al., 2018); the Thermoplasmatota phyl. nov. from the class Thermoplasmata (Anna-Louise Reysenbach, 2001), and the Halobacteriota phyl. nov. from the class Halobacteria (**Table S18**) (Grant et al., 2001). The recently proposed Hydrothermarchaeota and Hadarchaeota (Baker et al., 2016; Chuvochina

et al., 2019; Jungbluth et al., 2017) and Altiarchaeota (Probst et al., 2018) comprise the other three phyla previously classified as euryarchaeotes (**Fig. 4**).

The TACK superphylum (Guy and Ettema, 2011) was reclassified as a single phylum based on rank normalization and its robust monophyly, for which we retain Crenarchaeota as a placeholder name (**Fig. 4**). However, we recognize that this name would also be illegitimate under the proposal of Whitman (Whitman et al., 2018) and Thermoproteota would be a suitable alternative. The RED-based reorganisation of this lineage requires the demotion of the NCBI-defined TACK phyla to lower-level lineages, as previously proposed (Petitjean et al., 2014). Thus the Thaumarchaeota was reclassified as a class-level lineage, for which we propose an emended description of the only validly described class in this lineage, the Nitrososphaeria (Stieglmeier et al., 2014) (**Fig. 5).** According to its RED value and concatenated protein phylogeny, the Aigarchaeota constitutes an order-level lineage within the Nitrososphaeria, for which we propose the order Caldarchaeales ord. nov. (**Table S15**) derived from the Candidatus species *Caldiarchaeum subterraneum* (Nunoura et al., 2011). We propose to reclassify the NCBI-defined Crenarchaeota as an emended description of the validly described class Thermoprotei (A-L Reysenbach, 2001), and the Korarchaeota as the Korarchaeia class. nov. (**Table S15**) based on the type species *Korarchaeum cryptofilum* (Elkins et al., 2008). Two additional class-level lineages are included within the redefined phylum Crenarchaeota based on their phylogenetic affiliation; Methanomethylicia class. nov. (**Table S15**) derived from *Candidatus* Methanomethylicus mesodigestum replacing the originally proposed phylum name for this lineage *Verstraetearchaeota* (Vanwonterghem et al., 2016), and the former phylum Bathyarchaeota (Meng et al., 2014) identified by the placeholder name Bathyarchaeia until type material is assigned for this lineage (Chuvochina et al., 2019).

Similar to the TACK superphylum, the recently described Asgard superphylum (Zaremba-Niedzwiedzka et al., 2017) was reclassified as a single phylum for which we use Asgardarchaeota as a placeholder name to retain its identity until type material has been proposed for this lineage (Chuvochina et al., 2019). By contrast, the DPANN superphylum (Rinke et al., 2013) still comprises multiple (nine) phyla after rank normalization (**Fig. 4**). However, many of these phyla are represented by only a few genomes on long branches often reflected by low bootstrap support (**Fig. S17**). We envision that the DPANN phyla will undergo further reclassification in future GTDB releases when additional genomes become available to populate this region of the tree.

Unlike the major changes required at the phylum-level, archaeal taxa with lower rank information (species to class) were more stable, with an average of only 11% name changes. Notable examples include the well known genus *Sulfolobus*, which was reported early on as potentially polyphyletic (Fuchs et al., 1996) and is comprised of strains differing in their metabolic repertoire and genome size (Quehenberger et al., 2017). The concatenated protein tree confirms that *Sulfolobus* is not monophyletic, being interspersed with species belonging to the genera *Acidanus, Metallosphaera* and *Sulfurisphaera.* We resolved this polyphyly by dividing the group into four separate genera (**Fig. S18**) reflecting previously reported differences between *Sulfolobus* species, including a high number of transposable elements in *Sulfolobus_A* species (*S. islandicus* and *S.*

solfataricus), which are mostly absent in the type species of the genus, *Sulfolobus acidocaldarius* (Quehenberger et al., 2017). An example of a more complicated situation is the taxonomy of genera belonging to the halophilic family Natrialbaceae. Some of these genera have been reported as polyphyletic such as *Natrinema* and *Haloterrigena* (Minegishi et al., 2010), or are polyphyletic in published trees without associated comment, including *Halopiger* (Sorokin et al., 2019) and *Natronolimnobius* (Sorokin et al., 2018). We confirmed this polyphyly in the concatenated protein tree, which required extensive reclassification guided by the type species of each genus (**Fig S19, Table S14)**.

Discussion

Here we present the Genome Taxonomy Database (GTDB) for the domain Archaea to provide a phylogenetically congruent and rank-normalized classification based on well-supported nodes in a phylogenomic tree of 122 concatenated conserved single copy marker proteins (**Table S2**). Unlike the bacterial GTDB reference tree comprising 23,458 species representatives (Parks et al., 2019), the relatively modest number of publicly available archaeal genomes representing only 1,248 species allowed us to use IQ-TREE with a protein mixture model that captures substitution site heterogeneity between sites and can resolve long branch attraction artifacts (Wang et al., 2018). The proposed taxonomy is stable under a range of standard phylogenetic variables, including alternate marker genes, inference methods, and tree rooting scenarios, due in part to the inherent flexibility of rank designations within RED intervals (**Fig. 1**). We endeavoured to preserve the existing archaeal taxonomy wherever possible, however the great majority (93.3%) of the 2,239 archaeal genomes in GTDB had one or more changes to their classification compared to the NCBI taxonomy (**Fig. 1c**). This high percentage of modifications, compared to 58% reported for genomes in the bacterial GTDB (Parks et al., 2018), could be attributed to extreme unevenness observed within archaeal ranks, particularly at the phylum level. For example, the division of the NCBI-defined Euryarchaeota alone affected over two thirds of all archaeal genomes (**Fig. 4**). Secondly, widespread missing rank information in NCBI r89, particularly amongst as-yet-uncultivated lineages, required numerous passive (rank filling) changes to the taxonomy (**Fig. 1**). And thirdly, the Bacteria have highly sampled species and genera that did not require taxonomic changes, including over 12,000 *Streptococcus* genomes in release 04-RS89, which effectively lowers the percentage of bacterial genomes with taxonomic differences to NCBI.

The archaeal GTDB taxonomy and associated tree and alignment files are available online (https://gtdb.ecogenomic.org) and via linked third party tools, e.g. AnnoTree (Mendler et al., 2019). Users can classify their own genomes against the archaeal GTDB taxonomy using GTDB-Tk (Chaumeil et al., 2019). We envisage that in the short term, the archaeal taxonomy will scale with genome deposition in the public repositories based on current accumulation rates (**Fig. S20**). However, in the longer term more efficient phylogenomic tools will be needed to scale with increasingly large genomic datasets and to allow for biologically meaningful phylogenetic inferences. We also expect that less stable parts of the concatenated

protein tree, e.g. the DPANN phyla, will become more robust with additional genome sequence representatives.

Methods

Genome dataset

For the archaeal GTDB taxonomy R04-RS89 we obtained 2,661 archaeal genomes from RefSeq/GenBank release 89 and augmented them with 187 MAGs derived from the Sequencing Read Archive (SRA, (Haft et al., 2018)), resulting in 2,848 genomes. This data set was refined by applying a quality threshold (completeness - 5x contamination >50% (Parks et al., 2019)), leaving 2,392 genomes to form species clusters (see below), resulting in a total 1,248 species representative genomes for the downstream analysis (**Table S1**).

Metadata

The NCBI taxonomy of all representative genomes of R04-RS89 was obtained from the NCBI Taxonomy FTP site on July 16th, 2018. The NCBI taxonomy was standardized to seven ranks (species to domain) by identifying missing standard ranks and filling these gaps with rank prefixes and by removing non-standard ranks. All standard ranks were prefixed with rank identifiers (e.g. “p__” for phylum) as previously described (McDonald et al., 2012).

Phylogenomic marker set

Archaeal multiple sequence alignments (MSAs) were created through the concatenation of 122 phylogenetically informative markers comprised of proteins or protein domains specified in the Pfam v27 or TIGRFAMs v15.0 databases. The 122 proteins were selected based on the criteria described in (Parks et al., 2017). In brief this included being present in ≥90% of archaeal genomes, and, when present, single-copy in ≥95% of genomes. Only genomes composed of ≤200 contigs with an N50 of ≥20 kb and with CheckM completeness and contamination estimates of ≥95% and ≤5%, respectively, were considered. Phylogenetically informative proteins were determined by filtering ubiquitous proteins whose gene trees had poor congruence with a set of subsampled concatenated genome trees (Parks et al., 2017). Gene calling was performed with Prodigal v2.6.3, and markers were identified and aligned using HMMER v3.1b1. To remove sites with weak phylogenetic signals, we created an amino acid alignment by trimming columns represented in <50% of the genomes and columns with less than 25% or more than 95% amino acid consensus, resulting in an initial 27,000 amino acid alignment (Suppl file). To reduce computational requirements, the alignment was further trimmed by randomly selecting 42 amino acids from the remaining columns of each marker (defined as min. consensus = 25%, max. consensus = 95%, minimum percentage of taxa required to retain column = 50%). The resulting ar122.r89 MSA included a total of 5124 (42*122) columns. This MSA filtering methodology is implemented in the `align` method of GTDB-Tk v1.0.2 (Chaumeil et al., 2019).

Alternative supermatrix marker sets

Alternative MSAs, resembling previously published datasets, were created through the concatenation of 16 ribosomal proteins, termed dataset rp1 (Hug et al., 2016) and 23 ribosomal proteins, termed rp2 (Rinke et al., 2013). After trimming columns represented by <50% of the genomes and with an amino acid consensus <25%, the resulting MSA spanned 1174 and 2377 amino acids for rp1 and rp2, respectively.

SSU rRNA gene

Archaeal SSU rRNA genes were identified from the 1,248 archaeal R04-RS89 GTDB genomes using nhmmer v3.1b2 (Wheeler and Eddy, 2013) with the SSU rRNA model (RF00177) from the RFAM database (Kalvari et al., 2018). Only the longest sequence was retained per genome. The resulting sequences were aligned with SSU-ALIGN 0.1.1 (Nawrocki, 2009), regions of low posterior probabilities which are indicative of high alignment ambiguity were pruned with ssu-mask (SSU-ALIGN 0.1.1), and the alignment was trimmed with anin-house script (trimSeqs.py v0.0.1; module bioscripts) to remove poorly represented leading and trailing positions along with short sequences below 900bp.

Species cluster

The 2,848 archaeal genomes were formed into species clusters as previously described (Parks et al., 2019). Briefly, a representative genome was selected for each of the 380 validly or effectively published archaeal species with one or more genomes passing quality control and genomes assigned to these representatives using average nucleotide identity (ANI) and alignment fraction (AF) criteria. Genomes not assigned to one of these 380 species were formed into 868 *de novo* species clusters, each specified by a single representative genome. Representative genomes were selected based on assembly quality with preference given to isolate genomes. Nonetheless, only 450 representatives were from isolate genomes with the remaining being MAGs (775) and SAGs (23).

Accounting for compositional bias

Each of the untrimmed 122 GTDB r89 archaeal single-copy marker protein alignments was filtered individually using BMGE 1.12 (Criscuolo and Gribaldo, 2010) and Divvier 1.0 (Ali et al., 2019). BMGE was executed using -t AA -s FAST -h 0.55 -m BLOSUM30, and Divvier was run using the recommended options: -divvy -mincol 4 -divvygap. Processing untrimmed protein alignments of individual markers ensures that all protein positions are considered when accounting for compositional bias. Next, each of the filtered marker gene alignments were concatenated into a single MSA supermatrix for BMGE and one for Divvier, respectively, whereby previously removed gap-only sequences were added again in the corresponding positions. Finally, the MSA was trimmed according to GTDB criteria mentioned above to a length of 7859 amino acids (BMGE) and 32061 amino acids (Divvier) and used for phylogenetic inferences (see below).

Phylogenetic inference

Phylogenomic trees were inferred with FastTree 2 (Price et al., 2010), ExaML (Kozlov et al., 2015), IQ-TREE (Nguyen et al., 2015), and PhyloBayes (Lartillot and Philippe, 2004) on different alignments and with a range

of models (**Table S10**). Note that we chose IQ-TREE as the GTDB standard inference, since it scales with our dataset, allows mixture models (see below), and because a previous study concluded that for concatenation-based species tree inference, IQ-TREE consistently achieved the best-observed likelihoods for all data sets, compared to RAxML/ExaML and FastTree (Zhou et al., 2018). More details about each inference program is provided below.

FastTree

FastTree v2.1.9 was executed in multithreaded mode with the WAG+GAMMA parameters.

IQ-TREE

IQ-TREE was executed employing mixture models, such as C10-C60 (Le et al. 2008), and a faster approximation of these models known as "posterior mean site frequency" (PMSF) model (Wang et al., 2018). True C10-C60 trees are computationally more demanding (**Table S10**), and we therefore opted for the faster PMSF model to calculate the ar122.r89 tree. The tree was calculated with IQ-TREE v1.6.12 based on the C10 mixture model and a starting tree (-ft), inferred by FastTree v2.1.9 as described above, to invoke the faster PMSF approximation with the following settings: *-m LG+C10+F+G -ft <starting tree>*.

ExaML

ExaML trees (gamma, JTT) were calculated from 10 different starting trees with random seeds using the mpi version 3.0.20 (settings: -m GAMMA), whereby the tree with the highest likelihood score was retained.

PhyloBayes

To accommodate the computationally demanding Bayesian inferences we subsampled our data set by reducing the number of taxa to one representative per genus resulting in 440 taxa. The genus representatives were selected by removing genomes with a quality score (CheckM completeness - 5*CheckM contamination) of <50; <50% of the 122 archaeal marker genes; an N50 <4 Kb; >2,500 contigs or >1,500 scaffolds). From the remaining genomes, the highest quality genome was selected giving preference to i) NCBI reference genome annotated as "complete" at NCBI, ii) NCBI reference genomes, iii) complete NCBI representative genomes, iv) NCBI representative genomes, and v) GTDB representative genomes. **Table S19** indicates which of these categories a genome falls into. The Bayesian trees were inferred with PhyloBayes-MPI version 1.10.2 with the following settings*: -cat -gtr -x 10 -1 -dgam 4*. Three independent chains were run and tested for convergence (maxdiff < 0.3) using bpcomp (part of pb_mpi) whereby the first 1,000 trees were eliminated as burn-in and the remaining trees were sampled every 10 trees.

Supertree

Supertrees were calculated from 1) 122 GTDB markers, and 2) from 253 markers which were present in >10% of archaeal genomes and were extracted from the PhyloPhlAn dataset (Segata et al., 2013). For the 122 marker set maximum-likelihood (ML) inference was performed for each marker gene using the C10 + PSMF model in IQ-Tree. For each marker gene, a guide tree was inferred via approximate ML using FastTree v2.1.9

with the WAG model and gamma distributed rate heterogeneity. A complete ML tree was then inferred using IQ-Tree v 1.6.9 with the C10 + PMSF model with 100 non-parametric bootstrap samplings to assess node support. A consensus tree was constructed from the individual gene trees using ASTRAL v5.6.3. For the 253 marker set, we first identified, 400 conserved, single-copy gene orthologues from the PhyloPhlAn data set (Segata et al., 2013) using PhyloPhlAn v0.99. Markers present in less than 10% of all archaeal genomes were identified and removed resulting in a final marker set of 253 protein markers. For each marker, multiple-sequence alignment was performed using MAFFT-LINSI v7.221 (Katoh and Standley, 2013) and the resulting amino acid alignment was filtered using trimAl 1.2rev59, removing columns present in less than half of all taxa in the alignment. Tree inference was performed using the same approach as for the 122 marker set.

SSU trees

Trees from the 16S rRNA gene alignments (small subunit; SSU) were inferred with IQ-TREE as described above, except the substitution model was determined by IQ-TREEs model finder to be SYM+R10.

Taxonomic assignment and rank standardisation

The assignment of higher taxonomic ranks was normalized based on the relative evolutionary distance (RED) calculated from the ar122.r89 tree using PhyloRank (v0.0.37; https://github.com/dparks1134/PhyloRank/) as described previously (Parks et al., 2018). In brief, PhyloRank linearly interpolates the RED values of internal nodes according to lineage-specific rates of evolution under the constraints of the root being defined as zero and the RED of all present taxa being defined as one. To account for the influence of the root placement on RED values PhyloRank roots a tree multiple times, at the midpoint of each phylum with two or more classes. The RED of a taxon is then calculated as the median RED over all these tree rootings, excluding the tree in which the taxon was the outgroup. The RED intervals for each rank were defined as the median RED value ±0.1 to serve as a guide for the normalization of taxonomic ranks from genus to phylum in GTDB. The rank of species was assigned using ANI and AF criteria (see above).

Assessment of phylogenetic congruence

We measured the phylogenetic tree similarity applying the normalized robinson-Foulds distance (RF) of each alternative tree compared to the ar122.89 tree. The Robinson-Foulds distance is defined as the number of splits that are present in one tree but not in the other one, and vice versa (Robinson and Foulds, 1981). The normalized Robinson-Foulds distance is a relative measure obtained by dividing the calculated Robison-Foulds distance by the maximal Robinson-Foulds distance. The resulting distance is a value between 0% and 100%, which can be interpreted as the percentage of different or missing splits in the alternative trees compared to the ar.122.89 tree (Kupczok et al., 2010).

Assessment of taxonomic congruence

The congruence of the GTDB taxonomy in different trees was assessed as (i) the percentage of taxa identified as monophyletic, operationally monophyletic or polyphyletic, (ii) the RED distributions for taxa at each rank relative to the median RED value of that rank, and (iii) the number of genomes with identical or conflicting

taxonomic assignments between compared trees. Thereby (i) was carried out by placing each taxon on the node with the highest resulting F measure, which is defined as the harmonic mean of precision and recall, and it has been proposed for decorating trees with a donor taxonomy (McDonald et al., 2012). Since only a few incongruent genomes can cause a large number of polyphyletic taxa, we classified taxa with an F measure ≥0.95 as operationally monophyletic.

Data availability

The GTDB taxonomy is available at the GTDB website (https://gtdb.ecogenomic.org/), including the ar122.r89 tree and the GTDB and NCBI taxonomic assignments for all 2,392 archaeal genomes in GTDB 04-RS89. The standalone tool GTDB-Tk, which enables researchers to classify their own genomes according to the GTDB taxonomy is available at GitHub (https://github.com/Ecogenomics/GTDBTk/) and through KBase (https://kbase.us/applist/apps/ kb_gtdbtk/run_kb_gtdbtk/release). Genome assemblies are available from the NCBI Assembly database (BioProject: PRJNA593905).

Acknowledgements

We thank Brian Kemish and Dinindu Senanayake for system administration support, Pelin Yilmaz for stimulating discussions on archaeal taxonomy, and the GTDB user community for their feedback. We also thank the Australian Centre for Ecogenomics (ACE) at The University of Queensland and the New Zealand eScience Infrastructure (NeSI) for providing high performance computing facilities. The project was supported by an Australian Research Council (ARC) Future Fellowship (FT170100213) awarded to C.R. and by an Australian Research Council Laureate Fellowship (FL150100038) awarded to P.H.

References

Adam, P.S., Borrel, G., Brochier-Armanet, C., Gribaldo, S., 2017. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. ISME J. https://doi.org/10.1038/ismej.2017.122

Ali, R.H., Bogusz, M., Whelan, S., 2019. Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments. Mol. Biol. Evol. 36, 2340–2351. https://doi.org/10.1093/molbev/msz142

Baker, B.J., Comolli, L.R., Dick, G.J., Hauser, L.J., Hyatt, D., Dill, B.D., Land, M.L., VerBerkmoes, N.C., Hettich, R.L., Banfield, J.F., 2010. Enigmatic, ultrasmall, uncultivated Archaea. PNAS 107, 8806–8811. https://doi.org/10.1073/pnas.0914470107

Baker, B.J., Saw, J.H., Lind, A.E., Lazar, C.S., Hinrichs, K.-U., Teske, A.P., Ettema, T.J.G., 2016. Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea. Nat Microbiol 1, 1–9. https://doi.org/10.1038/nmicrobiol.2016.2

Barns, S.M., Delwiche, C.F., Palmer, J.D., Pace, N.R., 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc Natl Acad Sci U S A 93, 9188–9193.

Brochier-Armanet, C., Boussau, B., Gribaldo, S., Forterre, P., 2008. Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Micro 6, 245–252. https://doi.org/10.1038/nrmicro1852

Castelle, C.J., Wrighton, K.C., Thomas, B.C., Hug, L.A., Brown, C.T., Wilkins, M.J., Frischkorn, K.R., Tringe, S.G., Singh, A., Markillie, L.M., Taylor, R.C., Williams, K.H., Banfield, J.F., 2015. Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. Current Biology 16, 690–701. https://doi.org/10.1016/j.cub.2015.01.014

Cavicchioli, R., 2011. Archaea — timeline of the third domain. Nat Rev Micro 9, 51–61. https://doi.org/10.1038/nrmicro2482

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz848

Chuvochina, M., Rinke, C., Parks, D.H., Rappé, M.S., Tyson, G.W., Yilmaz, P., Whitman, W.B., Hugenholtz, P., 2019. The importance of designating type material for uncultured taxa. Systematic and Applied Microbiology, Taxonomy of uncultivated Bacteria and Archaea 42, 15–21. https://doi.org/10.1016/j.syapm.2018.07.003

Criscuolo, A., Gribaldo, S., 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evolutionary Biology 10, 210. https://doi.org/10.1186/1471-2148-10-210

Elkins, J.G., Podar, M., Graham, D.E., Makarova, K.S., Wolf, Y., Randau, L., Hedlund, B.P., Brochier-Armanet, C., Kunin, V., Anderson, I., Lapidus, A., Goltsman, E., Barry, K., Koonin, E.V., Hugenholtz, P., Kyrpides, N., Wanner, G., Richardson, P., Keller, M., Stetter, K.O., 2008. A korarchaeal genome reveals insights into the evolution of the Archaea. Proceedings of the National Academy of Sciences 105, 8102–8107. https://doi.org/10.1073/pnas.0801980105

Federhen, S., 2012. The NCBI Taxonomy database. Nucleic Acids Res 40, D136–D143. https://doi.org/10.1093/nar/gkr1178

Fuchs, T., Huber, H., Burggraf, S., Stetter, K.O., 1996. 16S rDNA-based Phylogeny of the Archaeal Order Sulfolobales and Reclassification of Desulfurolobus ambivalens as Acidianus ambivalens comb. nov. Systematic and Applied Microbiology 19, 56–60. https://doi.org/10.1016/S0723-2020(96)80009-9

Grant, W.D., Kamekura, M., McGENITY, T.J., Ventosa, A., 2001. Class III. Halobacteria class. nov, in: Bergey's Manual of Systematic Bacteriology, 2nd Ed., Vol. 1 (The Archaea and the Deeply Branching and Phototrophic Bacteria) (D.R. Boone and R.W. Castenholz, Eds.). Springer-Verlag, New York, p. 294.

Gribaldo, S., Brochier-Armanet, C., 2006. The origin and evolution of Archaea: a state of the art. Philos Trans R Soc Lond B Biol Sci 361, 1007–1022. https://doi.org/10.1098/rstb.2006.1841

Guy, L., Ettema, T.J.G., 2011. The archaeal 'TACK' superphylum and the origin of eukaryotes. Trends in Microbiology 19, 580–587. https://doi.org/10.1016/j.tim.2011.09.002

Guy, L., Spang, A., Saw, J.H., Ettema, T.J.G., 2014. 'Geoarchaeote NAG1'' is a deeply rooting lineage of the archaeal order Thermoproteales rather than a new phylum.' ISME J 8, 1353–1357. https://doi.org/10.1038/ismej.2014.6

Haft, D.H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., Chitsaz, F., Derbyshire, M.K., Gonzales, N.R., Gwadz, M., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Yamashita, R.A., Zheng, C., Thibaud-Nissen, F., Geer, L.Y., Marchler-Bauer, A., Pruitt, K.D., 2018. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res 46, D851–D860. https://doi.org/10.1093/nar/gkx1068

Huber, H., Hohn, M.J., Rachel, R., Fuchs, T., Wimmer, V.C., Stetter, K.O., 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 417, 63–67. https://doi.org/10.1038/417063a

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. Nature Microbiology 1, 16048. https://doi.org/10.1038/nmicrobiol.2016.48

Hugenholtz, P., Skarshewski, A., Parks, D.H., 2016. Genome-Based Microbial Taxonomy Coming of Age. Cold Spring Harb Perspect Biol 8, a018085. https://doi.org/10.1101/cshperspect.a018085

Jungbluth, S.P., Amend, J.P., Rappé, M.S., 2017. Metagenome sequencing and 98 microbial genomes from Juan de Fuca Ridge flank subsurface fluids. Scientific Data 4, sdata201737. https://doi.org/10.1038/sdata.2017.37

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., Petrov, A.I., 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 46, D335–D342. https://doi.org/10.1093/nar/gkx1038

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30, 772–780. https://doi.org/10.1093/molbev/mst010

Kozlov, A.M., Aberer, A.J., Stamatakis, A., 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics 31, 2577–2579. https://doi.org/10.1093/bioinformatics/btv184

Kozubal, M.A., Romine, M., Jennings, R. deM, Jay, Z.J., Tringe, S.G., Rusch, D.B., Beam, J.P., McCue, L.A., Inskeep, W.P., 2013. Geoarchaeota: a new candidate phylum in the Archaea from high-temperature acidic iron mats in Yellowstone National Park. ISME J 7, 622–634. https://doi.org/10.1038/ismej.2012.132

Kupczok, A., Schmidt, H.A., von Haeseler, A., 2010. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. Algorithms Mol Biol 5, 37. https://doi.org/10.1186/1748-7188-5-37

Lartillot, N., Philippe, H., 2004. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Mol Biol Evol 21, 1095–1109. https://doi.org/10.1093/molbev/msh112

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., Hugenholtz, P., 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6, 610–618. https://doi.org/10.1038/ismej.2011.139

Mendler, K., Chen, H., Parks, D.H., Lobb, B., Hug, L.A., Doxey, A.C., 2019. AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz246

Meng, J., Xu, J., Qin, D., He, Y., Xiao, X., Wang, F., 2014. Genetic and functional properties of uncultivated MCG archaea assessed by metagenome and gene expression analyses. ISME J 8, 650–659. https://doi.org/10.1038/ismej.2013.174

Minegishi, H., Kamekura, M., Itoh, T., Echigo, A., Usami, R., Hashimoto, T., 2010. Further refinement of the phylogeny of the Halobacteriaceae based on the full-length RNA polymerase subunit B′ (rpoB′) gene. International Journal of Systematic and Evolutionary Microbiology 60, 2398–2408. https://doi.org/10.1099/ijs.0.017160-0

Murray, R.G.E., Stackebrandt, E., 1995. Taxonomic Note: Implementation of the Provisional Status Candidatus for Incompletely Described Procaryotes. International Journal of Systematic and Evolutionary Microbiology 45, 186–187. https://doi.org/10.1099/00207713-45-1-186

Nawrocki, E., 2009. Structural RNA Homology Search and Alignment Using Covariance Models. All Theses and Dissertations (ETDs). https://doi.org/10.7936/K78050MP

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol 32, 268–274. https://doi.org/10.1093/molbev/msu300

Nunoura, T., Takaki, Y., Kakuta, J., Nishi, S., Sugahara, J., Kazama, H., Chee, G.-J., Hattori, M., Kanai, A., Atomi, H., Takai, K., Takami, H., 2011. Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. Nucleic Acids Res 39, 3204–3223. https://doi.org/10.1093/nar/gkq1228

Oren, A., 2017. A plea for linguistic accuracy – also for Candidatus taxa. International Journal of Systematic and Evolutionary Microbiology 67, 1085–1094. https://doi.org/10.1099/ijsem.0.001715

Oren, A., da Costa, M.S., Garrity, G.M., Rainey, F.A., Rosselló-Móra, R., Schink, B., Sutcliffe, I., Trujillo, M.E., Whitman, W.B., 2015. Proposal to include the rank of phylum in the International Code of Nomenclature of Prokaryotes. International Journal of Systematic and Evolutionary Microbiology 65, 4284–4287. https://doi.org/10.1099/ijsem.0.000664

Parker, C.T., Tindall, B.J., Garrity, G.M., 2019. International Code of Nomenclature of Prokaryotes. International Journal of Systematic and Evolutionary Microbiology, 69, S1–S111. https://doi.org/10.1099/ijsem.0.000778

Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., Hugenholtz, P., 2019. Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy. bioRxiv 771964. https://doi.org/10.1101/771964

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., Hugenholtz, P., 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology 36, 996–1004. https://doi.org/10.1038/nbt.4229

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., Tyson, G.W., 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology 2, 1533–1542. https://doi.org/10.1038/s41564-017-0012-7

Petitjean, C., Deschamps, P., López-García, P., Moreira, D., 2014. Rooting the Domain Archaea by Phylogenomic Analysis Supports the Foundation of the New Kingdom Proteoarchaeota. Genome Biol Evol 7, 191–204. https://doi.org/10.1093/gbe/evu274

Petitjean, C., Deschamps, P., López-García, P., Moreira, D., Brochier-Armanet, C., 2015. Extending the conserved phylogenetic core of archaea disentangles the evolution of the third domain of life. Mol. Biol. Evol. 32, 1242–1254. https://doi.org/10.1093/molbev/msv015

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 5, e9490. https://doi.org/10.1371/journal.pone.0009490

Probst, A.J., Ladd, B., Jarett, J.K., Geller-McGrath, D.E., Sieber, C.M.K., Emerson, J.B., Anantharaman, K., Thomas, B.C., Malmstrom, R.R., Stieglmeier, M., Klingl, A., Woyke, T., Ryan, M.C., Banfield, J.F., 2018. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. Nature Microbiology 3, 328. https://doi.org/10.1038/s41564-017-0098-y

Quehenberger, J., Shen, L., Albers, S.-V., Siebers, B., Spadiut, O., 2017. Sulfolobus – A Potential Key Organism in Future Biotechnology. Front Microbiol 8. https://doi.org/10.3389/fmicb.2017.02474

Raymann, K., Brochier-Armanet, C., Gribaldo, S., 2015. The two-domain tree of life is linked to a new root for the Archaea. PNAS 112, 6670–6675. https://doi.org/10.1073/pnas.1420858112

Reysenbach, Anna-Louise, 2001. Class IV. Thermoplasmata class nov, in: Bergey's Manual of Systematic Bacteriology. Springer-Verlag, New York, p. p 335–338.

Reysenbach, A-L, 2001. Class I. Thermoprotei class. nov, in: In DR Boone and RW Castenholz, Eds.. Bergey's Manual of Systematic Bacteriology Volume 1: The Archaea and the Deeply Branching and Phototrophic Bacteria. Springer Verlag, p. 169.

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 advance online publication.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53, 131–147. https://doi.org/10.1016/0025-5564(81)90043-2

Segata, N., Börnigen, D., Morgan, X.C., Huttenhower, C., 2013. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun 4, 1–11. https://doi.org/10.1038/ncomms3304

Seitz, K.W., Dombrowski, N., Eme, L., Spang, A., Lombard, J., Sieber, J.R., Teske, A.P., Ettema, T.J.G., Baker, B.J., 2019. Asgard archaea capable of anaerobic hydrocarbon cycling. Nature Communications 10, 1822. https://doi.org/10.1038/s41467-019-09364-x

Sorokin, D.Y., Messina, E., La Cono, V., Ferrer, M., Ciordia, S., Mena, M.C., Toshchakov, S.V., Golyshin, P.N., Yakimov, M.M., 2018. Sulfur Respiration in a Group of Facultatively Anaerobic Natronoarchaea Ubiquitous in Hypersaline Soda Lakes. Front. Microbiol. 9. https://doi.org/10.3389/fmicb.2018.02359

Sorokin, D.Y., Yakimov, M., Messina, E., Merkel, A.Y., Bale, N.J., Sinninghe Damsté, J.S., 2019. Natronolimnobius sulfurireducens sp. nov. and Halalkaliarchaeum desulfuricum gen. nov., sp. nov., the first sulfur-respiring alkaliphilic haloarchaea from hypersaline alkaline lakes. International Journal of Systematic and Evolutionary Microbiology, 69, 2662–2673. https://doi.org/10.1099/ijsem.0.003506

Spang, A., Caceres, E.F., Ettema, T.J.G., 2017. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. Science 357, eaaf3883. https://doi.org/10.1126/science.aaf3883

Stieglmeier, M., Klingl, A., Alves, R.J.E., Rittmann, S.K.-M.R., Melcher, M., Leisch, N., Schleper, C., 2014. Nitrososphaera viennensis gen. nov., sp. nov., an aerobic and mesophilic, ammonia-oxidizing archaeon from soil and a member of the archaeal phylum Thaumarchaeota. International Journal of Systematic and Evolutionary Microbiology 64, 2738–2752. https://doi.org/10.1099/ijs.0.063172-0

Vanwonterghem, I., Evans, P.N., Parks, D.H., Jensen, P.D., Woodcroft, B.J., Hugenholtz, P., Tyson, G.W., 2016. Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. Nature Microbiology 1, 16170. https://doi.org/10.1038/nmicrobiol.2016.170

Wang, H.-C., Minh, B.Q., Susko, E., Roger, A.J., 2018. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Syst Biol 67, 216–235. https://doi.org/10.1093/sysbio/syx068

Wheeler, T.J., Eddy, S.R., 2013. nhmmer: DNA homology search with profile HMMs. Bioinformatics 29, 2487–2489. https://doi.org/10.1093/bioinformatics/btt403

Whitman, W.B., 2016. Modest proposals to expand the type material for naming of prokaryotes. Int. J. Syst. Evol. Microbiol. 66, 2108–2112. https://doi.org/10.1099/ijsem.0.000980

Whitman, W.B., Oren, A., Chuvochina, M., da Costa, M.S., Garrity, G.M., Rainey, F.A., Rossello-Mora, R., Schink, B., Sutcliffe, I., Trujillo, M.E., Ventura, S., 2018. Proposal of the suffix –ota to denote phyla. Addendum to 'Proposal to include the rank of phylum in the International Code of Nomenclature of Prokaryotes.' International Journal of Systematic and Evolutionary Microbiology 68, 967–969. https://doi.org/10.1099/ijsem.0.002593

Williams, T.A., Szöllősi, G.J., Spang, A., Foster, P.G., Heaps, S.E., Boussau, B., Ettema, T.J.G., Embley, T.M., 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. PNAS 114, E4602–E4611. https://doi.org/10.1073/pnas.1618463114

Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74, 5088–5090.

Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U.S.A. 87, 4576–4579.

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. https://doi.org/10.1038/nature21031

Zhou, X., Shen, X.-X., Hittinger, C.T., Rokas, A., 2018. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. Mol Biol Evol 35, 486–503. https://doi.org/10.1093/molbev/msx302

Zuo, G., Xu, Z., Hao, B., 2015. Phylogeny and Taxonomy of Archaea: A Comparison of the Whole-Genome-Based CVTree Approach with 16S rRNA Sequence Analysis. Life 5, 949–968. https://doi.org/10.3390/life5010949

Tables and Figures

Figure 1 | Comparison of rank normalized archaeal GTDB and NCBI taxonomies. (a) Relative evolutionary divergence (RED) of taxa defined by the NCBI taxonomy; **(b)** RED of taxa defined by the curated GTDB taxonomy (04-RS89). In (a) and (b) each data point (black circle) represents a taxon distributed according to its RED value (x-axis) and its rank (y-axis). The fill colors of the circle (green, orange, or red) indicate that a taxon is monophyletic, operationally monophyletic, or polyphyletic, respectively in the underlying genome tree. An overlaid histogram shows the relative abundance of monophyletic, operationally monophyletic, and polyphyletic taxa for each 0.025 RED interval. A blue bar shows the median RED value, and two black bars the RED interval (+/- 0.1) for each rank. Note, that in the NCBI taxonomy the values of the higher ranks (order and above) are very unevenly distributed, to the point where the median distributions were not even in the expected order, i.e. the median RED value for classes was higher than the median order value. The GTDB taxonomy uses the RED value to resolve over- and under-classified taxa by moving them to a new interior node (horizontal shift in plot) or by assigning them to a new rank (vertical shift in plot). Note that only monophyletic or operationally monophyletic taxa were used to calculate the median RED values for each rank. In addition, only taxa with a minimum of two children were considered for the GTDB tree (--min_children 2), however, a more lenient approach (--min_children 0) was necessary for the NCBI tree since, except for the Euryarchaeota, none of the other NCBI phyla had the required minimum of two classes. RED values were calculated based on the ar122.r89 tree, inferred from 122 concatenated proteins, decorated with the NCBI and GTDB taxonomy, respectively. **(c)** Rank comparison of GTDB and NCBI taxonomies. Shown are changes in GTDB compared to the NCBI taxonomic assignments across 2,392 archaeal genomes from RefSeq/GenBank release 89. In the bars on the left, a taxon is shown as unchanged if its name was identical in both taxonomies, as a passive change if the GTDB taxonomy provided name information absent in the NCBI taxonomy, or as active change if the name was different between the two taxonomies. The right bar shows the changes of the entire tax string (consisting of seven ranks) per genome, indicating that most genomes had active and passive changes in their taxonomy. **Acronyms:** phylum Euryarchaeota (Eur), phylum Candidatus Marsarchaeota (Mar), phylum Candidatus Thorarchaeota (Tho), class Thermoprotei (Tpr), class Thermoplasmata (Tpl), class Halobacteria (Hal), genus *Thermofilum* (Thm), genus *Methanobrevibacter* (Meth), genus Sulfolobus (Sulf).

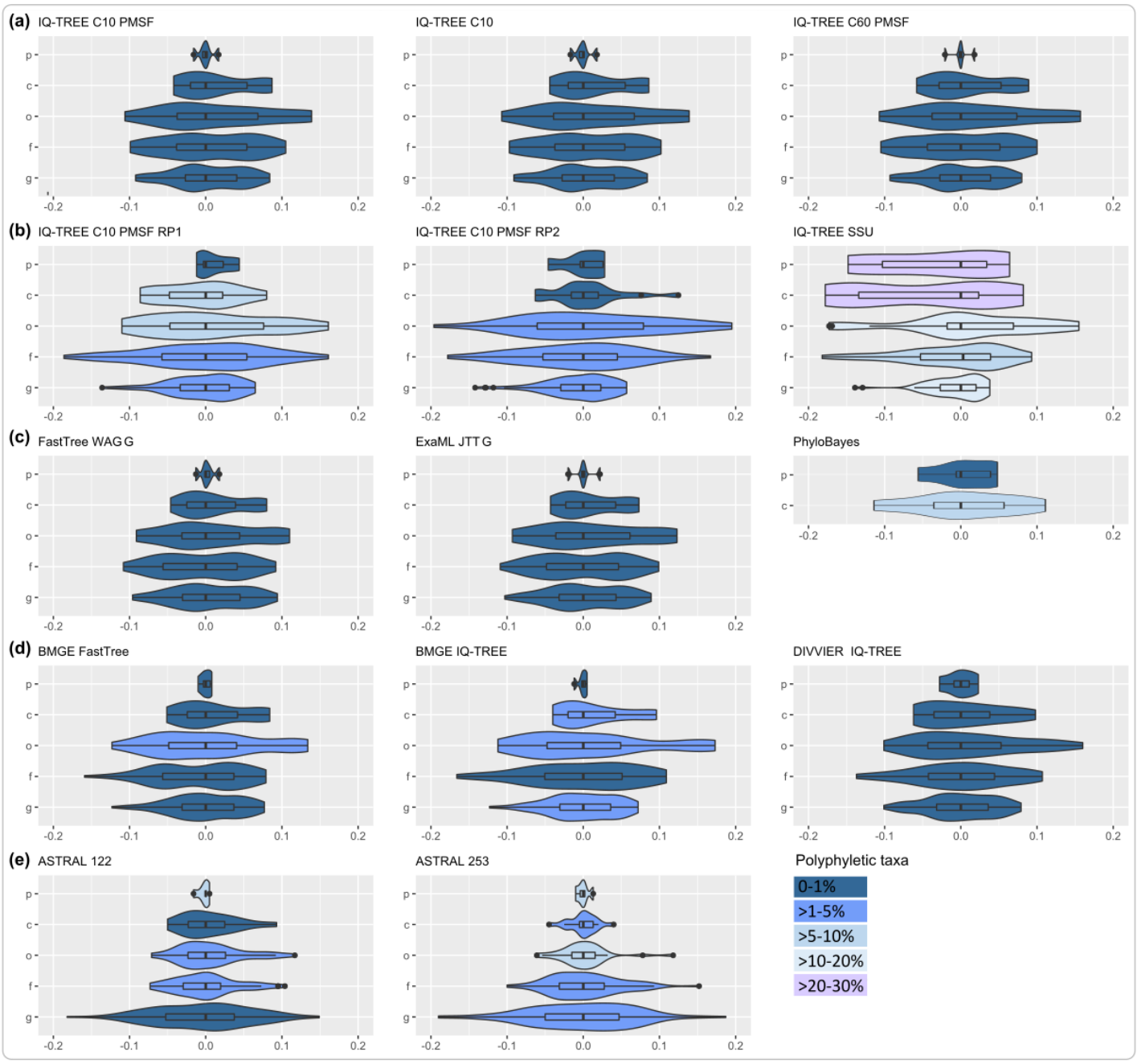

Figure 2 | Comparison of marker sets, inference methods and models. Phylogenetic trees inferred with different methods, from varying concatenated alignments, or via supertree approaches were decorated with the GTDB 04-RS89 taxonomy. RED distributions for taxa at each rank (p, phylum; c, class; o, order; f, family; g, genus) are shown relative to the median RED value of the rank. The legend indicates the percentage of polyphyletic taxa per rank, defined as an F measure <0.95. Note that only taxa with two or more genomes were included. **(a)** trees inferred with IQ-TREE from a concatenated alignment of the 122 GTDB markers with ~5,000 alignment columns. **(b)** Trees inferred from alternative protein markers, including 16 ribosomal proteins (rp1), 23 ribosomal proteins (rp2), and the SSU rRNA gene. **(c)** Trees inferred from 122 marker alignments using different inference software and models. Note PhyloBayes was calculated from the order dereplicated data set, allowing the evaluation of the ranks phylum and class only. **(d)** Trees inferred from a modified concatenated alignment of the 122 GTDB markers to account for compositional bias. **(e)** Trees inferred with the ASTRAL supertree approach. More details about inference models and methods are given in **Table S10**.
Abbreviations: IQ-TREE C10 122 (2.3.iqtree.c10.slow), IQ-TREE C60 PMSF 122 (2.4.iqtree.c60.pmsf.fasttreeStart), IQ-TREE C10 PMSF RP1 (05.rp1.iqtree.c10.pmsf.fasttreeStart.rp1), IQ-TREE C10 PMSF RP2 (07.rp2.iqtree.c10.pmsf.fasttreeStart), IQ-SSU (14.3.SSU.IQtree.900bp), FastTree WAGG 122 (01fasttree1248xwagg.tree), ExaML JTTG 122 (12.ExaML_JTT_G), PhyloBayes CAT 122 (10.phylobayes), BMGE FastTree 122 (11.1.bmge.FastTree), BMGE IQ-TREE 122 (11.2.bmge.IQTree), DIVVIER IQ-TREE 122 (11.5.divvier.mincol_4.IQtree), ASTRAL 122 (18.1.astral.ar122.lpp), ASTRAL 253 (18.2.astral.253x.phylophlan.lpp) .

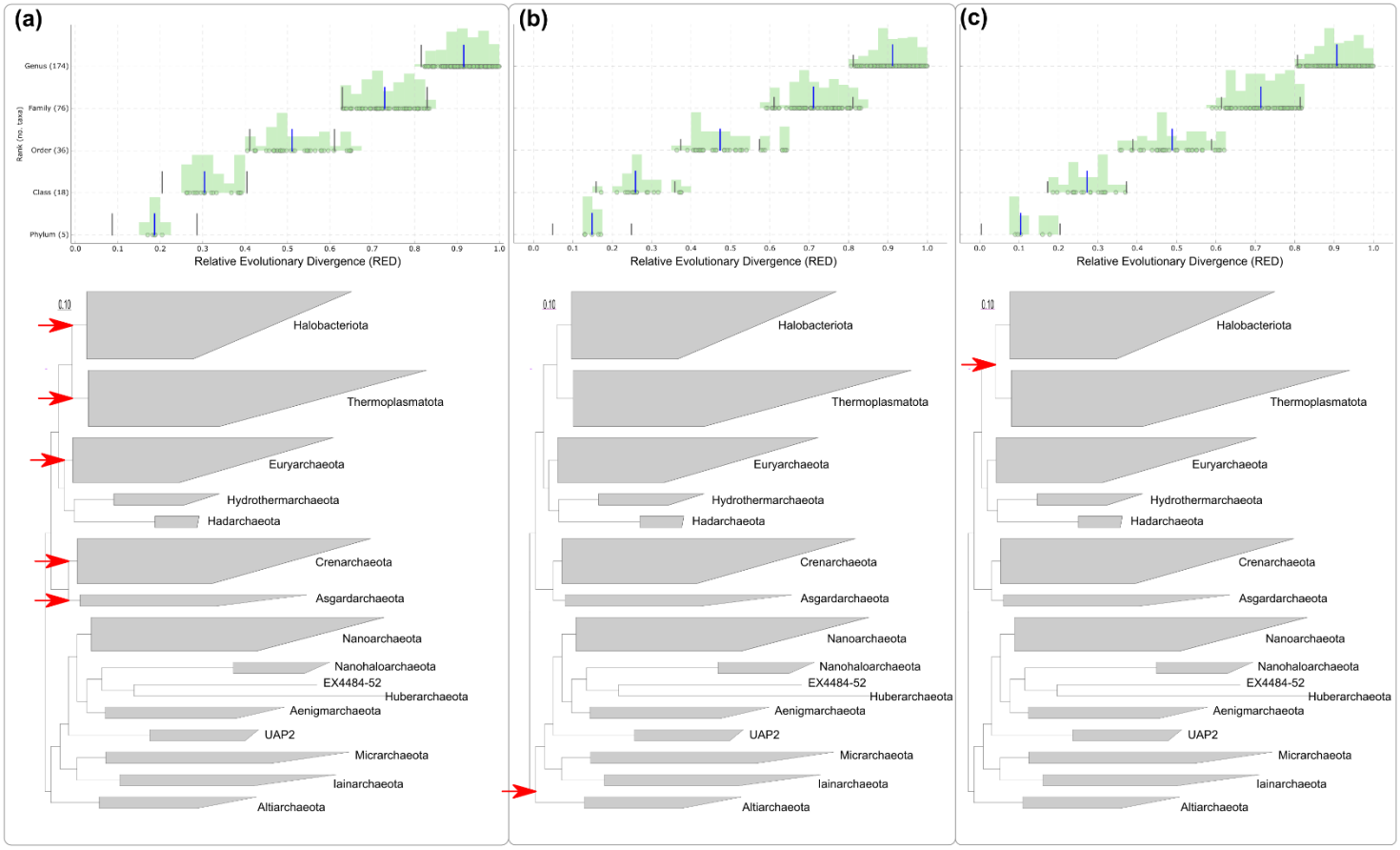

Figure 3 | Impact of different rooting scenarios on the relative evolutionary divergence (RED). The rooting approach implemented in GTDB **(a),** which calculates the relative evolutionary divergence (RED) as the median of all possible rootings of phyla with at least two children (red arrows), is compared to a fixed root between the DPANN superphylum (red arrow) and the remaining Archaea **(b)**, and to a fixed root within the NCBI phylum Euryarchatoa, which translates to a root between the two phyla Thermoplasmata and Halobacteriota (red arrow) and the rest of the Archaea in the GTDB taxonomy **(c)**. In the upper bar charts each data point (black circle) represents a taxon distributed according to its RED value (x-axis) and its rank (y-axis). An overlaid histogram shows the relative abundance of taxa for each 0.025 RED interval, a blue bar shows the median RED value, and two black bars the RED interval (+/- 0.1) for each rank. Note that, overall, the ranks can be distinguished based on their RED value, regardless of the applied rooting scenario. The scale bars indicate 0.1 substitutions.

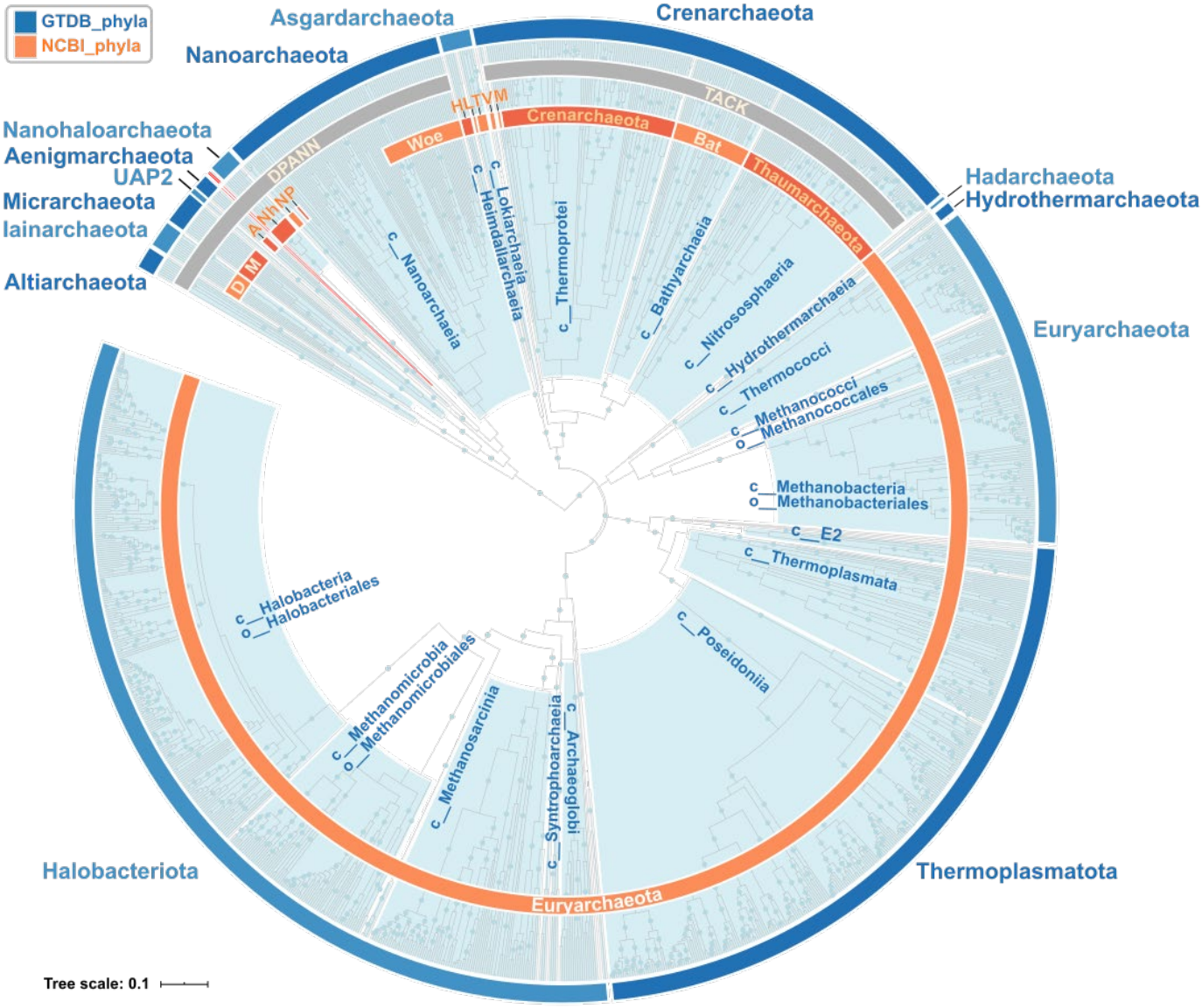

Figure 4 | Rank normalized archaeal GTDB taxonomy. Species dereplicated scaled ar122.r89 tree decorated with the archaeal GTDB taxonomy R04-RS89. The outer blue ring denotes the rank normalized phyla, and the light blue clades indicate the classes in the rank normalized GTDB taxonomy. Classes with 10 or more taxa are labelled, and a below class divergence is indicated by providing the lower rank (e.g. order) for a given class. The two GTDB phyla consisting of only a single species each, namely Huberarchaeota and EX4484-52 are highlighted by red branches indicating their uncertain placement in the ar122.r89 tree. The inner orange ring denotes the r89 NCBI phyla with 2 or more taxa. The NCBI superphyla TACK and DPANN are indicated with gray color strips. Abbreviations are the following: Bat (*Ca.* Bathyarchaeota), M (*Ca.* Marsarchaeota), V (*Ca.* Verstraetearchaeota), T (*Ca*. Thorarchaeota), L (*Ca*. Lokiarchaeota), H (*Ca.* Heimdallarchaeota), Woe (*Ca.* Woesearchaeota), P (Ca. Parvarchaeota), N (Nanoarchaeota), Nh (*Ca.* Nanohaloarchaeota), A (*Ca*. Aenigmarchaeota), M (*Ca.* Micrarchaeota), D (*Ca*. Diapherotrites). Note: the scaled tree was generated by replacing the branch lengths with the median relative evolutionary distance (RED), calculated across all plausible rootings. Bootstrap values over 90% are indicated by blue dots. Scale bar indicates 0.1 RED.

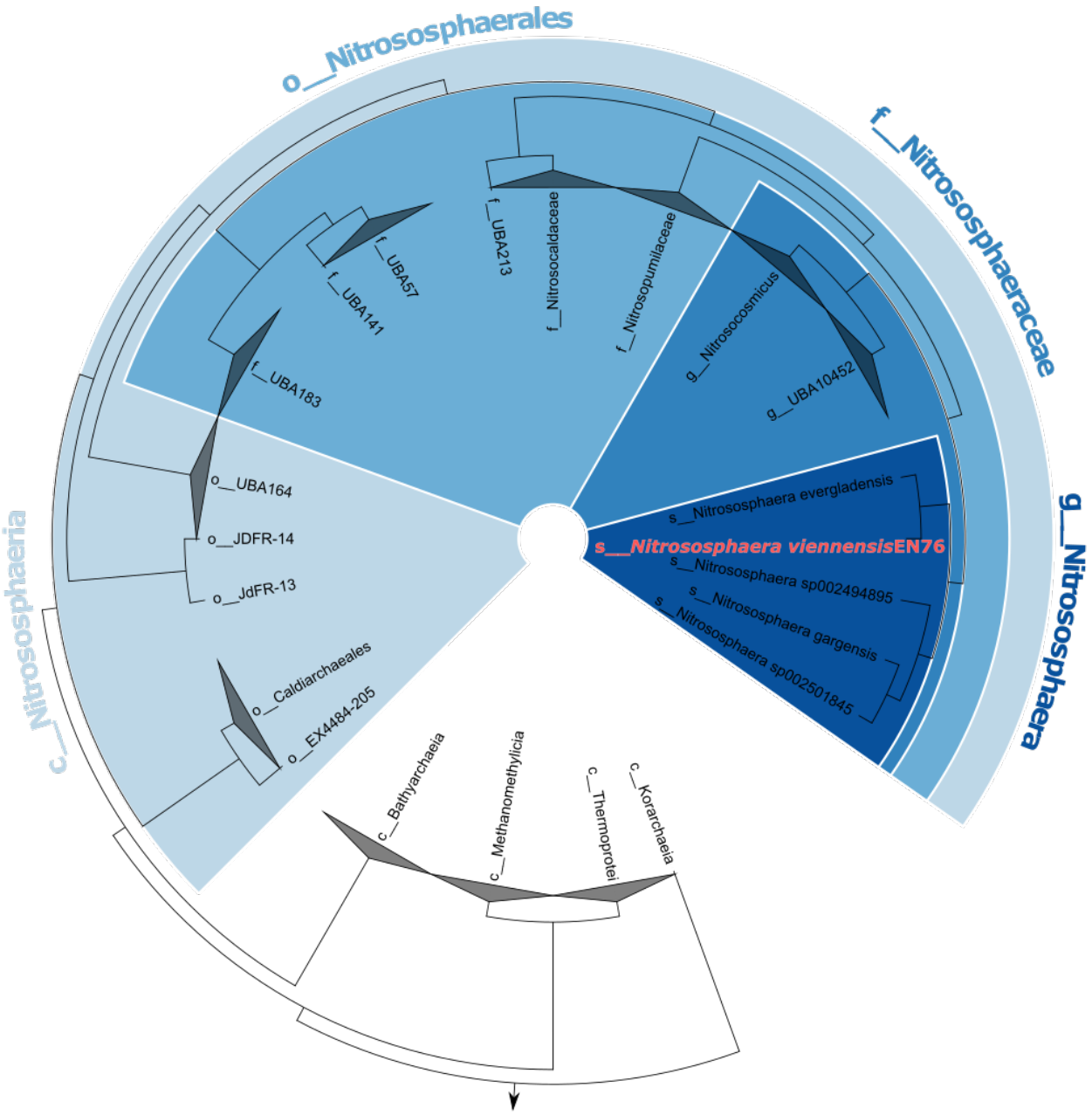

Figure 5 | Reclassification of the members of the Thaumarchaeota. Inverse cladogram based on the ar122.r89 tree showing the GTDB phylum Crenarchaeota with the classes Korarchaeia, Thermoprotei, Methanomethylicia (former phylum *Candidatus* Verstraetearchaeota), Bathyarchaeia, and Nitrososphaeria. The validly published class Nitrososphaeria (light blue) was emended in GTDB to include all taxa assigned to the phylum Thaumarchaeota in the NCBI r89 taxonomy. The type species of this lineage is *Nitrososphaera viennensis* (Stieglmeier et al. 2014), which serves as the type for higher taxa including the genus *Nitrososphaera*, the family *Nitrososphaeraceae*, the order *Nitrososphaerales*, and the class *Nitrososphaeria*. The genome of the *Nitrososphaera viennensis* type strain EN76T (=DSM 26422T=JMC 19564T), is highlighted in red. Arrow points to outgroup.

Supplementary information

Supplementary Tables and Figures

See files "*Supplementary Tables and Figures*" 1, 2, and 3.

Supplementary data

All GTDB files for Release 04-RS89 are available through the GTDB repository https://data.ace.uq.edu.au/public/gtdb/data/releases/release89/89.0/

Supplementary File 01 | The archaeal r89 taxonomy (release R04-RS89). Taxonomy file for all 2392 archaeal genomes. File name: *ar122_taxonomy_r89.tsv*

Supplementary File 02 | GTDB r89 archaeal synonyms. Synonyms and synonym representatives for archaeal GTDB species (release R04-RS89). File name: *Suppl_file_02_GTDB_r89_archaeal_synonyms.xlsx*