Anomalous phylogenetic behavior of ribosomal proteins in metagenome assembled asgard archaea

Sriram G. Garg[1*†], Nils Kapust[1†], Weili Lin[2†], Michael Knopp[1], Fernando D. K. Tria[1], Shijulal Nelson-Sathi[3], Sven B. Gould[1], Lu Fan[4,5,6], Ruixin Zhu[2], Chuanlun Zhang[4,5,7], William F. Martin[1*]

[1] Institute for Molecular Evolution, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

[2] Putuo people's Hospital, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

[3] Transdisciplinary Biology, Computational Biology Laboratory Rajiv Gandhi Centre for Biotechnology, Thiruvananthapuram, India

[4] Shenzhen Key Laboratory of Marine Archaea Geo-Omics, Southern University of Science and Technology, Shenzhen 518055, China

[5] Department of Ocean Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China

[6] SUSTech Academy for Advanced Interdisciplinary Studies, Southern University of Science and Technology, Shenzhen 518055, China

[7] Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou 510000, China

† These authors contributed equally

Authors for correspondence: gargs@hhu.de, bill@hhu.de

Abstract

Metagenomic studies permit the exploration of microbial diversity in a defined habitat and binning procedures enable phylogenomic analyses, taxon description and even phenotypic characterizations in the absence of morphological evidence. Such lineages include asgard archaea, which were initially reported to represent archaea with eukaryotic cell complexity, although the first images of such an archaeon show simple cells with prokaryotic characteristics. However, these metagenome-assembled genomes (MAGs) might suffer from data quality problems not encountered in sequences from cultured organisms due to two common analytical procedures of bioinformatics: assembly of metagenomic sequences and binning of assembled sequences on the basis of innate sequence properties and abundance across samples. Consequently, genomic sequences of distantly related taxa, or domains, can in principle be assigned to the same MAG and result in chimeric sequences. The impacts of low-quality or chimeric MAGs on phylogenomic and metabolic prediction remain unknown. Debates that asgard archaeal data are contaminated with eukaryotic sequences are overshadowed by the lack of evidence indicating that individual asgard MAGs stem from the same chromosome. Here we show that universal proteins including ribosomal proteins of asgard archaeal MAGs fail to meet the basic phylogenetic criterion fulfilled by genome sequences of cultured archaea investigated to date: these proteins do not share common evolutionary histories to the same extent as pure culture genomes (PCGs) do, pointing to a chimeric nature of asgard archaeal MAGs. Our analysis suggests that some asgard archaeal MAGs represent unnatural constructs, genome-like patchworks of genes resulting from assembly and/or the binning process. (249 words)

Significance statement

Metagenomics while permitting microbiologists to study complex environments and unculturable microbes, is fundamentally a series of computer algorithms extracting information from sequence data. Using a basic principle underpinning our current understanding of phylogenetics we show that some MAGs can be chimeric. This issue although mitigated by careful curation and generation of closed genomes warrants immediate attention in order to mitigate the effects of chimeric MAGs like those of asgard archaea, in bioinformatic databases leading to snowballing errors in further analyses.

Keywords: Metagenomics, binning, eukaryogenesis, phylogenetics, asgard archaea, candidate phylum radiation (CPR)

Introduction

The sequencing of environmental microbial DNA (metagenomics) allows microbiologists to uncover the existence of genes from a species mix in environments such as marine sediment or the deep biosphere, from which representatives cannot readily be cultured (Torsvik et al. 1996; Handelsman et al. 1998) revealing what is commonly referred to as microbial dark matter (Marcy et al., 2007; Rinke et al., 2013; Bernard et al., 2018). Metagenomic investigations have led to the reconstruction of archaeal and bacterial MAGs from environmental shotgun sequencing data by using metagenomic assembly that is followed by the binning of assembled sequences based on GC content, nucleotide frequency and stoichiometry co-occurrence across samples (Breitwieser et al. 2017). Binning is a crucial step in the MAGs reconstruction process because it assigns metagenomic sequences to a genome, a procedure that is not required in the case of a cultured organisms.

Because rRNA has limited phylogenetic resolution, concatenated sequences of ribosomal proteins (r-proteins) and other universally distributed proteins are now commonly used for phylogeny even though concatenation has its own caveats (Thiergart et al. 2014). This practice is well established with over 20 years of tradition, whereby the validity of using concatenated r-proteins for phylogeny lies in the reproducible crosscheck that individual r-proteins from the same sequenced genome give the same or very similar trees (Hansmann and Martin 2000; Brown et al. 2001; Nesbø et al. 2001; Daubin et al. 2002; Matte-Tailliez et al. 2002; Martin 2003). Based on such precedence, it became common practice to use concatenated r-proteins from sequenced genomes for microbial phylogeny without first crosschecking whether the individual proteins gave similar trees. That practice has been extended to MAGs, which is potentially problematic because there is no independent evidence that the 20 to 30 r-proteins used for phylogeny in a given MAG are encoded in one and the same genome. Potentially, a MAG could stem from DNA fragments of multiple genomes that occur in the same environment in which up to 90% of DNA could be extracellular (Dell'Anno et al. 2002; Dell'Anno and Danovaro 2005; Torti et al. 2015; Nagler et al. 2018). Consequently, the basic assumption of common binning algorithms – sequences from the same cell share the same abundance profile across samples – might not always hold true.

Given the unprecedented speculations regarding the cellular complexity of asgard archaea (Spang et al. 2015; Eme et al. 2017; Zaremba-Niedzwiedzka et al. 2017; Akıl and Robinson 2018; Cunha et al. 2018); especially in light of the non-complex phenotype of the first enriched

and imaged cells from the these metagenomic assemblies (Imachi et al., 2020), methods to validate phylogenies based on universal and r-proteins from MAGs are needed. We turned to a fundamental principle known from the earliest days of phylogenetic testing: proteins with a shared evolutionary history should generate the same or similar trees (Robinson and Foulds 1981; Penny et al. 1982). For genomic DNA from cultivated and isolated prokaryotic organisms, hereafter referred to as pure culture genomes (PCGs) it is known that different r-proteins produce the same tree or similar trees (Hansmann and Martin 2000; Brown et al. 2001; Nesbø et al. 2001; Daubin et al. 2002; Matte-Tailliez et al. 2002; Martin 2003). Yet even for r-proteins that share a common evolutionary history, their trees will differ to some extent owing to practical and theoretical limitations of phylogenetic inference (Hansmann and Martin 2000; Lockh 2000; Brown et al. 2001; Nesbø et al. 2001; Daubin et al. 2002; Matte-Tailliez et al. 2002; Martin 2003; Shen et al. 2017). The extent of natural variation across r-protein trees for PCGs can however be determined empirically by simply comparing the r-protein trees for a given genome set. Using that natural distribution as a reference, one can ask whether MAGs fulfill the same criterion, that is, do trees for r-proteins from MAGs resemble each other to the same or to a lesser degree than trees based on PCGs, and if they differ, is the difference significant?

Results

We performed individual alignments of proteins that are universal, or nearly so, among archaeal genomes from PCGs (Archaeal PCGs) and a set of selected archaeal MAGs containing asgard archaea (hereafter asgard MAGs) (see Methods; Supplementary Figure S1, S2; Supplementary Table S1,S2) in order to compare their phylogenetic properties with regard to consistent phylogenies among proteins. We first generated alignments for 39 different proteins that are present in all members of a large and diverse sample of archaeal PCGs and asgard MAGs. For the Archaeal PCGs, we generated phylogenetic trees for each alignment individually and asked how similar the trees are to one another using the robust Robison-Foulds pairwise distance measure (Robinson and Foulds 1981) (Figure 1A). We asked the same question for the same proteins using asgard MAGs. The result (Figure 1B) shows that in phylogenetic terms, asgard MAGs behave in a manner fundamentally different from Archaeal PCGs in two aspects. First, archaeal PCGs trees are more similar to one another than asgard MAG trees are. Second, the 23 ribosomal proteins (including secY, which is ribosomal for co-translational insertion) and 16 other proteins universally distributed within the genome sample show little difference in their ability to recover approximately the same tree for archaeal PCGs, but obvious differences

for asgard MAGs (light shading in Figure 1B, see scale bar). The evident bimodal distribution of phylogenetic behavior for the 39 asgard MAG proteins is not observed for PCGs.

To quantify these differences, we plotted the distribution of pairwise similarity across 39 trees for ten different samples of archaeal PCGs containing 30 genomes each from RefSeq archaea and compared it with the asgard MAGs sample. The distribution for asgard MAGs, which include Lokiarchaeum and other asgard archaea, appears bimodal using four different tree comparison methods (Figure 2A). Note that the distributions for trees from archaeal MAGs are always shifted towards higher pairwise distances between trees. We plotted the curves for the 23 ribosomal proteins against the 16 other universal proteins (including RNaseH, DNA primase, polA and polB) (Figure 2B, Figure 2C). The phylogenetic behavior of asgard MAGs appears different from all ten Archaeal PCG samples. The distribution of tree dissimilarity scores across r-proteins from asgard MAGs is significantly larger than the corresponding values for archaeal PCGs, with p-values ranging from 10^{-26} to 10^{-139} (two-tailed Kolmogorov-Smirnov test) depending on the tree comparison metric (Figure 2B) (see Methods and Supplementary Table S3B). The probability that the distribution of tree dissimilarities across non-ribosomal proteins from asgard MAGs is drawn from the same distribution as the corresponding value for archaeal PCGs ranges from 10^{-48} to 10^{-66}, depending on the tree comparison metric (see Methods and Supplementary Table S3C).

This difference in phylogenetic behavior does not appear to be due to greater phylogenetic depth or greater sequence divergence of ribosomal proteins in asgard MAGs relative to PCGs, as shown by the distributions of pairwise uncorrected p-distances for asgard MAGs and PCGs for each protein in the sample (Supplementary Figure S3, S4). Despite our careful genome selection procedure to assemble bacteria and archaea datasets to reflect the known phylum-level diversity (Supplementary Figure S1, S2) (not order-level or any other taxon-level because the debates are about novel phyla), a tenacious critic could still suggest that the higher incongruence of gene trees from MAGs may be readily explained by differences in phylogenetic distribution in the data. If MAGs are deeper-branching in comparison to RefSeq genomes, than MAG-derived trees will inevitably show reduced congruence as a result of artefacts from phylogenetic reconstructions (not sequence binning errors) and possibly the more pronounced effects of lateral gene transfer. This possibility, however, is resoundingly rejected by the results of a simple yet informative experiment: for each matched sample we quantified the phylogenetic depth of MAGs and RefSeq trees independently, measured as the mean root-to-tip distances across the rooted topologies (see methods for details). For the asgard

metagenomes we found that while the root to tip distances for the 16 non-ribosomal proteins are significantly different from the archaeal PCGs the 23 ribosomal proteins including SecY agree with each other (Figure 3A; Supplementary Table S4). This recapitulates what we have observed with the bimodal distribution for tree dissimilarity scores (Figure 1, 2A, 2B, 2C) when looking at the 39 universal single copy genes of archaea.

Many phylogenetic analyses involving MAG data employ site-filtering procedures to remove sites from the sequence alignment (Hansmann and Martin 2000; Talavera and Castresana 2007; Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017; Fan et al. 2020). To check whether site filtering affects the phylogenetic anomaly of asgard MAGs, we trimmed alignments (see Methods) as in earlier analyses and recalculated the trees and comparisons. As shown in Supplementary Figure S7, site filtering does not improve MAG phylogenetic behavior (p-values in Supplementary Table S5). The asgard MAGs have a systematic and previously undetected phylogenetic anomaly.

Is the MAG phylogenetic anomaly asgard specific? We examined ten additional non-asgard archaeal MAGs from various metagenome sequencing projects by repeating the tree comparison metrics (see Methods). Ribosomal protein samples from non-asgard archaeal MAGs from marine surface metagenomes that were generated by binning methods are not fundamentally anomalous in their phylogenetic behavior, although trees for non-r-proteins from this extended dataset are significantly different (Supplementary Figure S8A-E; Supplementary Table S6). Furthermore, within the asgard Metagenomes sample we have used 16 published asgard MAGs together with 14 non-asgard archaeal MAGs. In order to determine the effect of mixing asgard and non-asgard MAGs, we replaced the r-proteins of the 14 non-asgard MAGs with non-asgard archaeal MAGs from the Marine group II (Supplementary Table 2) and with r-proteins from archaeal PCGs. While the degree of phylogenetic incongruence is lower due to the addition of marine surface MAGs or RefSeq PCGs, the difference is still significantly higher (p-value $< 10^{-27}$; Supplementary Figure S9; Supplementary Table S7) for two of the tree comparison metrics. Thus, it is the asgard MAGs in the sample that are phylogenetically anomalous with minimal contribution to the r-protein tree incongruence by the other non-asgard archaeal MAGs or archaeal PCGs.

For bacteria, we investigated whether MAGs from the Candidate Phyla Radiation (CPR) group of bacteria showed a similar effect. The probability that the distribution of tree similarities

across 16 ribosomal proteins from CPR MAGs is drawn from the same distribution as the corresponding value for bacterial PCGs is also significantly different, in all tree comparison metrics (Figure 2D). Similar to CPR MAGs, for bacterial MAGs that do not stem from the CPR, the distribution of tree similarities across r-proteins from MAGs is significantly different from bacterial PCGs for the four tree comparison metrics when compared to the reference (Figure 2E). The investigation of 10 additional samples of non-CPR and CPR MAGs generated similar results (Supplementary Figure S8D, S8E). We also looked at the contribution of phylogenetic depth using the root to tip distances for the bacterial comparisons. For the CPR MAG sample and the bacterial MAG sample, while there are no differences between the metagenomes and PCGs, a few of the Reference Samples show significant differences in the distribution of root-to-tip distances within themselves (Supplementary Table S4). Since we are primarily comparing metagenome samples to PCGs, this difference can be overlooked for the purposes of our conclusions (Figure 3B, 3C Supplementary Table S4). Together these results are in line with our findings of aberrant phylogenetic behavior of MAG derived trees stemming from sequence-binning errors.

The quality of MAGs is typically assessed using bioinformatic tools such as CheckM (Parks et al. 2015), a pipeline that uses lineage specific marker proteins to determine completeness and contamination of a given MAG. While both parameters are informative, CheckM does not assess the phylogeny of the marker proteins. As a consequence, chimeric MAGs containing ribosomal proteins of diverse ancestry can receive a high completeness or low contamination score. We compared archaeal MAGs spanning a range of CheckM completeness values (Figure 4). Phylogenetic incongruence of trees in MAGs with 90% completeness or more is higher than those with a completeness between 70–80% (Figure 4, Supplementary Table S8). This reveals that while CheckM looks for the completeness of a genome by looking at single copy marker genes it does not look for phylogenetic congrunce of those marker genes. Furthermore, several closed circular genomes have been recovered from metagenomic surveys (Anantharaman et al. 2016). MAGs reported as closed and high-quality draft genomes from the CPR reveal distributions of tree similarity scores of r-proteins that are similar to bacterial PCGs (Supplementary Figure S10, Supplementary Table S9). These analyses indicate that while r-proteins from asgard MAGs fail to provide congruent phylogenies, it is possible for metagenomic binning to generate draft and closed genomes whose r-proteins show evidence for coevolution.

In the worst case, the phylogenies generated by r-proteins of MAGs might share no better than random similarity, reflecting trees of proteins encoded by DNA of similar sequence properties but perhaps of unlinked phylogenetic histories. To visualize an effect of randomization we employed a standard network method, Neighbor-Net (NNet) (Bryant and Moulton 2004). The NNet for r-proteins sampling crenarchaeal and euryarchaeal PCGs is extremely tree-like, with 16 strong, low conflict splits indicated in the figure (Figure 5A), reflecting highly congruent phylogenetic signals across r-proteins from diverse cultivated archaea (Figure 1A, Figure 2B). An NNet for asgard MAGs is shown in Figure 5B, revealing one split in the data, that separating Thorachaea MAGs from the rest. The phylogenetic congruence across r-proteins in the remaining 21 archaeal MAGs is close to random, as shown by randomizing the source of sequences in individual ribosomal protein alignments, used for concatenation and NNet plots (Figure 5C). Reversing the alignments using the Head or Tails method (Landan and Graur, 2007) shows the same general NNet result, namely a tree-like structure for PCG data but a star-like structure for MAGs (Supplementary Figure 13), indicating that variation in the underlying signal is far greater than variation introduced by alignment procedures. The strong tree-like NNet is also recovered when using the non-asgard archaeal MAG sample (Supplementary Figure S11) which demonstrates the ability of the NNets to recapitulate the tree congruence metrics.

As a further independent test for tree similarity, we examined the phylogenetic compatibility within MAG and PCGs tree sets (see Methods) by scoring the compatibility of each r-protein tree against all other trees from the same set (Nelson-Sathi et al. 2012). Again, r-proteins from genomes of cultured organisms produced trees that are very similar to each other, as they should be (Hansmann and Martin 2000; Brown et al. 2001; Nesbø et al. 2001; Daubin et al. 2002; Matte-Tailliez et al. 2002; Martin 2003), while trees for r-proteins from asgard MAGs differ from each other (Figure 6, Supplementary Figure S12). The differences in cumulative distribution frequencies are obvious (Figure 6) and highly significant ($p = 10^{-18}$, two-tailed Kolmogorov-Smirnov test; Supplementary Tables S10, S11).

Our analyses reveal that the archaeal MAGs designated as Lokiarchaea (Spang et al. 2015) and asgard archaea (Zaremba-Niedzwiedzka et al. 2017) and their r-proteins may perhaps be error prone by bioinformatic artefacts. While the names and rank of asgard archaea is defined by their r-protein phylogenies (Spang et al. 2015; Eme et al. 2017; Zaremba-Niedzwiedzka et al. 2017; Spang et al. 2018), these r-proteins, however, may not be linked by common ancestry,

rather they are stitched together from sequences in same environment, to a degree that remains unknown. Alternatively, it is possible that ribosomal proteins in asgard archaea do not share common histories and thus evolve in a fundamentally different manner from those archaeal PCGs studied here. If the latter is true, concatenated ribosomal proteins of asgard archaea can hardly be used for phylogenomic analyses in the first place.

For the non-ribosomal proteins, the issue might be even more serious (Figure 1; Figure 2). This has gone hitherto unnoticed, because MAGs have not been rigorously tested for congruent phylogenetic properties as are sequences from closed genomes (Hansmann and Martin 2000; Brown et al. 2001; Nesbø et al. 2001; Daubin et al. 2002; Matte-Tailliez et al. 2002; Martin 2003). Notably, a closed genome was recently reported for an anaerobic, fermenting archaeon from sediment (sister to Crenarchaeota, termed *Candidatus* Prometheoarchaeum syntrophicum strain MK-D1) that branches as the sister to eukaryotes in the absence of MAG data (Imachi et al. 2020). It is small (0.5 μm diameter), not phagocytosing and lacks any signs of eukaryote-like complexity. This is in contrast to earlier inferences regarding asgard MAGs, which were speculated to represent complex, eukaryote-like phagocytosing archaea based on binned genomes alone (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017; Eme et al. 2017). There have been challenges posed to individual protein sequences present in asgard MAGs (Cunha et al. 2018; Cunha et al. 2017; Martin et al. 2017). Previous experiments on Lokiarchaea have shown that individual trees for the 36 universal proteins (Spang et al., 2015) recover different positions for lokiarchaea (Cunha et al., 2017). Our findings are congruent with that finding and further suggests that the chimeric nature of lokiarchaea and asgard MAG data are likely systematic. They also predict that proteins from a majority of currently available asgard MAGs will not overlap with those present in *Ca.* P. syntrophicum. The recently emerged images of MK-D1 demonstrate that early skepticism brought forth regarding interpretations of asgard complexity was warranted (Dey et al. 2016; Gould et al. 2016). In the absence of cell-biological data the same caution is generally warranted regarding phylogenetic inference and metabolic predictions of any new group identified solely through metagenomic means. (2816 words)

Methods:

Sources of archaeal metagenomics data

The description and list of organisms comprising each MAG sample is given in Supplementary Table S2. Archaeal Metagenomes comprising of asgard archaea were downloaded from IMG (https://img.jgi.doe.gov/) and NCBI databases (https://www.ncbi.nlm.nih.gov). 30 archaeal

MAGs that contain 39 universal archaeal proteins (23 ribosomal proteins + 16 non-ribosomal proteins) were retained and constitute asgard MAG dataset. In the extended dataset, MAGs were obtained by the following three ways: (1) archaeal MAGs were downloaded from IMG database. (2) Assembled contigs/scaffolds of sediment samples were downloaded from IMG database, and were binned with MetaBAT (version: 2.12.1) (Kang et al., 2019) (3) The raw reads from selected *Tara* Oceans, biofilm and rhizosphere samples were downloaded from Sequence Read Archive (SRA) database. For each sample, clean reads were obtained from raw data by trimming adaptors and low-qualified bases with Trimmomatic (version: 0.38) (Bolger et al 2014). Abundance profiles of assembled contigs were generated by mapping of reads in each sample with BamM (version: 1.7.3; https://ecogenomics.github.io/BamM/) and "jgi_summarize_bam_contig_depths" script in MetaBAT. Using MetaBAT, binning was performed with abundance profiles on assembled contigs. In the extended dataset, all qualified MAGs were screened with a criterion of completeness >= 70% and contamination <= 5% using CheckM (version: 1.0.13) (Parks et al. 2015). 114 archaeal MAGs that contain all 39 universal proteins were retained and used to generate 10 archaeal MAG samples containing 30 genomes each. Samples with different CheckM completeness ranges were subsampled from the afore mentioned extended dataset (Supplementary Table S1), corresponding to completeness values of 70–80%, 80–90 and 90–100% (Supplementary Table S8).

Sources of CPR metagenomics data

The MAGs from Candidate Phyla Radiation were obtained from the authors of Hug et al., 2016. 140 CPR MAGs containing all 16 universal ribosomal proteins were used to create 10 CPR MAG samples of 30 genomes each (Supplementary Table S2). In addition, 23 closed genomes were kindly provided by Prof. Karthik Anantharaman which along with 7 high-quality draft genomes from Anantharaman et al., 2016 were used to create a MAG sample comprising of high-quality MAGs.

Sources of bacterial metagenomics data

The bacterial MAG Dataset was downloaded from the BioProjects with the Accessions PRJNA288027 (2545 Assemblies) and PRJNA270657 (103 Assemblies). This set was further subsampled for 10 groups of 30 organisms that have all 20 bacterial, universal ribosomal protein families, to generate samples each with 30 species or operational taxonomic units (OTUs) (Figure S2; Table S2).

Identification of homologs of ribosomal proteins in RefSeq and metagenomics genomes

Universal protein clusters for Archaea were obtained from Nelson-Sathi et al., 2015 while ribosomal protein clusters Bacteria were retrieved from UniProt (Jan 2019, The UniProt Consortium 2019). These clusters were used for a BLAST against the RefSeq 2016 database (O'Leary et al. 2016) consisting of 5443 bacteria and 212 archaea with an identity threshold of 20% and an e-value cut-off of 10^{-5}. The BLAST searches were also performed on metagenomics genomes. The archaeal RefSeq dataset was further subsampled for groups of 30 organisms that reflect the full breadth of taxonomic distribution for the complete dataset for bacteria and archaea, respectively, to generate 10 reference samples each with 30 species or operational taxonomic units (OTUs) for archaea. For bacteria, the sampling was restricted to only include a maximum of two organism from each phylum to generate 10 reference samples each with 30 species (OTUs) for bacteria. That is, the reference PCGs samples were chosen to sample as much phylogenetic diversity, depth and breadth as the MAG samples (for more information see Supplementary Table S1; Supplementary Figure S1; Supplementary Figure S2).

Phylogenetic tree construction and uncorrected p-distances

For each group (Archaea, Bacteria and CPR) matching sets of ribosomal proteins for each sample were chosen based on their universal presence in all 30 OTUs in the 10 reference sets as well as in the metagenomes. Maximum-likelihood trees were calculated using IQ-tree with the model set to the General matrix model by Le and Gascuel (LG) following an alignment performed using MAFFT (linsi, Katoh et al. 2013). Uncorrected p-distances for each alignment used to build the trees used for the pairwise tree comparisons were calculated using the mdist package from EMBOSS.

Comparisons of phylogenetic trees

The pairwise distances between trees were calculated by 4 different tree comparison methods (ALIGN, Nye et al. 2006, NODE, Williams et al. 1971, RF, Robinson and Foulds 1981 and RFK, Kuhner et al. 1994) (described in Kuhner and Yamato, 2015) and the Kernel Density Estimate (KDE) of the histogram resulting from the pairwise distances between trees (the lower triangular matrix from Figure 1) is plotted in Figure 2. Pairs of distributions were compared using a two-sample KS test to test if the two distributions are similar. (Please see supplementary table S1 for the full list of p-values for each comparison)

Neighbour-Net analysis

Alignments for the ribosomal proteins from PCGs and asgard MAG sample were concatenated and used to draw a Neighbor-Net (Bryant and Moulton 2004) using SplitsTree5. Randomization of the taxonomic assignments for the PCGs alignments were achieved using a custom python script to generate an alignment where r-proteins from different organisms have the same label. For the case where MK-D1 was added to the concatenated alignment the sequences for the 39 universal archaeal proteins were identified using BLAST with an identity of 25% and e-value of 1.0 x e^{-10} against its genome (PRJNA557562) and then subsequently added to the 30 MAGs in the asgard MAGs and Archaeal PCG sample respectively. The concatenated alignment was used to draw Neighbor-Nets using SplitsTree5. The Neighbour-Net for the reversed alignments (Supplementary Figure S13) was constructed using the same procedure. The ribosomal proteins from PCGs and the asgard MAG sample were reversed using an inhouse perl script, and aligned according to the Heads or Tails method (Landan and Graur, 2007) for phylogenetic analysis. The aligned columns in the reverse alignments were reversed once more to allow visual comparison of site patterns in the forward and reverse alignment.

BMGE based Trimming

In order to check if trimming the alignment to only retain sites with sufficient information Block Mapping and Gathering Entropy (BMGE, Criscuolo et al. 2010) was used with the default settings with the BLOSUM30 substitution matrix. To ensure uniformity all sequences of a protein (from each of the 10 reference samples and the MAGs) were combined together and aligned again with MAFFT (linsi). This combined alignment was then trimmed with BMGE and then separated into the 11 samples respectively. Trees were then drawn from the trimmed alignments as described before and the trees were compared.

Incompatibility scores for a set of phylogenetic trees with equal OTUs

For a set of phylogenetic trees T we calculated incompatibility scores for each tree t in the set similarly to (Nelson-Sathi et al, 2015). Each n OTU tree in the set was decomposed into its (n-3) splits. A split, s_1 from tree t_1 is considered incompatible with tree t_2 if s_1 is incompatible with at least one split from t_2. The incompatibility of a split s with the complete set of trees T is defined as the fraction of trees in T that are incompatible with s. Finally, the incompatibility of a tree t with a reference set of trees T is defined as the mean incompatibility observed among its splits. The differences in the distributions of tree incompatibility scores for the two sets of trees was assessed using the two-tailed Kolmogorov–Smirnov test.

Root to tip distance

All the phylogenetic trees computed as described above were rooted using MAD (Version 2.2) (Tria et al. 2017). The rooted trees were used to calculate the distance from the root to the 30 tips in each tree to calculate the average root to tip distance. The average root to tip distances for each of the genes in the metagenome samples were compared to their respective reference samples. A Kolmogorov–Smirnov test was performed to verify whether the root to tip distances for the two sets of trees are drawn from the same distribution.

Acknowledgements

This study was supported by the National Science Foundation of China (Nos. 91851210, 41530105), the Shenzhen Key Laboratory of Marine Archaea Geo-Omics, the Shenzhen Science and Technology Innovation Commission (JCYJ20180305123458107), Southern University of Science and Technology (No. ZDSYS20180208184349083), and the Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (No. K19313901). This research was also funded by the Moore–Simons Project on the Origin of the Eukaryotic Cell GBMF9743 to W.F.M. W.F.M. also would like to thank the ERC (666053), the VW Foundation (93046 and 96742), and the DFG (Ma 1426/21-1) for funding.

Author contributions.

S.G.G., N.K. and W.L. collected data and performed the computations and tests and contributed equally towards the manuscript. F.D.K.T. implemented the split incompatibility scores. All authors read the manuscript contributed to the final version of the manuscript. There are no conflicts of interests between the co-authors.

Data Availability Statements

The data underlying this article are available upon request and on https:// researchdata.hhu.de/.

References

Akıl, C. & Robinson, R. C. Genomes of asgard archaea encode profilins that regulate actin. *Nature* **562**, 439–443 (2018).

Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* **7**, 13219 (2016).

Bernard, G., Pathmanathan, J. S., Lannes, R., Lopez, P. & Bapteste, E. Microbial dark matter investigations: How microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery. *Genome Biol Evol*. **10,** 707-715 (2018).

Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

Breitwieser, F.P, Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* **20**, 1125–1136 (2017).

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. Universal trees based on large combined protein sequence data sets. *Nat Genet* **28**, 281–285 (2001).

Bryant, D. & Moulton, V. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* **21**, 255–265 (2004).

Criscuolo, A. & Gribaldo, S. BMGE (Block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* **10**, 210 (2010).

Cunha, V. D., Gaia, M., Gadelle, D., Nasir, A. & Forterre, P. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *Plos Genet* **13**, e1006810 (2017).

Cunha, V., Gaia, M., Nasir, A. & Forterre, P. Asgard archaea do not close the debate about the universal tree of life topology. *PLOS Genet* **14**, e1007215 (2018).

Daubin, V., Gouy, M. & Perrière, G. A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. *Genome Res* **12**, 1080–1090 (2002).

Dell'Anno, A. & Danovaro, R. Extracellular DNA plays a key role in deep-sea ecosystem functioning. *Science* **309**, 2179–2179 (2005).

Dell'Anno, A., Stefano, B. & Danovaro, R. Quantification, base composition, and fate of extracellular DNA in marine sediments. *Limnol Oceanogr* **47**, 899–905 (2002).

Dey, G., Thattai, M. & Baum, B. On the archaeal origins of eukaryotes and the challenges of inferring phenotype from genotype. *Trends Cell Biol.* **26**, 476–485 (2016).

Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nature Rev Micro* **15**, 711–723 (2017).

Fan, L. *et al*. Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria. *Nat Ecol Evol*, **4,** 1213–1219 (2020).

Gould, S. B., Garg, S. G. & Martin, W. F. Bacterial vesicle secretion and the evolutionary origin of the eukaryotic endomembrane system *Trends Microbiol* **24**, 525-534 (2016).

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5**, R245–R249 (1998).

Hansmann, S. & Martin, W. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *Int J Syst Evol Microbiol* **50**, 1655–1663 (2000).

Hug, L. A. *et al.* A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).

Imachi, H. *et al*. Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).

Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, (2019).

Katoh, K. & Standley, D. M. MAFFT Multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**, 772–780 (2013).

Kuhner, M. & Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol* **11**, 459–468 (1994).

Kuhner, M. K. & Yamato, J. Practical performance of tree comparison metrics. *Syst Biol* **64**, 205 214 (2015).

Landan, G., & Graur, D. Heads or Tails: A simple reliability check for multiple sequence alignments. *Mol Biol Evol* **24**, 1380–1383 (2007).

Marcy, Y. *et al*. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci USA* **104**, 11889 (2007).

Martin, W. *et al.* Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* **393**, 162–165 (1998).

Martin, W. F., Tielens, A. G. M., Mentel, M., Garg, S. G. & Gould, S. B. The physiology of phagocytosis in the context of mitochondrial origin. *Microbiol Mol Biol R* **81**, e00008-17 (2017).

Martin, W. Gene transfer from organelles to the nucleus: Frequent and in big chunks. *Proc National Acad Sci* **100**, 8612–8614 (2003).

Matte-Tailliez, O., Brochier, C., Forterre, P. & Philippe, H. Archaeal phylogeny based on ribosomal proteins. *Mol Biol Evol* **19**, 631–639 (2002).

Nagler, M., Insam, H., Pietramellara, G. & Ascher-Jenull, J. Extracellular DNA in natural environments: features, relevance and applications. *Appl Microbiol Biot* **102**, 6343–6356 (2018).

Nelson-Sathi, S. *et al.* Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc Natl Acad Sci USA* **109**, 20537–20542 (2012).

Nelson-Sathi, S. *et al.* Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**, 77–80 (2015).

Nesbø, C. L., Boucher, Y. & Doolittle, F. W. Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. *J Mol Evol* **53**, 340–350 (2001).

Nye, T., Liò, P. & Gilks, W. R. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics* **22**, 117–119 (2006).

O'Leary, N. A. *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**.** *Nucleic Acids Res* **4**, 733-745 (2016).

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

Penny, D., Foulds, L. R. & Hendy, M. D. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature* **297**, 197–200 (1982).

Rinke, C. *et al*. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431-437 (2013).

Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Mathematical Biosciences* **53**, 131–147 (1981).

Shen, X.-X., Hittinger, C. T. & Rokas, A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecol Evol* **1**, s41559-017–0126 (2017).

Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173-179 (2015).

Spang, A., *et al*. Asgard archaea are the closest prokaryotic relatives of eukaryotes. *Plos Genet* **14**, e1007080 (2018).

Steel, M., Huson, D. & Lockhart, P. J. Invariable sites models and their use in phylogeny reconstruction. *Systematic Biol* **49**, 225–232 (2000).

Stephen, J., McCaig, A., Smith, Z., Prosser, J. & Embley, T. Molecular diversity of soil and marine 16S rRNA gene sequences related to beta-subgroup ammonia-oxidizing bacteria. *Appl Environ Microbiol* **62**, 4147–54 (1996).

Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biol* **56**, 564–577 (2007).

The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**, D506-D515 (2019).

Thiergart, T., Landan, G. & Martin, W. F. Concatenated alignments and the case of the disappearing tree. *Bmc Evol Biol* **14**, 266 (2014).

Torsvik, V., Sørheim, R. & Goksøyr, J. Total bacterial diversity in soil and sediment communities—A review. *J Ind Microbiol* **17**, 170–178 (1996).

Torti, A., Lever, M. A. & Jørgensen, B. B. Origin, dynamics, and implications of extracellular DNA pools in marine sediments. *Mar Genomics* **24**, 185-196 (2015)

Tria, F. D. K., Landan, G. & Dagan, T. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol* **1**, 0193 (2017).

Williams, W. & Clifford, H. On the comparison of two classifications of the same set of elements. *Taxon* **20**, 519–522 (1971).

Zaremba-Niedzwiedzka, K. *et al*. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

Figure Legends

Figure 1 | Pairwise Robinson-Foulds distance between trees for universal archaeal proteins. Pairwise distance between phylogenetic trees of 39 universal proteins were calculated using the Robinson-Foulds metric and plotted in **(A)** for a sample of 30 archaeal genomes from RefSeq (archaeal PCGs; Supplementary Table S1) and **(B)** for a sample of 30 archaeal metagenomes including asgards (asgard archaeal MAGs; Method section; Supplementary Table S2). The differences among the phylogenetic trees for the proteins in archaeal PCGs reflect the natural variation for sequenced genomes from cultured archaea. The asgard archaeal MAGs, while having a lower degree of congruence between the trees overall, cluster into two major discernable groups with one composed largely of ribosomal proteins. It is evident that r-protein trees in archaeal MAGs are more similar to each other than trees for non-ribosomal proteins. The scale bar applies to both panels.

Figure 2 | Distribution of pairwise tree-distances between PCGs and MAGs: The pairwise comparisons of tree distances computed using four different metrics (see Methods) are shown. In each case, a matched set of proteins present in MAGs and 10 random samples from RefSeq to make up the PCG samples are taken to plot comparable distributions. The MAG sample is always shown in red. **(A)** Trees for 39 universal proteins from 30 asgard archaeal MAGs are compared with trees for the 39 homologues from 10 samples of 30 archaeal PCGs **(B)** Trees for 23 ribosomal proteins from 30 asgard archaeal MAGs are compared with those from 10 samples of 30 archaeal PCGs. Note that the mean topological distance between trees is higher for the asgard archaeal MAGs when compared with archaeal PCGs. **(C)** Trees for 16 non-ribosomal proteins from 30 asgard archaeal MAGs are compared with those from 10 samples of 30 archaeal PCGs. **(D)** Trees for 16 ribosomal proteins from 30 candidate phyla radiation (CPR) MAGs from Hug et al., 2016 are compared with those from 10 samples comprising of 30 bacterial PCGs each. **(E)** Trees for 20 ribosomal proteins from a non-CPR bacterial MAG sample are compared with those from 10 samples of 30 bacterial PCGs each. In all panels, blue curves represent the respective 10 independent PCG samples while the red curve represents the respective MAGs. The largest FDR corrected p-value (two-tailed Kolmogorow-Smirnow test) from comparisons between the MAG sample and the respective PCG samples is indicated in red while the smallest p-value is indicated in blue (see Methods).

Figure 3 | Comparison of mean root to tip distances between MAGs and PCGs. (A) The mean root to tip distances of rooted trees for each of the 39 universal archaeal proteins from the asgard archaeal metagenomes vs an archaeal reference sample (the 16 non-ribosomal proteins are shown in purple) (B) The mean root to tip distance of rooted trees for each of the 16 universal ribosomal proteins from CPR metagenomes vs a bacterial reference sample (C) The mean root to tip distance of rooted trees for each of the 21 universal ribosomal proteins from bacterial metagenomes vs a bacterial reference sample. The FDR corrected p-value (two-tailed Kolmogorow-Smirnow test) from comparisons between the MAG sample and each RefSeq sample is indicated. The p-values for the 16 non ribosomal proteins is indicated separately in purple (see Methods).

Figure 4 | Distribution of pairwise tree-distances between RefSeq and metagenomes evaluated by CheckM: The pairwise comparisons of tree distances computed using four different metrics (see Methods) are shown. In each case, a matched set of proteins present a MAG sample and 10 samples from RefSeq are taken to plot comparable distributions. The 30 MAGs in each sample is taken such that their completeness as evaluated by CheckM is at least 70, 80 or 90% as shown in the inset. (A) Trees for 39 universal proteins from 3 samples of 30 archaeal MAGs each are compared with trees for the 39 homologues from 10 samples of 30 archaeal RefSeq genomes (B) Trees for 23 ribosomal proteins from 3 samples of 30 archaeal MAGs each are compared with trees for the 39 homologues from 10 samples of 30 archaeal RefSeq genomes (C) Trees for 16 non-ribosomal proteins from 3 samples of 30 archaeal MAGs each are compared with trees for the 39 homologues from 10 samples of 30 archaeal RefSeq genomes. Individual p-values for each comparison are given in Table S8.

Figure 5 | Neighbor-Nets reconstructed from concatenated alignments of 23 ribosomal proteins for Archaeal PCGs and asgard archaeal MAGs. (A) The Neighbor-Net of a concatenated alignment of 23 ribosomal proteins in the archaeal PCGs sample shows very little conflict throughout, resulting in a tree-like network with 16 well supported splits (indicated with red dots). (B) A Neighbor-Net drawn from a concatenated alignment of the same 23 ribosomal proteins from asgard archaeal MAGs results in a network with a star-like structure. The insets magnify the central area of interest to better highlight the difference of signals of the two networks (C) Before generating a concatenated alignment, the 23 r-proteins from the 30 genomes in the archaeal PCGs sample were randomly redistributed. These scrambled genomes,

indicated with the prefix ‘rnd’, were used to reconstruct a Neighbor-Net, which generated a star-like structure very similar to that of the asgard archaeal MAGs Neighbor-Net shown in panel B.

Figure 6 | Tree compatibility scores for samples of tree reconstructed from PCGs and MAGs. Cumulative distribution of tree incompatibility scores within sets of gene trees. In each case every curve represents a set of 30 organisms where the RefSeq samples are shown in shades of blue and the MAG sample is always shown in red (A) Trees for 39 universal proteins sampled from 10 archaeal RefSeq genomes vs asgard achaeal MAGs, (B) Trees for a subset of 23 ribosomal proteins sampled from 10 archaeal RefSeq genomes vs asgard archaeal MAGs (C) Trees for the complement set of 16 non-ribosomal proteins sampled from 10 archaeal RefSeq genomes vs asgard archaeal MAGs. (D) Trees for 16 ribosomal proteins sampled from 10 bacterial RefSeq genomes vs CPR MAGs and (e) Trees for 20 ribosomal proteins sampled from 10 bacterial RefSeq genomes vs non-CPR bacterial MAGs.

Figure 1

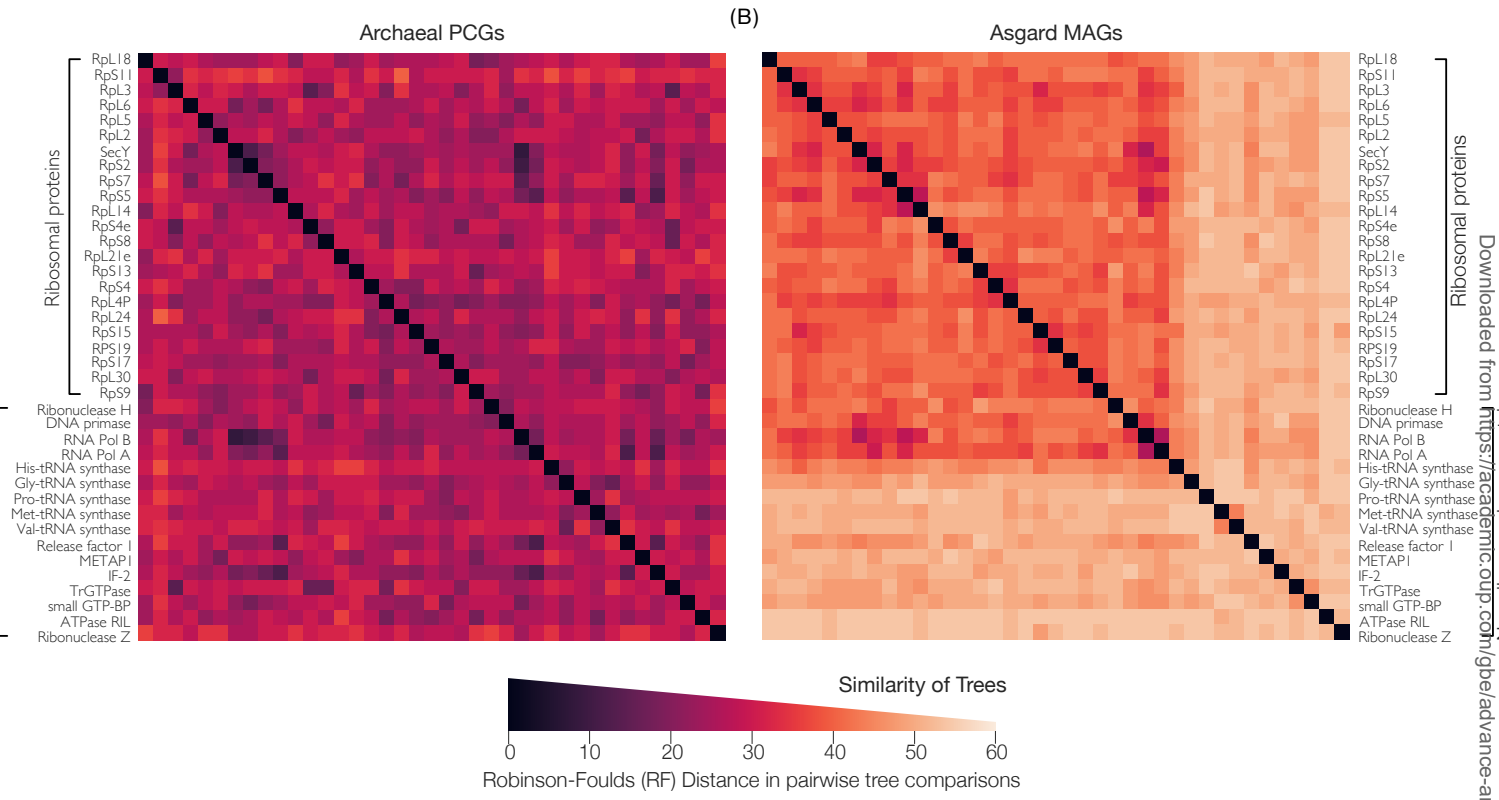


Figure 2

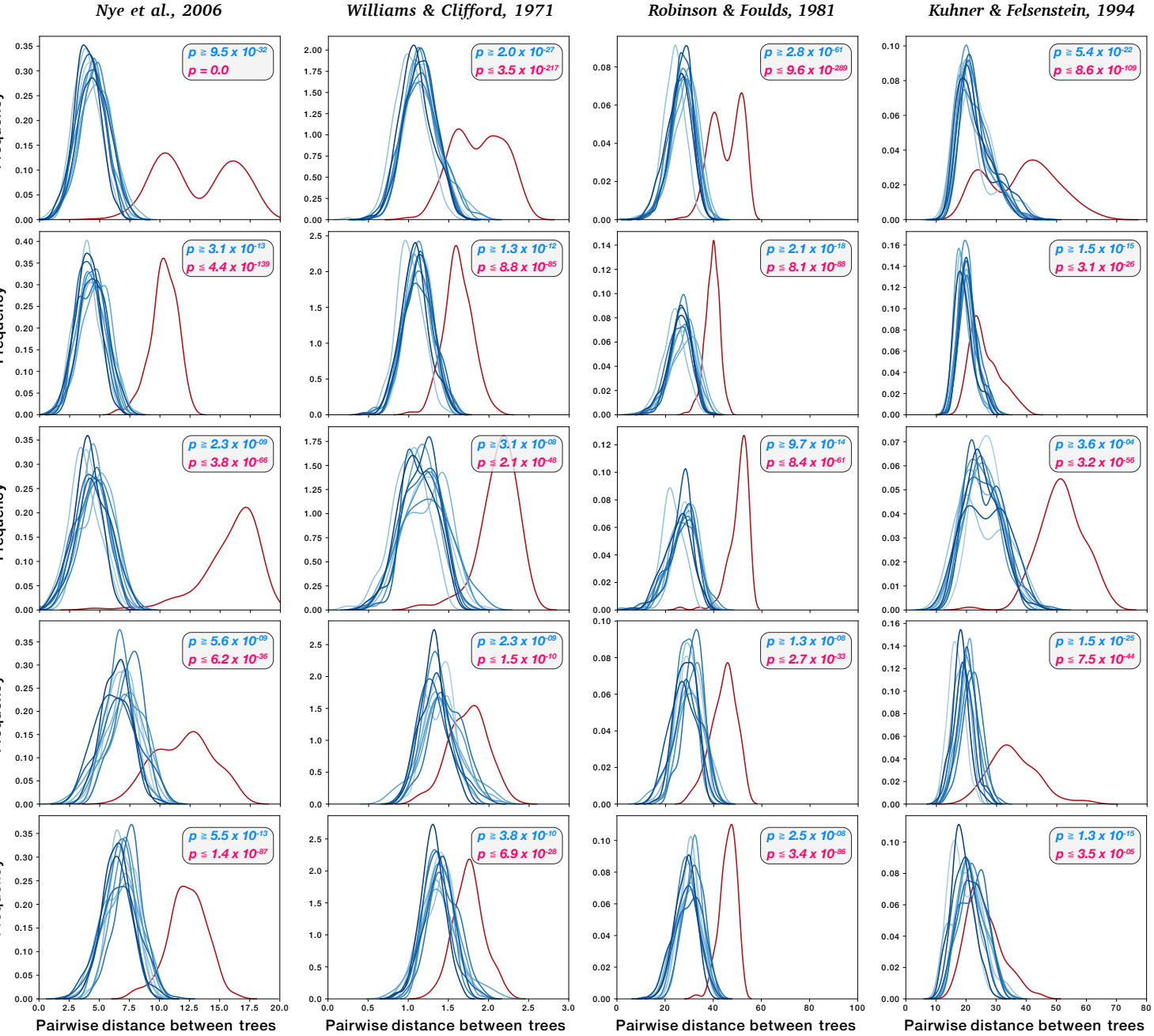


Figure 3

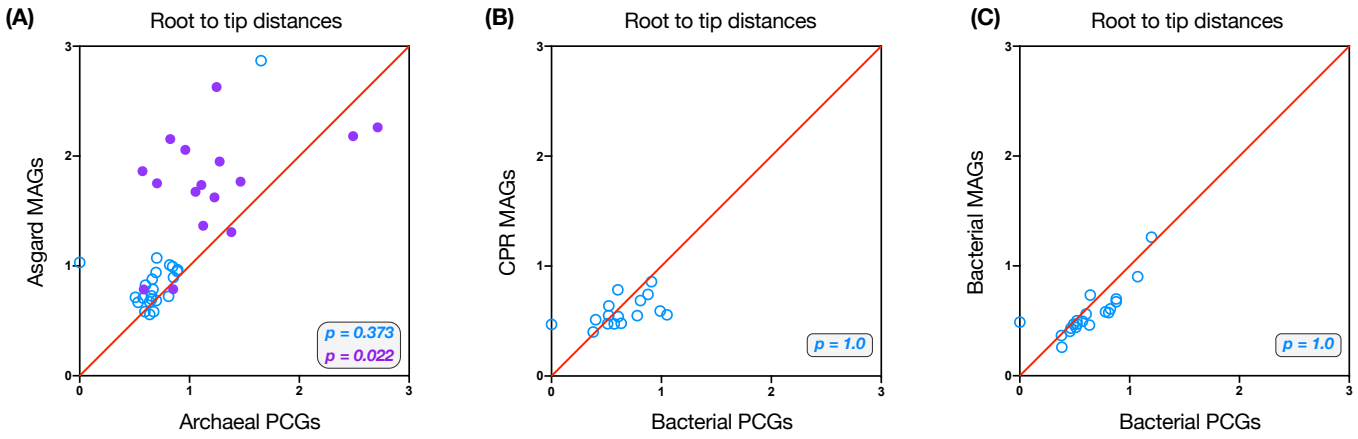
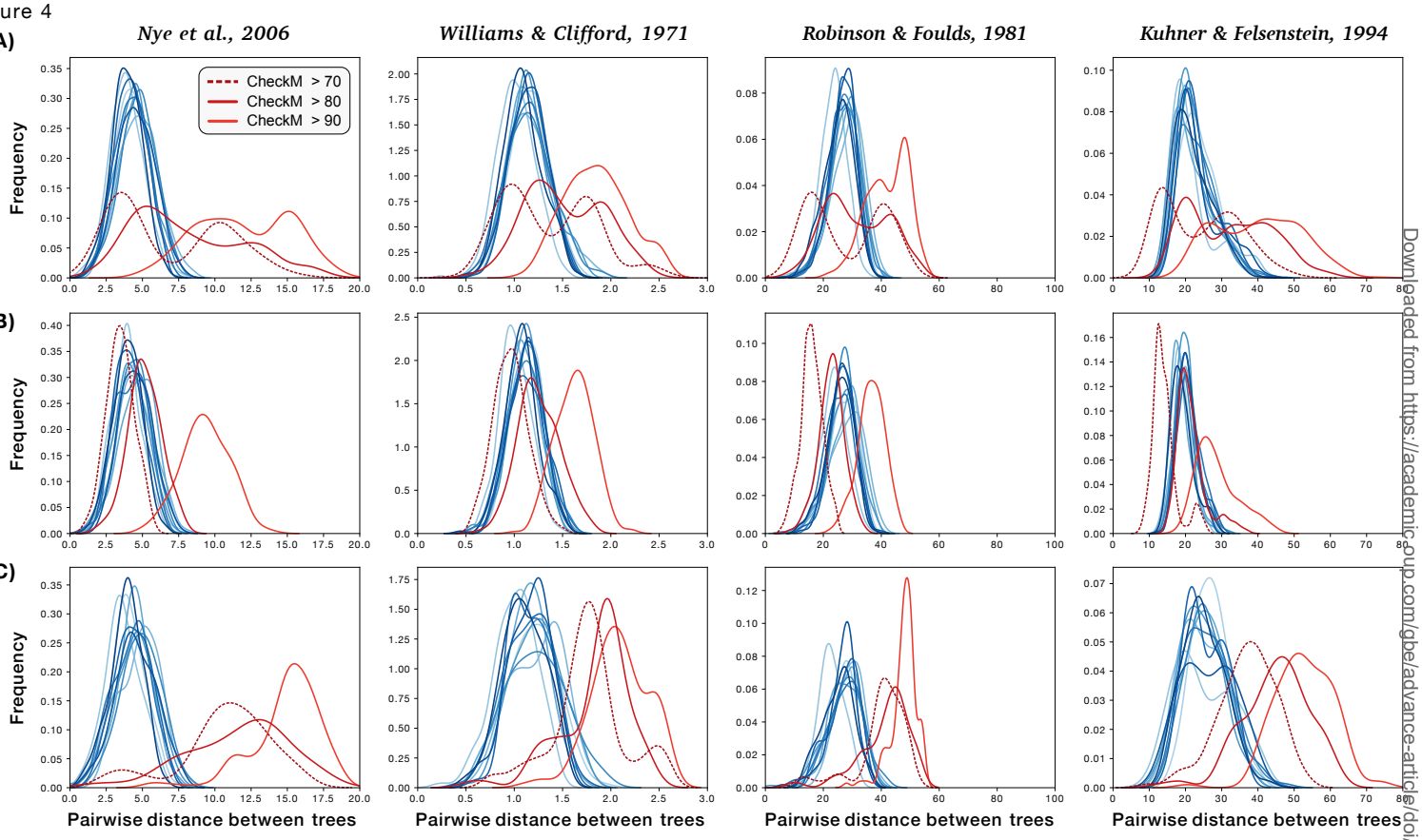


Figure 4



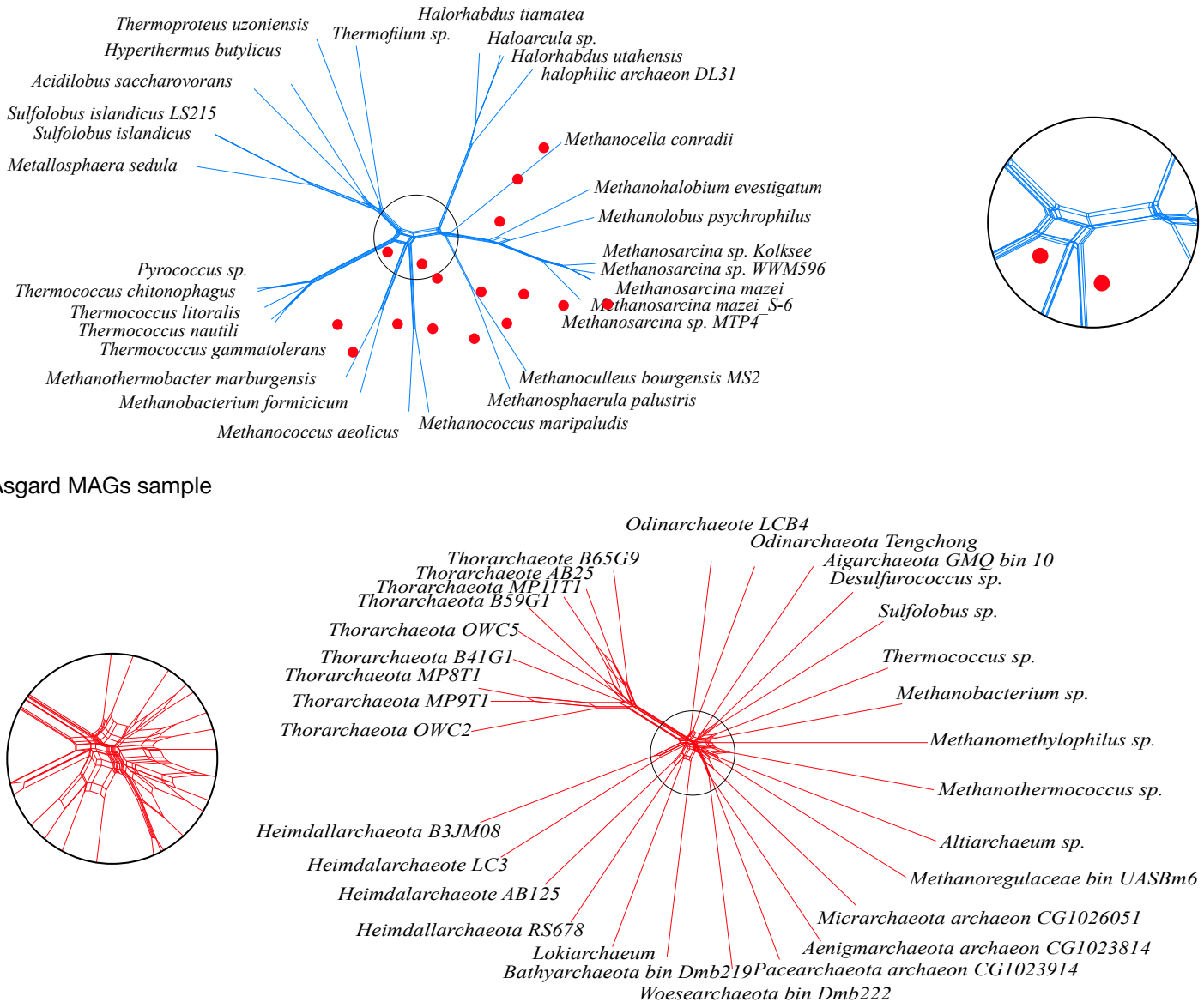


Figure 6