











# Anomalous Phylogenetic Behavior of Ribosomal Proteins in Metagenome-Assembled Asgard Archaea

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

<sup>1</sup>Institute for Molecular Evolution, Heinrich Heine University Düsseldorf, Germany

<sup>2</sup>Putuo People's Hospital, School of Life Sciences and Technology, Tongji University, Shanghai, China

<sup>3</sup>Transdisciplinary Biology, Computational Biology Laboratory, Rajiv Gandhi Centre for Biotechnology, Thiruvananthapuram, India

<sup>4</sup>Shenzhen Key Laboratory of Marine Archaea Geo-Omics, Southern University of Science and Technology, Shenzhen, China

<sup>5</sup>Department of Ocean Science and Engineering, Southern University of Science and Technology, Shenzhen, China

<sup>6</sup>SUSTech Academy for Advanced Interdisciplinary Studies, Southern University of Science and Technology, Shenzhen, China

<sup>7</sup>Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou, China

<sup>†</sup>These authors contributed equally to this work.

\*Corresponding authors: E-mails: gargs@hhu.de; bill@hhu.de.

Accepted: 3 November 2020

## 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 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.

**Key words:** metagenomics, binning, eukaryogenesis, phylogenetics, asgard archaea, candidate phylum radiation (CPR).

## Significance

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.

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This article is published and distributed under the terms of the Oxford University Press, Standard Journals Publication Model ([https://academic.oup.com/journals/pages/open\\_access/funder\\_policies/chorus/standard\\_publication\\_model](https://academic.oup.com/journals/pages/open_access/funder_policies/chorus/standard_publication_model))

## 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 metagenome-assembled genomes (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. 2019). Binning is a crucial step in the reconstruction of MAGs because it assigns metagenomic sequences to a genome, a procedure that is not required in the case of cultured organisms.

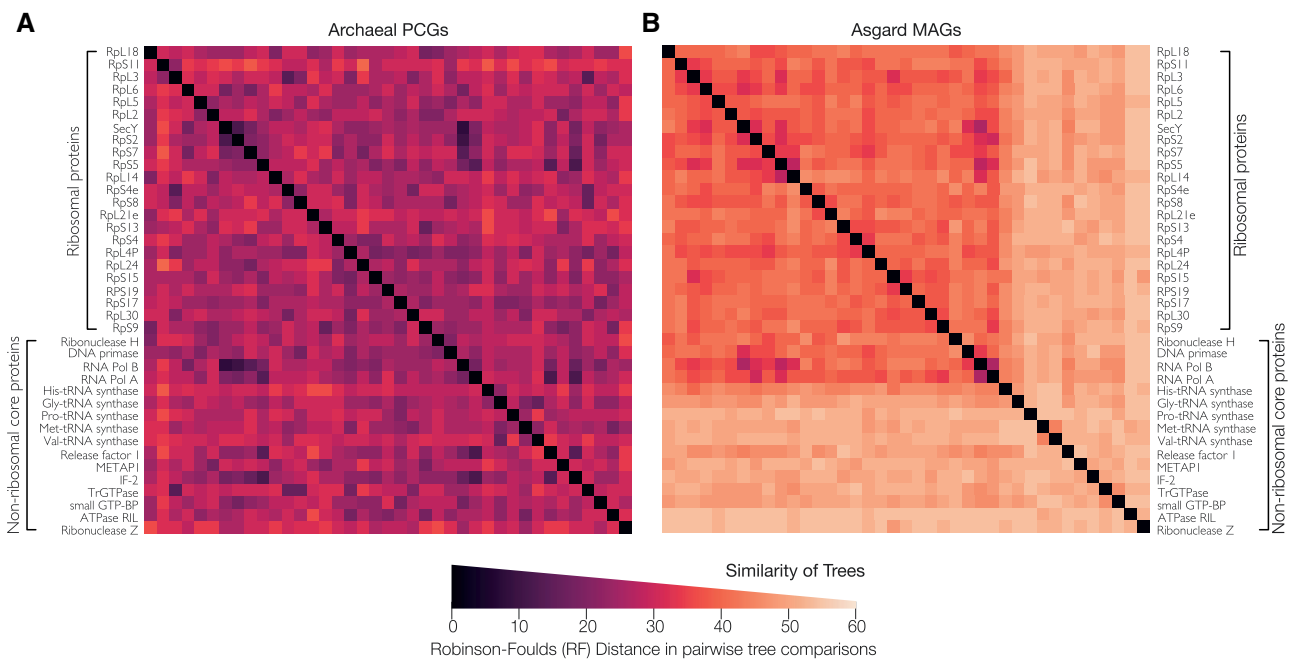
Because rRNA has limited phylogenetic resolution, concatenated sequences of ribosomal proteins (r-proteins) and other universally distributed proteins are now commonly used for phylogeny even though concatenation has its own caveats (Thiergart et al. 2014). This practice is well established with over 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 20–30 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; Akil and Robinson 2018; Cunha et al. 2018); especially in light of the noncomplex phenotype of the first enriched and imaged cells from 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; Stephen et al., 1996; Martin, 1998 (Steel et al., 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 Materials and Methods; [supplementary figs. S1 and S2](#) and [tables S1 and S2, Supplementary Material](#) online) 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 Robinson–Foulds pairwise distance measure (Robinson and Foulds 1981) (fig. 1A). We asked the same question for the same proteins for asgard MAGs. The result (fig. 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 cotranslational 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 fig. 1B, see scale bar). The evident bimodal distribution of phylogenetic behavior for the 39 asgard MAG proteins is not observed for PCGs.

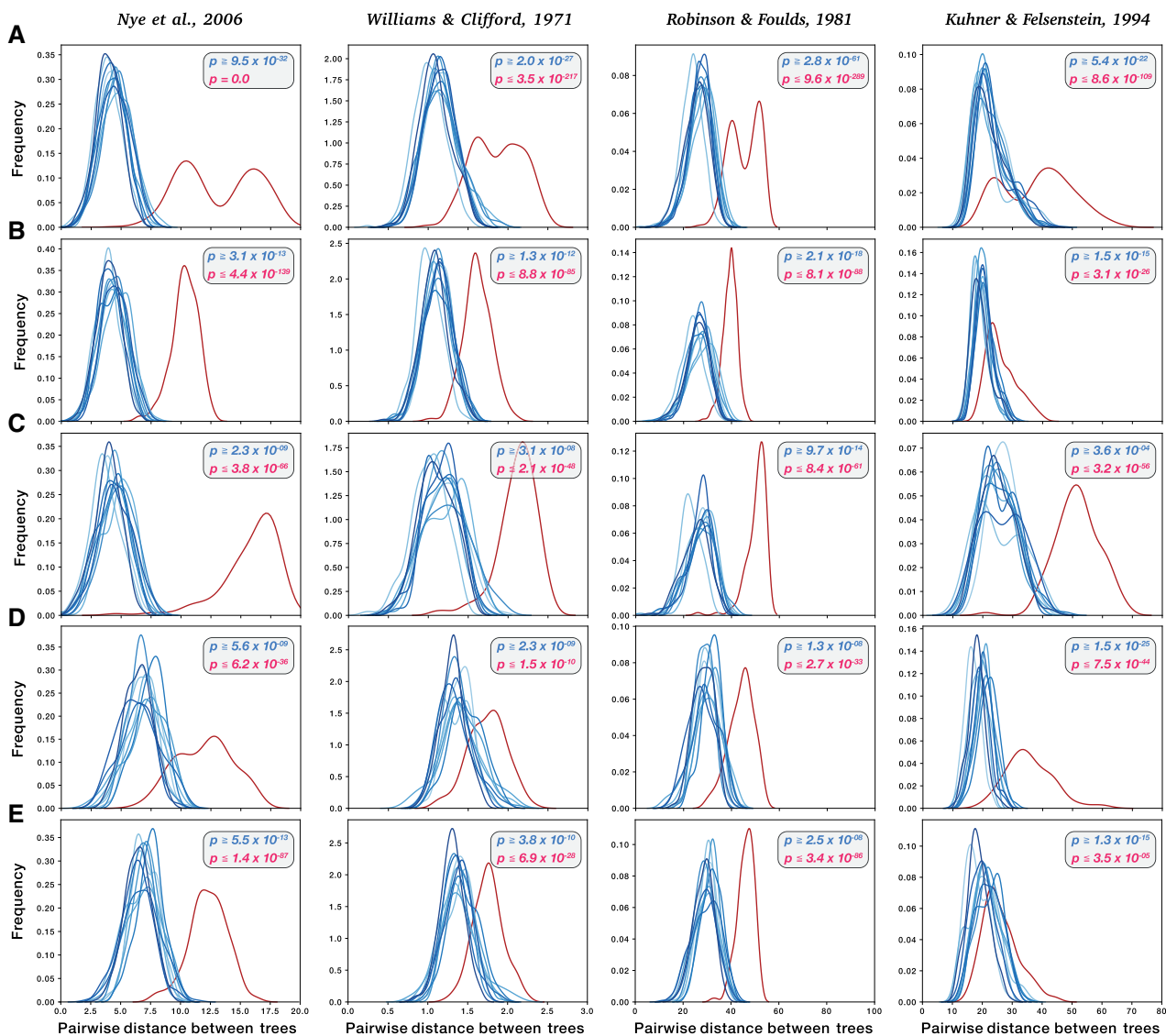


**FIG. 1.**—Pairwise Robinson–Foulds distance between trees for universal archaeal proteins. Pairwise distance between phylogenetic trees of 39 universal proteins was calculated using the Robinson–Foulds metric and plotted in (A) for a sample of 30 archaeal genomes from RefSeq (archaeal PCGs; [supplementary table S1, Supplementary Material](#) online) and (B) for a sample of 30 archaeal metagenomes including asgards (asgard archaeal MAGs; Materials and Methods section; [supplementary table S2, Supplementary Material](#) online). 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 nonribosomal proteins. The scale bar applies to both panels.

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 ([fig. 2A](#)). Note that the distributions for trees from archaeal MAGs are always shifted toward 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) ([fig. 2B and C](#)). 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 ([fig. 2B](#)) (see Materials and Methods and [supplementary table S3B, Supplementary Material](#) online). The probability that the distribution of tree dissimilarities across nonribosomal 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

Materials and Methods and [supplementary table S3C, Supplementary Material](#) online).

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 figs. S3 and S4, Supplementary Material](#) online). Despite our careful genome selection procedure to assemble bacteria and archaea data sets to reflect the known phylum-level diversity ([supplementary figs. S1 and S2, Supplementary Material](#) online) (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, then MAG-derived trees will inevitably show reduced congruence as a result of artifacts 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

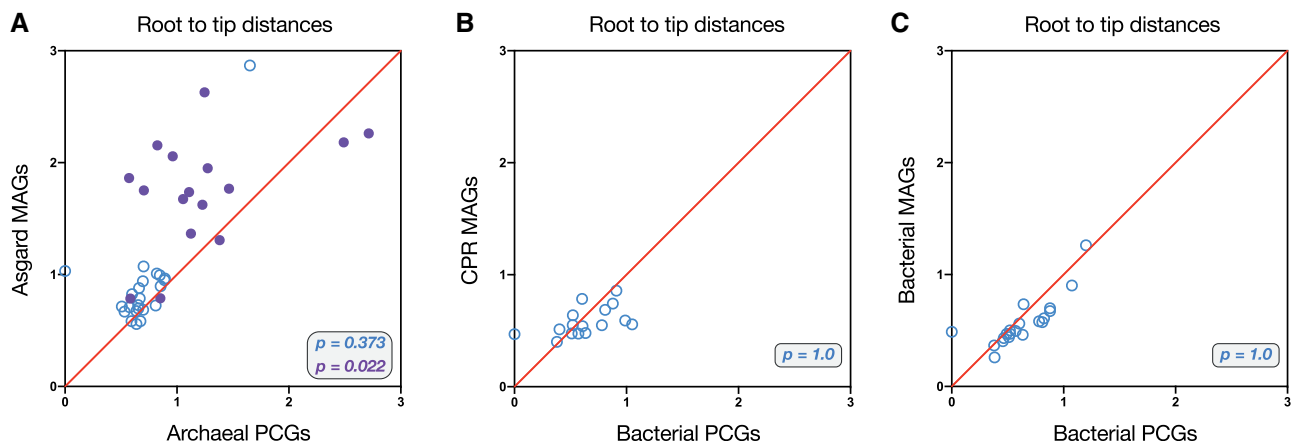


**FIG. 2.**—Distribution of pairwise tree-distances between PCGs and MAGs: The pairwise comparisons of tree distances computed using four different metrics (see Materials and Methods) are shown. In each case, a matched set of proteins present in MAGs and ten random samples from RefSeq to make up the 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 ten samples of 30 archaeal PCGs. (B) Trees for 23 ribosomal proteins from 30 asgard archaeal MAGs are compared with those from ten 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 nonribosomal proteins from 30 asgard archaeal MAGs are compared with those from ten 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 ten samples comprising 30 bacterial PCGs each. (E) Trees for 20 ribosomal proteins from a non-CPR bacterial MAG sample are compared with those from ten samples of 30 bacterial PCGs each. In all panels, blue curves represent the respective ten independent PCG samples, whereas the red curve represents the respective MAGs. The largest FDR corrected  $P$  value (two-tailed Kolmogorow–Smirnov test) from comparisons between the MAG sample and the respective PCG samples is indicated in red, whereas the smallest  $P$  value is indicated in blue (see Materials and Methods).

independently, measured as the mean root-to-tip distances across the rooted topologies (see Materials and Methods for details). For the asgard metagenomes, we found that although the root to tip distances for the 16 nonribosomal proteins are significantly different from the archaeal PCGs, the 23 ribosomal proteins including SecY agree with each other (fig. 3A and [supplementary table S4, Supplementary](#)

[Material online](#)). This recapitulates what we have observed with the bimodal distribution for tree dissimilarity scores (figs. 1 and 2A–C) 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



**Fig. 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 versus an archaeal reference sample (the 16 nonribosomal 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 versus 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 versus 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 nonribosomal proteins is indicated separately in purple (see Materials and Methods).

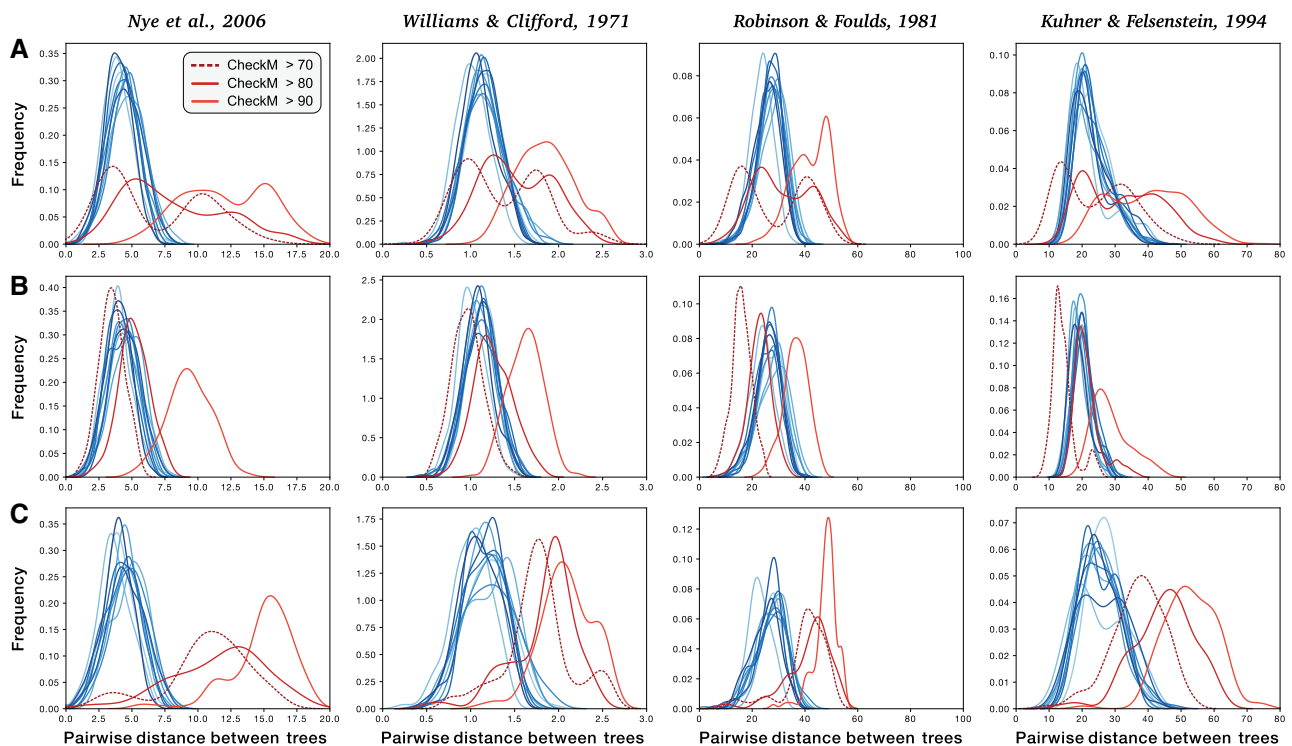
Castresana 2007; Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017; Fan et al. 2020). To check whether site filtering affects the phylogenetic anomaly of asgard MAGs, we trimmed alignments (see Materials and Methods) as in earlier analyses and recalculated the trees and comparisons. As shown in [supplementary figure S7, Supplementary Material online](#), site filtering does not improve MAG phylogenetic behavior ( $P$  values in [supplementary table S5, Supplementary Material online](#)). The asgard MAGs have a systematic and previously undetected phylogenetic anomaly.

Is the MAG phylogenetic anomaly asgard specific? We examined ten additional nonasgard archaeal MAGs from various metagenome sequencing projects by repeating the tree comparison metrics (see Materials and Methods). Ribosomal protein samples from nonasgard 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 data set are significantly different ([supplementary fig. S8A–E and table S6, Supplementary Material online](#)). Furthermore, within the asgard Metagenomes sample, we have used 16 published asgard MAGs together with 14 nonasgard archaeal MAGs. In order to determine the effect of mixing asgard and nonasgard MAGs, we replaced the  $r$ -proteins of the 14 nonasgard MAGs with nonasgard archaeal MAGs from the Marine group II ([supplementary table S2, Supplementary Material online](#)) and with  $r$ -proteins from archaeal PCGs. Although 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 fig. S9 and table S7, Supplementary Material online](#)) 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 nonasgard 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 ([fig. 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 with the reference ([fig. 2E](#)). The investigation of ten additional samples of non-CPR and CPR MAGs generated similar results ([supplementary fig. S8D and E, Supplementary Material online](#)). 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, whereas 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, Supplementary Material online](#)). Since we are primarily comparing metagenome samples to PCGs, this difference can be overlooked for the purposes of our conclusions ([fig. 3B and C; supplementary table S4, Supplementary Material online](#)). 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



**FIG. 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 Materials and Methods) are shown. In each case, a matched set of proteins present a MAG sample and ten 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 three samples of 30 archaeal MAGs each are compared with trees for the 39 homologues from ten samples of 30 archaeal RefSeq genomes. (B) Trees for 23 ribosomal proteins from three samples of 30 archaeal MAGs each are compared with trees for the 39 homologues from ten samples of 30 archaeal RefSeq genomes. (C) Trees for 16 nonribosomal proteins from three samples of 30 archaeal MAGs each are compared with trees for the 39 homologues from ten samples of 30 archaeal RefSeq genomes. Individual *P* values for each comparison are given in [supplementary table S8, Supplementary Material](#) online.

completeness and contamination of a given MAG. Although 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 (fig. 4). Phylogenetic incongruence of trees in MAGs with 90% completeness or more is higher than those with a completeness between 70% and 80% (fig. 4 and [supplementary table S8, Supplementary Material](#) online). 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 congruence 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 fig. S10 and table S9, Supplementary Material](#) online). 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 2003). The NNet for r-proteins sampling crenarchaeal and euryarchaeal PCGs is extremely tree-like, with 16 strong, low conflict splits indicated in the figure (fig. 5A), reflecting highly congruent phylogenetic signals across r-proteins from diverse cultivated archaea (figs. 1A and 2B). A NNet for asgard MAGs is shown in figure 5B, revealing one split in the data, which separates Thorarchaea 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 (fig. 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 fig. S13, Supplementary Material online), 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 nonasgard archaeal MAG sample (supplementary fig. S11, Supplementary Material online) 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 Materials and 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), whereas trees for r-proteins from asgard MAGs differ from each other (fig. 6 and supplementary fig. S12, Supplementary Material online). The differences in cumulative distribution frequencies are obvious (fig. 6) and highly significant ( $P = 10^{-18}$ , two-tailed Kolmogorov–Smirnov test; supplementary tables S10 and S11, Supplementary Material online).

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 artifacts. Although the names and rank of asgard archaea is defined by their r-protein phylogenies (Spang et al. 2015, 2018; Eme et al. 2017; Zaremba-Niedzwiedzka et al. 2017), 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 nonribosomal proteins, the issue might be even more serious (figs. 1 and 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  $\mu\text{m}$  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; Eme et al. 2017; Zaremba-Niedzwiedzka et al. 2017). There have been challenges posed to individual protein sequences present in asgard MAGs (Cunha et al. 2017, 2018; 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. 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.

## Materials and Methods

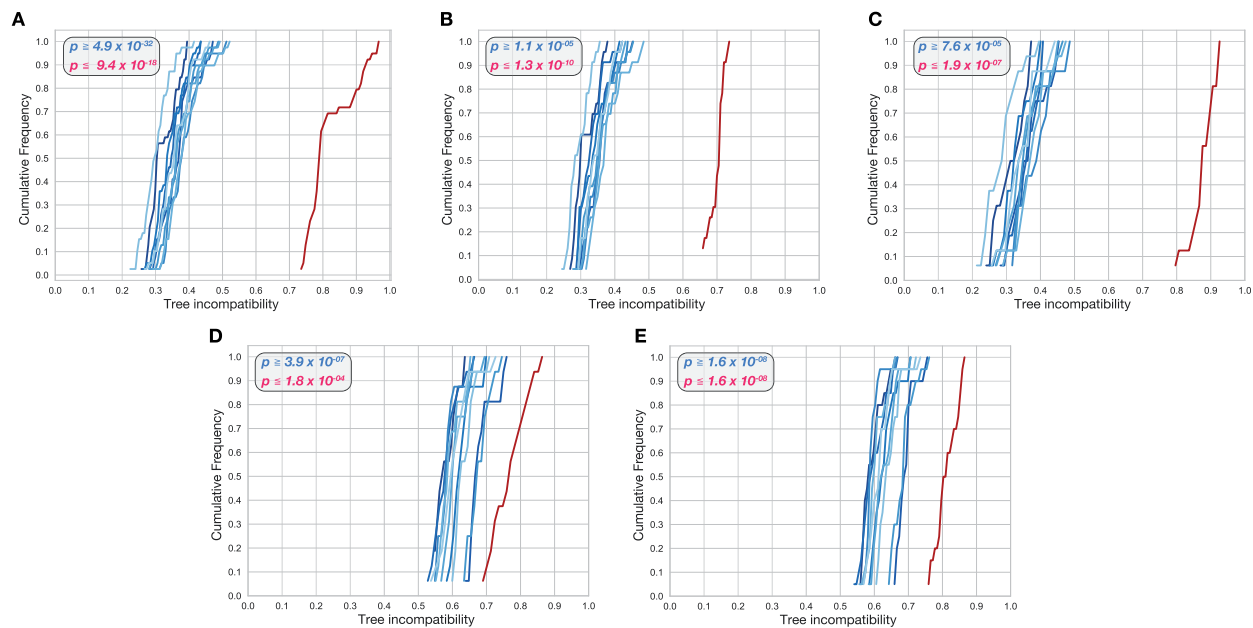
### Sources of Archaeal Metagenomics Data

The description and list of organisms comprising each MAG sample are given in supplementary table S2, Supplementary Material online. Archaeal Metagenomes comprising asgard archaea were downloaded from IMG (<https://img.jgi.doe.gov/>) and NCBI databases (<https://www.ncbi.nlm.nih.gov/>). About 30 archaeal MAGs that contain 39 universal archaeal proteins (23 ribosomal proteins + 16 nonribosomal proteins) were retained and constitute asgard MAG data set. In the extended data set, 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

FIG. 5—Continued

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).





**FIG. 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 ten archaeal RefSeq genomes versus asgard archaeal MAGs. (B) Trees for a subset of 23 ribosomal proteins sampled from ten archaeal RefSeq genomes versus asgard archaeal MAGs. (C) Trees for the complement set of 16 nonribosomal proteins sampled from ten archaeal RefSeq genomes versus asgard archaeal MAGs. (D) Trees for 16 ribosomal proteins sampled from ten bacterial RefSeq genomes versus CPR MAGs and (E) trees for 20 ribosomal proteins sampled from ten bacterial RefSeq genomes versus non-CPR bacterial MAGs.

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 data set, all qualified MAGs were screened with a criterion of completeness  $\geq 70\%$  and contamination  $\leq 5\%$  using CheckM (version: 1.0.13) (Parks et al. 2015). About 114 archaeal MAGs that contain all 39 universal proteins were retained and used to generate ten archaeal MAG samples containing 30 genomes each. Samples with different CheckM completeness ranges were subsampled from the afore-mentioned extended data set (supplementary table S1, Supplementary Material online), corresponding to completeness values of 70–80%, 80–90%, and 90–100% (supplementary table S8, Supplementary Material online).

#### Sources of CPR Metagenomics Data

The MAGs from CPR were obtained from the authors of Hug et al. (2016). About 140 CPR MAGs containing all 16 universal ribosomal proteins were used to create ten CPR MAG samples of 30 genomes each (supplementary table S2, Supplementary Material online). In addition, 23 closed genomes were kindly

provided by Prof. Karthik Anantharaman which along with seven high-quality draft genomes from Anantharaman et al. (2016) were used to create a MAG sample comprising high-quality MAGs.

#### Sources of Bacterial Metagenomics Data

The bacterial MAG data set was downloaded from the BioProjects with the accessions PRJNA288027 (2,545 assemblies) and PRJNA270657 (103 assemblies). This set was further subsampled for ten 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) (supplementary fig. S2 and table S2, Supplementary Material online).

#### Identification of Homologs of Ribosomal Proteins in RefSeq and Metagenomics Genomes

Universal protein clusters for Archaea were obtained from Nelson-Sathi et al. (2015), whereas 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 5,443 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 data set was further sub-sampled for groups of 30 organisms that reflect the full breadth of taxonomic distribution for the complete data set for bacteria and archaea, respectively, to generate ten 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 organisms from each phylum to generate ten 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](#) and [figs. S1 and S2](#), [Supplementary Material](#) online).

### 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 ten 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, [Kato and Standley 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 four different tree comparison methods (ALIGN, [Nye et al. 2006](#), NODE, [Williams and Clifford 1971](#), RF, [Robinson and Foulds 1981](#) and RFK, [Kuhner and Felsenstein 1994](#)) (as 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 [fig. 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](#), [Supplementary Material](#) online, for the full list of *P* values for each comparison.)

### Neighbor-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 2003](#)) using SplitsTree4. Randomization of the taxonomic assignments for the PCGs alignments was 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  $1.0 \times e^{-10}$  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 Neighbor-Net for the reversed alignments ([supplementary fig. S13](#), [Supplementary Material](#) online) 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 and visualised the Neighbor-Net using SplitsTree4 ([supplementary fig. S13](#)).

### BMGE-Based Trimming

In order to check if trimming the alignment to only retain sites with sufficient information, Block Mapping and Gathering Entropy (BMGE, [Crisuolo and Gribaldo 2010](#)) was used with the default settings with the BLOSUM30 substitution matrix. To ensure uniformity, all sequences of a protein (from each of the ten 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 were 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 with 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.

## Supplementary Material

Supplementary data is available at *Genome Biology and Evolution* online.

## Acknowledgments

This study was supported 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. Participation of RZ, WL, LF, and CZ in this research was supported by National Science Foundation of China (Nos. 8177415291851210, 91951120), 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).

## Author Contributions

S.G.G., N.K., and W.L. collected data and performed the computations and tests and contributed equally toward 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 interest between the coauthors

## Data Availability

The data underlying this article are available upon request and on <http://dx.doi.org/10.25838/d5p-14>.

## Literature Cited

- Akil C, Robinson RC. 2018. Genomes of asgard archaea encode profilins that regulate actin. *Nature* 562(7727):439–443.
- Anantharaman K, et al. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun.* 7(1):13219.
- Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E. 2018. Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery. *Genome Biol Evol.* 10(3):707–715.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.
- Breitwieser FP, Lu J, Salzberg SL. 2019. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform.* 20(4):1125–1136.
- Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. 2001. Universal trees based on large combined protein sequence data sets. *Nat Genet.* 28(3):281–285.
- Bryant D, Moulton V. 2003. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol.* 21(2):255–265.
- Crisuolo 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 Evol Biol.* 10(1):210.
- Cunha VD, Gaia M, Gadelle D, Nasir A, Forterre P. 2017. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.* 13(6):e1006810.
- Cunha V, Gaia M, Nasir A, Forterre P. 2018. Asgard archaea do not close the debate about the universal tree of life topology. *PLoS Genet.* 14(3):e1007215.
- Daubin V, Gouy M, Perrière G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.* 12(7):1080–1090.
- Dell'Anno A, Danovaro R. 2005. Extracellular DNA plays a key role in deep-sea ecosystem functioning. *Science* 309:2179–2179.
- Dell'Anno A, Stefano B, Danovaro R. 2002. Quantification, base composition, and fate of extracellular DNA in marine sediments. *Limnol Oceanogr.* 47(3):899–905.
- Dey G, Thattai M, Baum B. 2016. On the archaeal origins of eukaryotes and the challenges of inferring phenotype from genotype. *Trends Cell Biol.* 26(7):476–485.
- Eme L, Spang A, Lombard J, Stairs CW, Ettema TJG. 2017. Archaea and the origin of eukaryotes. *Nat Rev Microbiol.* 15(12):711–723.
- Fan L, et al. 2020. Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria. *Nat Ecol Evol.* 4(9):1213–1219.
- Gould SB, Garg SG, Martin WF. 2016. Bacterial vesicle secretion and the evolutionary origin of the eukaryotic endomembrane system. *Trends Microbiol.* 24(7):525–534.
- Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol.* 5(10):R245–R249.
- Hansmann S, Martin W. 2000. 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(4):1655–1663.
- Hug LA, et al. 2016. A new view of the tree of life. *Nat Microbiol.* 1(5):16048.
- Imachi H, et al. 2020. Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* 577(7791):519–525.
- Kang DD, et al. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7:e7359.
- Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4):772–780.
- Kuhner M, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol.* 11(3):459–468.
- Kuhner MK, Yamato J. 2015. Practical performance of tree comparison metrics. *Syst Biol.* 64(2):205–214.
- Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol.* 24(6):1380–1383.
- Marcy Y, et al. 2007. Dissecting biological “dark matter” with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci U S A.* 104(29):11889–11894.
- Martin W. 2003. Gene transfer from organelles to the nucleus: frequent and in big chunks. *Proc Natl Acad Sci U S A.* 100(15):8612–8614.
- Martin W, et al. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393(6681):162–165.
- Martin WF, Tielens AGM, Mentel M, Garg SG, Gould SB. 2017. The physiology of phagocytosis in the context of mitochondrial origin. *Microbiol Mol Biol R.* 81:e00008-17.

- Matte-Tailliez O, Brochier C, Forterre P, Philippe H. 2002. Archaeal phylogeny based on ribosomal proteins. *Mol Biol Evol.* 19(5):631–639.
- Nagler M, Insam H, Pietramellara G, Ascher-Jenull J. 2018. Extracellular DNA in natural environments: features, relevance and applications. *Appl Microbiol Biotechnol.* 102(15):6343–6356.
- Nelson-Sathi S, et al. 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc Natl Acad Sci U S A.* 109(50):20537–20542.
- Nelson-Sathi S, et al. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517(7532):77–80.
- Nesbø CL, Boucher Y, Doolittle FW. 2001. Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. *J Mol Evol.* 53(4–5):340–350.
- Nye T, Liò P, Gilks WR. 2006. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics* 22(1):117–119.
- O’Leary NA, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44(D1):D733–D745.
- Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25(7):1043–1055.
- Penny D, Foulds LR, Hendy MD. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature* 297(5863):197–200.
- Rinke C, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499(7459):431–437.
- Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci.* 53(1–2):131–147.
- Shen X-X, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat Ecol Evol.* 1:s41559-017–0126.
- Spang A, et al. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521(7551):173–179.
- Spang A, et al. 2018. Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLoS Genet.* 14(3):e1007080.
- Steel M, Huson D, Lockhart PJ. 2000. Invariable sites models and their use in phylogeny reconstruction. *Syst Biol.* 49(2):225–232.
- Stephen J, McCaig A, Smith Z, Prosser J, Embley T. 1996. Molecular diversity of soil and marine 16S rRNA gene sequences related to beta-subgroup ammonia-oxidizing bacteria. *Appl Environ Microbiol.* 62(11):4147–4154.
- Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* 56(4):564–577.
- The UniProt Consortium 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47:D506–D515.
- Thiergart T, Landan G, Martin WF. 2014. Concatenated alignments and the case of the disappearing tree. *BMC Evol Biol.* 14(1):266.
- Torsvik V, Sørheim R, Goksøyr J. 1996. Total bacterial diversity in soil and sediment communities—a review. *J Ind Microbiol.* 17(3–4):170–178.
- Torti A, Lever MA, Jørgensen BB. 2015. Origin, dynamics, and implications of extracellular DNA pools in marine sediments. *Mar Genomics.* 24:185–196.
- Tria FDK, Landan G, Dagan T. 2017. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol.* 1(1):0193.
- Williams W, Clifford H. 1971. On the comparison of two classifications of the same set of elements. *Taxon.* 20(4):519–522.
- Zaremba-Niedzwiedzka K, et al. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541(7637):353–358.

**Associate editor:** Rebecca Zufall