


# An Efficient, Nonphylogenetic Method for Detecting Genes Sharing Evolutionary Signals in Phylogenomic Data Sets

Luiz Thibério Rangel<sup>1,\*</sup>, Shannon M. Soucy<sup>2</sup>, João C. Setubal<sup>3</sup>, Johann Peter Gogarten <sup>4,5</sup>, and Gregory P. Fournier<sup>1</sup>

<sup>1</sup>Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

<sup>2</sup>Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, USA

<sup>3</sup>Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, Brasil

<sup>4</sup>Department of Molecular and Cell Biology, University of Connecticut, USA

<sup>5</sup>Institute for Systems Genomics, University of Connecticut, USA

\*Corresponding author: E-mail: lthiberiol@gmail.com.

Accepted: 11 August 2021

## Abstract

Assessing the compatibility between gene family phylogenies is a crucial and often computationally demanding step in many phylogenomic analyses. Here, we describe the Evolutionary Similarity Index ( $I_{ES}$ ), a means to assess shared evolution between gene families using a weighted orthogonal distance regression model applied to sequence distances. The utilization of pairwise distance matrices circumvents comparisons between gene tree topologies, which are inherently uncertain and sensitive to evolutionary model choice, phylogenetic reconstruction artifacts, and other sources of error. Furthermore,  $I_{ES}$  enables the many-to-many pairing of multiple copies between similarly evolving gene families. This is done by selecting non-overlapping pairs of copies, one from each assessed family, and yielding the least sum of squared residuals. Analyses of simulated gene family data sets show that  $I_{ES}$ 's accuracy is on par with popular tree-based methods while also less susceptible to noise introduced by sequence alignment and evolutionary model fitting. Applying  $I_{ES}$  to an empirical data set of 1,322 genes from 42 archaeal genomes identified eight major clusters of gene families with compatible evolutionary trends. The most cohesive cluster consisted of 62 genes with compatible evolutionary signal, which occur as both single-copy and multiple homologs per genome; phylogenetic analysis of concatenated alignments from this cluster produced a tree closely matching previously published species trees for Archaea. Four other clusters are mainly composed of accessory genes with limited distribution among Archaea and enriched toward specific metabolic functions. Pairwise evolutionary distances obtained from these accessory gene clusters suggest patterns of interphyla horizontal gene transfer. An  $I_{ES}$  implementation is available at <https://github.com/lthiberiol/evolSimIndex>.

**Key words:** bioinformatics, phylogenomics, gene coevolution, clustering, Archaea.

## Significance

Detecting shared evolutionary trends of gene families is necessary for distinguishing genes with incompatible evolutionary signals from those suitable to reconstruct reliable phylogenomic trees. Commonly used methods to achieve this directly compare the topologies of gene trees, which tend to be inaccurate given the inherent uncertainty of phylogenetic reconstruction. We propose the Evolutionary Similarity Index ( $I_{ES}$ ), based on orthogonal distance regressions between evolutionary distance matrices. Simulations show that  $I_{ES}$  substantially outperforms tree-based methods, at a fraction of the computational effort, and is able to evaluate similarities between gene families containing duplication events. We used  $I_{ES}$  to assess the compatibility between archaeal gene families, producing a cohesive cluster of 62 genes with similar evolutionary signals, likely representing a central evolutionary trend of archaeal genomes.

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

## Introduction

Phylogenies reconstructed from single genes are known to poorly reflect the underlying history of whole genomes; consequently the detectable phylogenetic signal from an isolated locus cannot be extrapolated to reflect the evolution of whole genomes (Dagan and Martin 2006; Baptiste et al. 2009; Koonin et al. 2009). To ameliorate this effect, it has become common practice to estimate species' evolutionary histories by concatenating multiple sequence alignments of single-copy genes widely conserved across sampled genomes. The preference toward using core genome sequences is due to their expected resistance to horizontal gene transfer (HGT) (Thomas and Nielsen 2005; Sorek et al. 2007; Popa and Dagan 2011); however, despite the lower frequency of HGT among some gene families, it has been shown that horizontal exchange also occasionally affects core genes. In fact, the slow substitution rate and corresponding high sequence conservation of the core genome may even favor HGT, permitting increases in neutral and nearly neutral HGT at the genus and species levels (Papke and Gogarten 2012; Shapiro et al. 2012).

Given this context, it is clear that more rigorous methods are needed to identify genes best reflecting the underlying vertical evolutionary signal in a group of species; such methods should seek to maximize the compatibility between evolutionary signals in order to provide a more robust basis for phylogenomic reconstruction. Many strategies have been proposed to assess similarities between phylogenetic signals obtained by individual gene trees—for example, Robinson–Foulds bipartition compatibility (RF) (Robinson and Foulds 1981), geodesic distance ( $D_{\text{geo}}$ ) (Kimmel and Sethian 1998; Billera et al. 2001; Kupczok et al. 2008; Owen and Provan 2011), matching split distance ( $D_{\text{ms}}$ ) (Lin et al. 2012; Bogdanowicz and Giaro 2012), and quartet distance ( $D_{\text{qt}}$ ) (Estabrook et al. 1985)—as well as other methods that assess similarities between phylogenetic profiles (Pellegrini et al. 1999; Vert 2002; Barker and Pagel 2005; Liu et al. 2018). The majority of tree-based methods rely on straightforward comparisons between tree topologies (Robinson and Foulds 1981; Kunin et al. 2005; Leigh et al. 2008; Puigbò et al. 2009; Lin et al. 2012; Bogdanowicz and Giaro 2012; Mirarab et al. 2014; Gori et al. 2016). However, although an intuitive solution, comparisons between tree topologies require phylogenetic trees of all assessed gene families to be accurately reconstructed, adding a substantial computational cost to an already computationally demanding task. Furthermore, the vastness of tree space, combined with the inherent uncertainty of phylogenetic reconstruction, provides multiple sources of error to tree-based evolutionary similarity assessments. Another method to assess the evolutionary compatibility of genes is based on similarities between patterns of presence and absence (phylogenetic profiles) of such genes among genomes of interest. Although more recent implementations displayed substantial improvements (Liu et al. 2018), reliance on an initial reference tree constitutes an obstacle to

its application to new taxon samplings. Phylogenetic profile-based methods also do not assess divergences between sequences of homologous genes, which limits the resolution of their results.

Accounting for uncertainty-based variations in tree topology (i.e., bipartition support) further increases the computational burden and decreases the resolution of the evaluated phylogenetic signal (e.g., collapsing low support bipartitions or weighing them based on support). A proposed solution to bypass the computational cost of tree similarity assessments is Pearson's correlation coefficient ( $r$ ) between evolutionary distance matrices (Goh et al. 2000; Pazos and Valencia 2001; Novichkov et al. 2004; Rangel et al. 2019). Unlike tree-based comparisons, methods based on Pearson's  $r$  enable simple implementations to detect similar evolutionary signals between gene families with histories complicated by multiple homologs within genomes. This is accomplished by estimating multiple correlation coefficients, each using distinct combinations of paralogs between gene families (Gertz et al. 2003; Ramani and Marcotte 2003). Despite its application in protein–protein interaction studies, the sensitivity of Pearson's  $r$  to noise in evolutionary distances and the granularity of its estimates have yet to be compared with those of tree-based metrics. Direct coupling analysis has also been used to pair gene copies between possibly coevolving gene families (Gueudré et al. 2016), but despite positive results the assumption that protein products of coevolving genes must be directly interacting may limit its applications.

Given the limitations of the aforementioned approaches and methods, we propose the  $I_{\text{ES}}$  to quantify the similarity between evolutionary histories based on weighted orthogonal distance regression (ODR) (wODR) between pairwise distance matrices. We show that  $I_{\text{ES}}$  is robust to dissimilarity saturation resulted by up to 50 simulated perturbations to underlying phylogenies. More common tree-based evolutionary similarity estimates must be corrected for distance saturation to display similar robustness and are significantly more susceptible to errors in evolutionary history reconstruction. As a case study for this new method, we assessed evolutionary similarities across 1,322 archaeal gene families and detected significant evolutionary incompatibilities between conserved single-copy genes, as well as a clear central evolutionary tendency involving 62 gene families that occur as both single and multiple copies across genomes.

## Results and Discussion

### Simulated Data Set

Similarities between evolutionary histories of simulated gene families were assessed using our newly proposed  $I_{\text{ES}}$  and four tree-based metrics:  $D_{\text{geo}}$ , RF,  $D_{\text{ms}}$ , and  $D_{\text{qt}}$ . Results from all five approaches successfully identified the monotonic decrease in similarity between simulated gene families as the number of perturbations increased (supplementary figs. S2 and S3,

Supplementary Material online); tree-based estimates, however, were heavily impacted by noise introduced through empirical sequence alignment methods and suboptimal evolutionary model (fig. 1 and supplementary figs. S2–S4, Supplementary Material online). The accuracy of each method in identifying the degree of similarity between simulated gene families was evaluated by calculating ordinary least squares (OLS)  $R^2$  between pairwise shared evolution estimates ( $1 - I_{ES}$  or  $d_{adj}$  for tree-based metrics) and number of simulated perturbations. OLS models were fitted using a Y-axis intercept of 0 as two gene families with identical histories should yield no differences according to any method (supplementary figs. S2 and S3, Supplementary Material online).

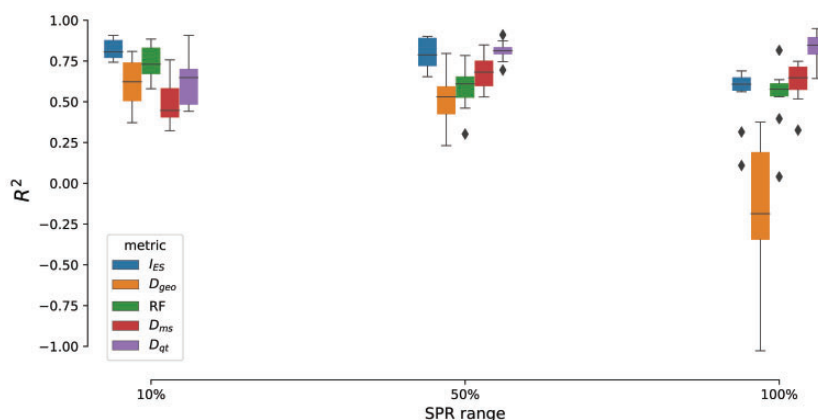
Under perfect conditions (i.e., no sequence alignment errors and optimal evolutionary model),  $I_{ES}$  performed on par with tree-based adjusted estimates (fig. 1). Among gene families simulated with weak phylogenetic perturbations,  $I_{ES}$  significantly outperformed tree-based methods as measured by OLS  $R^2$ . Adjusted  $D_{qt}$  performed similarly to  $I_{ES}$  among gene families with medium perturbations. In simulations generated using strong phylogenetic perturbations, adjusted  $D_{qt}$  significantly outperformed  $I_{ES}$  as a predictor of differences between simulated gene families; for simulations with strong phylogenetic perturbations no significant difference was detected between  $I_{ES}$ , RF, and  $D_{ms}$ . Among simulated genes reconstructed with perfect information,  $I_{ES}$  displayed greater accuracy when compared with established tree-based methods even though three out of four tree-based methods ignore branch lengths (i.e., RF,  $D_{ms}$ , and  $D_{qt}$ ), and consequently are not affected by perturbations applied to branch lengths. Among tree-based methods,  $D_{geo}$  was shown to be the most negatively impacted by strong phylogenetic perturbations between simulated gene trees. Although  $D_{geo}$  distributions obtained from SPR10 and SPR50 data sets are not significantly different from other tree-based methods,  $D_{geo}$  showed an abrupt decrease in

performance among SPR100 gene families. This may be due to a high sensitivity to long SPR moves or suggest that the applied saturation correction is not well suited for its full range of distance estimates.

Phylogenetic reconstruction using a simulated sequence alignment (*tree\_1\_TRUE.phy*, from SPR10 and replicate\_1) under LG+G model by IQTree 1.6.7 took 393.6 s in a single thread; computing the pairwise distance matrix for the same alignment took 6.487 s. Both computations were performed on a 3 GHz Intel Xeon W processor. Although  $D_{qt}$  may provide marginal gains in extreme scenarios in comparison to  $I_{ES}$ , it comes with an added computing time of almost 150 $\times$ , without bipartition support assessment. When assessing large data sets, the quadratic nature of distance matrices will decrease the reported disparity between computing times of  $I_{ES}$  and tree-based methods, even though the latter still rely in bipartition support values for enhanced accuracy. Both the time and computing resources required for reasonable phylogenetic reconstruction constitute prohibitive factors toward the assessment of shared evolution in large multi-genome data sets.

#### Robustness Assessment between Approaches

The dichotomic pattern in a cladogram is extremely susceptible to uncertainties in phylogenetic reconstruction. Combined with the vast tree space available for 50 taxa, this can cause noise-induced topological variations to be not directly distinguishable from real deviations in evolutionary history (Szöllösi et al. 2013). Pairwise maximum likelihood distance matrices used to estimate  $I_{ES}$  are less prone to such uncertainty as they bypass forming hypotheses about the evolutionary relationships between taxa. This assumption is corroborated by comparing evolutionary similarity estimates obtained using error-free sequence alignments and optimal evolutionary models to those using realigned sequences and suboptimal evolutionary



**Fig. 1.**—Boxplots of OLS  $R^2$  for shared evolution estimates from perfectly aligned simulated gene families for each data set of simulated phylogenetic perturbations. Each data set was replicated ten times, and scatterplot and fitted OLS regressions are available in supplementary figure S2, Supplementary Material online. Negative  $R^2$  values occur as fitted linear regressions do not explain the association between variables, and in this scenario reflect strong saturation of evolutionary similarity measurements.

models. Although alignment errors (average SP-score of 0.87) and evolutionary model suboptimality (JTT) were kept to the minimum one could expect in empirical data sets, the resulting noise is sufficient to negatively impact RF,  $D_{ms}$ , and  $D_{qt}$  estimates. In all three simulated data sets (SPR10, SPR50, and SPR100) tree-based methods, except  $D_{geo}$ , displayed significantly smaller OLS  $R^2$  when compared with the number of phylogenetic perturbations between simulated gene families (supplementary fig. S5, Supplementary Material online). On the other hand,  $I_{ES}$  estimates were shown to be robust toward error-induced noise in both SPR50 and SPR100 data sets (supplementary fig. S5, Supplementary Material online).

### Evolutionary Similarities within Archaeal Gene Families

In order to test  $I_{ES}$  performance when estimating shared phylogenetic signal in an empirical set of gene families, we evaluated 1,322 families of homologous proteins assembled from annotated coding sequences (CDS) extracted from 42 archaeal genomes (supplementary table S1, Supplementary Material online). This empirical data set contains conserved and accessory gene families with different sizes due to gene losses, duplications, and transfers.

$I_{ES}$  was estimated for all pairwise combinations of gene families present in at least ten genomes, with 2,142 out of 748,712 comparisons having  $I_{ES}$  values of at least 0.7 (supplementary fig. S6a, Supplementary Material online). Pairs of gene families with an  $I_{ES} \geq 0.7$  were added as nodes to a network with pairwise  $I_{ES}$  as edge weights connecting gene families. In total 419 unique archaeal gene families were added to the network, whereas the remaining 903 gene families did not display any  $I_{ES} \geq 0.7$  with other gene families. The 0.7  $I_{ES}$  threshold was selected as it is the most robust to threshold increases, as measured by Variation of Information (Meilă 2007), while also yielding the greater cluster modularity than networks obtained using lower thresholds (supplementary table S2, Supplementary Material online). Although increasing the  $I_{ES}$  threshold from 0.7 to 0.75 leads to a 26% increase in cluster modularity, it does so by removing 71.15% of edges (supplementary table S2, Supplementary Material online). Previous applications based on Pearson's  $r$  between evolutionary distances have also suggested a 0.7 threshold (Goh et al. 2000; Pazos and Valencia 2001). The resulting evolutionary similarity network (fig. 2) is heavily imbalanced, with just 11% of nodes involved in 50% of network edges. The majority of gene families (68%) did not display any  $I_{ES}$  above the 0.7 threshold with other gene families, suggesting a general incompatibility between evolutionary signals, or an inability to detect compatibility with this method. However, the high edge concentration within just a few nodes suggests a strong central signal present among few gene families, from which the evolutionary trajectories of others have diverged (Puigbò et al. 2009).

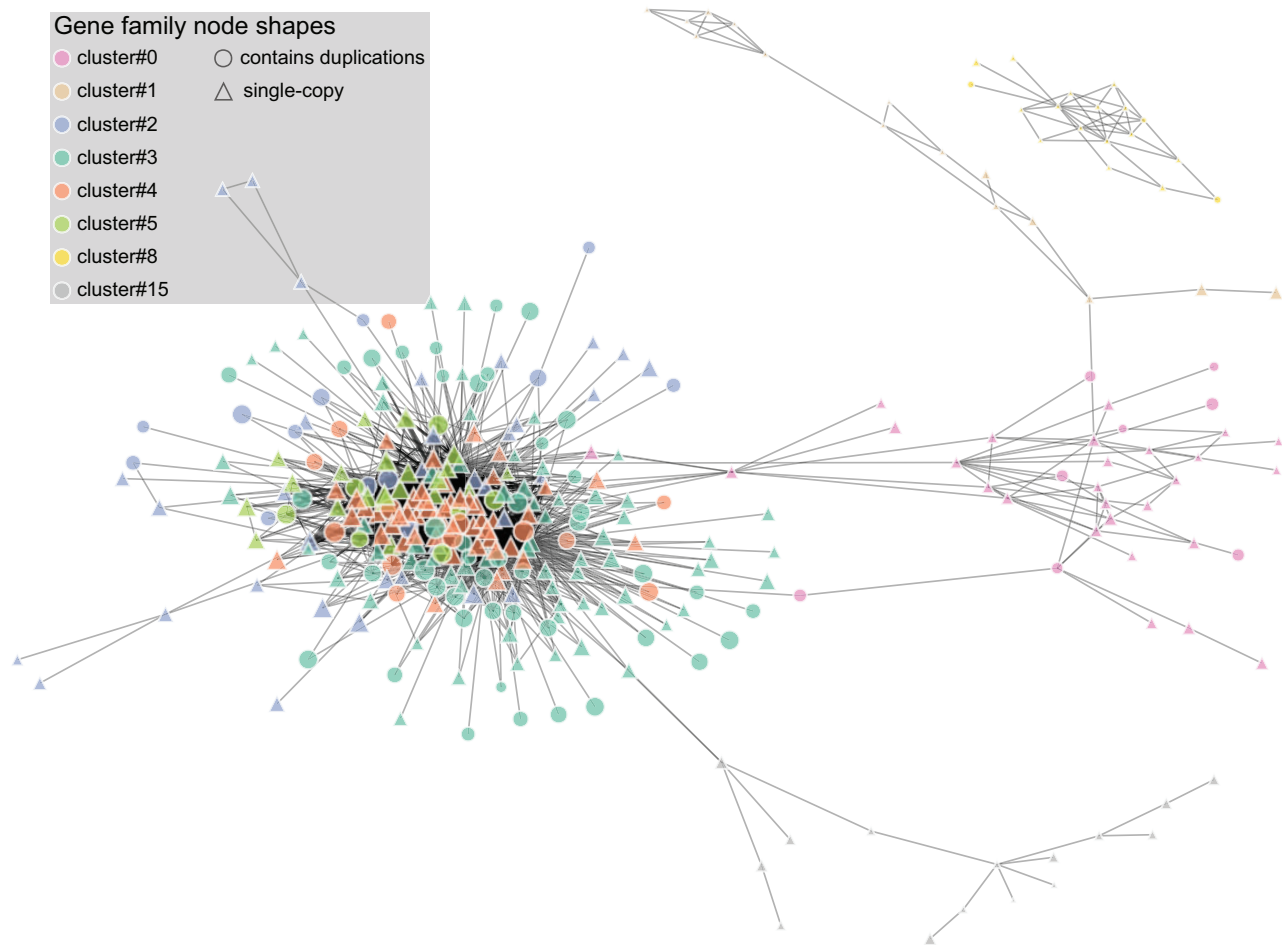
A hierarchical clustering of pairwise  $I_{ES}$  based on average nucleotide distances between gene families within individual

genomes suggests that shared evolution estimates are strongly impacted by genetic linkage (supplementary fig. S6b, Supplementary Material online). Pairs of gene families within close genomic proximity (i.e., apart from each other by fewer than 10,000 bp in at least 21 genomes) have significantly more similar evolutionary trends than pairs separated by long genomic distances (i.e., apart from each other by more than 100,000 bp in at least 21 genomes), as depicted in figure 3a ( $P = 6.18e^{-75}$  and  $f = 0.86$ ). Pairs of gene families with strong genetic linkage displayed significantly greater  $I_{ES}$  relative to pairs of gene families with weak linkage.

### Clusters of Gene Families with Compatible Evolutionary Signals

Evolutionary trends shared across gene families were grouped using Louvain community detection, recovering 41 compatible evolutionary signal (CES) clusters, of which 25 comprise only two gene families and eight major clusters contain ten or more similarly evolving gene families. As evidence of the effectiveness of clustering gene families using pairwise  $I_{ES}$ , we observe that CES clusters are strongly associated with genetic linkage within short genetic distances (fig. 3b). Across short nucleotide distances between loci, linkage is a strong predictor of CES relations between genes, but its predictive power rapidly decreases as the genomic distance between two given loci increases (fig. 3b), displaying a linear log-log relationship (supplementary fig. S7, Supplementary Material online). Comparing frequencies of intra- and intercluster gene pairs across distinct windows of genomic distances showed that the proportion of CES genes within 1,000 bp of each other is three times the proportion of non-CES genes within the same window. Increasing the window size led to abrupt decreases in proportion; within a 10,000-bp window, the ratio of CES genes is reduced to 1.8 the ratio of non-CES, and at a 100,000-bp window this difference in proportion falls to 1.2 (fig. 3b). Given the strong genetic linkage and functional associations of operons, these results suggest that the evolutionary signal shared by gene pairs in known operons might be even stronger than that shown by figure 3.

Among the eight CES clusters with ten or more gene families, four are comprised mostly core genes, and four are composed of mostly accessory genes (fig. 4). The four CES clusters of core genes (cluster#2, cluster#3, cluster#4, and cluster#5) are promising candidates for reconstructing a representative phylogenetic signal present within sampled Archaea. These four core CES clusters are composed of 102 extended core genes (single copy and present in at least 35 genomes) and 146 broadly distributed gene families present on average in 33 genomes, both as single and multiple copies. CES clusters of accessory genes (cluster#0, cluster#1, cluster#8, and cluster#15 in fig. 5) include specific archaeal clades, but do not map to well-established phylogenetic relationships; rather, they show polyphyletic gene distributions, likely caused by HGTs and/or



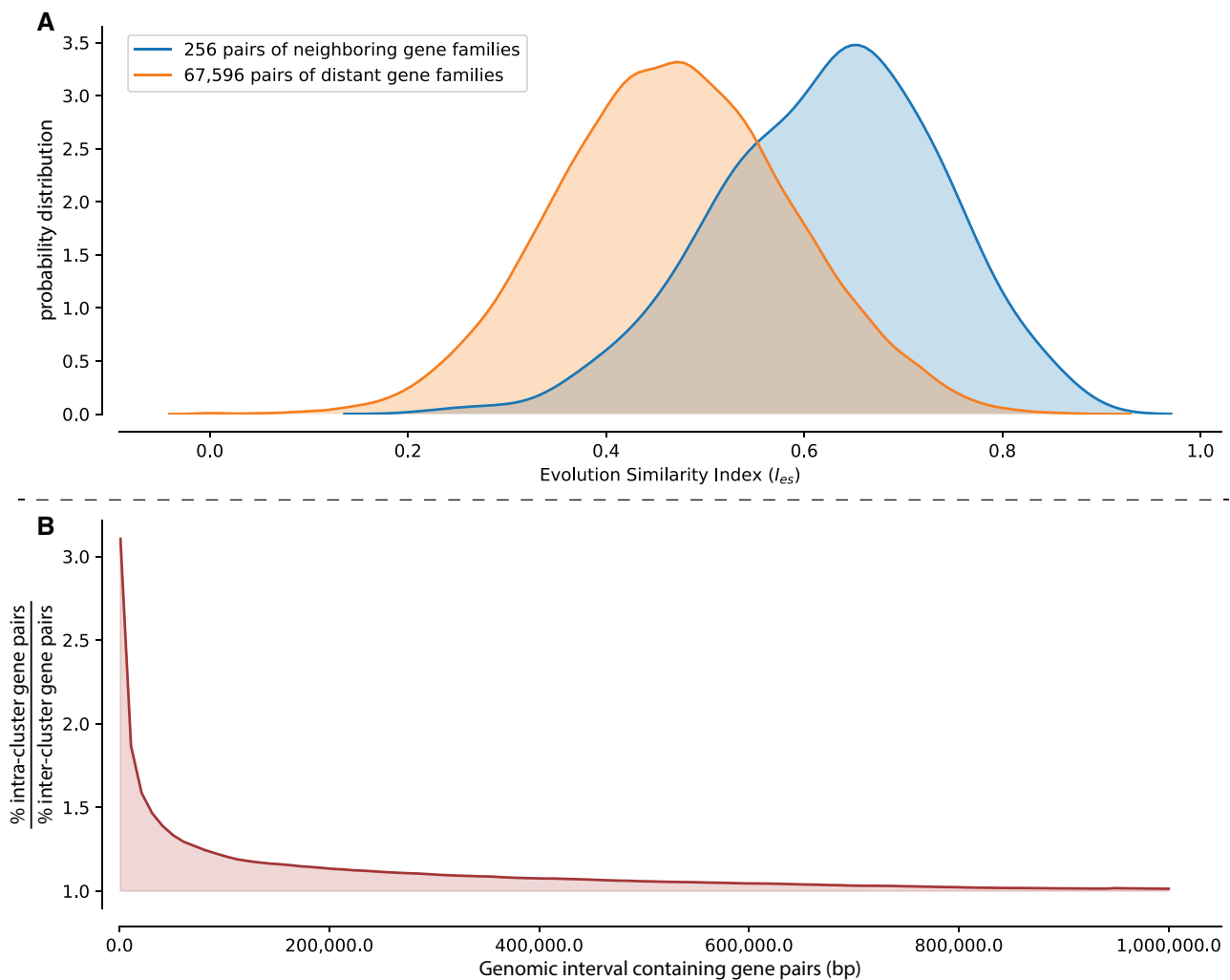
**FIG. 2.**—A compatible evolutionary signal network, with each node representing a gene family, and edges connecting nodes representing shared evolutionary signals ( $I_{ES} \geq 0.7$ ). Nodes of the same colors have similar evolutionary trends, as identified by Louvain community detection. Triangular nodes represent single-copy genes, and circular nodes are gene families containing duplications. Clusters of CES gene families with less than ten members are not represented.

gene losses shared by CES gene families. For example, cluster#0 is well represented amongst Euryarchaeota and hyperthermophilic TACK; cluster#15 comprises gene families with shared evolutionary trends mainly occurring within Crenarchaeota and hyperthermophilic Euryarchaeota; CES accessory gene families in cluster#1 and cluster#8 display congruent signals grouping methanogenic Euryarchaeota with Thaumarchaeota and Asgardarchaeota, respectively. Besides the eight CES clusters with ten or more gene families, the CES network community detection yielded 34 other clusters containing between two and nine gene families (see [Supplementary Material](#) online). These 34 CES clusters contain a total of 88 gene families, with degree centralities much smaller than the 331 gene families within the eight major CES clusters (averages of 1.36 and 17.79, respectively).

CDSs from 21 out of 42 sampled genomes have functional annotations available in StringDB (see [Supplementary Material](#) online), which was used to identify annotated KEGG Pathways enriched within homologs of CES gene families from each

genome. In the dendrogram and heatmap depicted in figure 4, we clearly identify two sets of opposing CES clusters of gene families: accessories (top three rows) and core (bottom four rows). All four CES clusters of core gene families are enriched with KEGG Pathways related to genetic information processing (e.g., Ribosome, DNA replication, and Aminoacyl-tRNA biosynthesis), whereas CES clusters of accessory gene families are enriched with KEGG Pathways related to metabolism (e.g., Methane metabolism, Microbial metabolism in diverse environments, and Biosynthesis of antibiotics in fig. 4). KEGG Pathways related to metabolism display minor enrichment signal within CES clusters of core gene families, and KEGG Pathways related to genetic information processing are not enriched within clusters of accessory genes (fig. 4). Cluster#1, restricted to methanogenic Euryarchaeota and Thaumarchaeota, is enriched for methane metabolism genes within six genomes. Similarly, gene families from cluster#8, restricted to methanogenic Euryarchaeota and Asgardarchaeota, are also enriched in methane metabolism in five genomes (fig. 4).



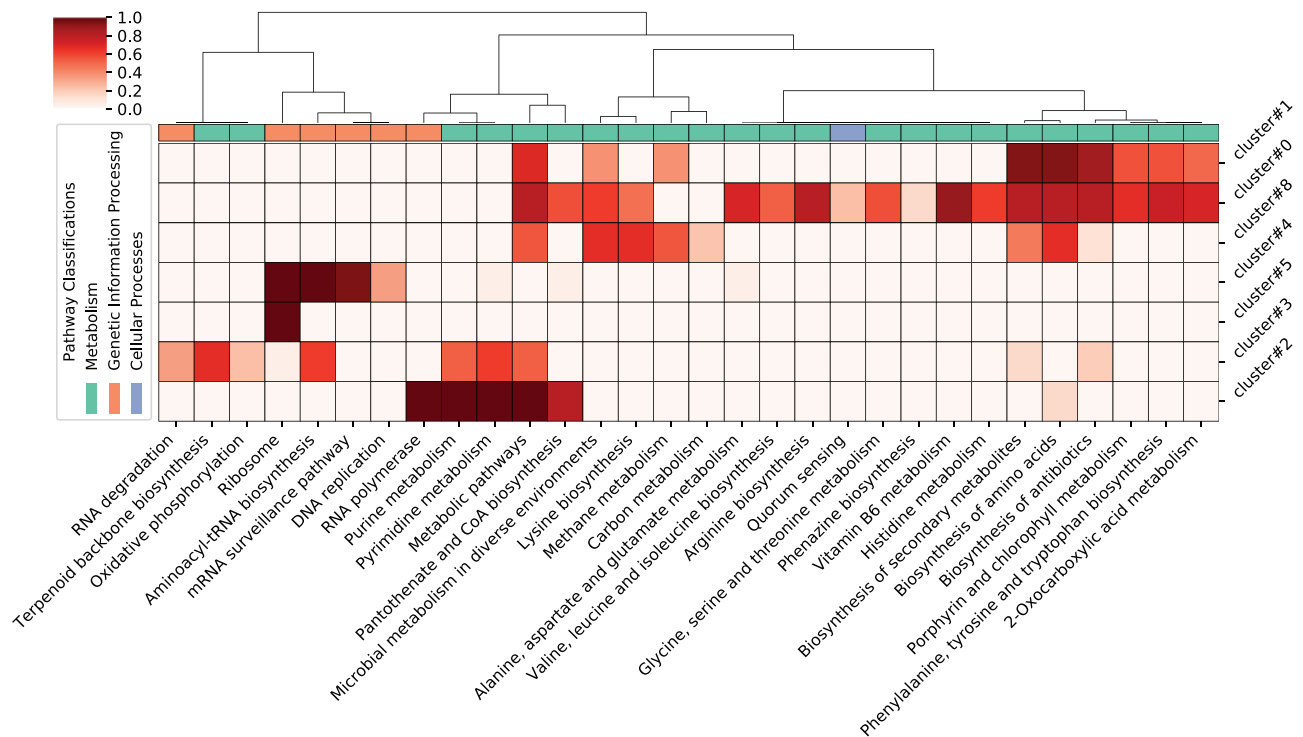


**FIG. 3.**—(A) Distributions of  $I_{ES}$  between pairs of genes within 10,000 bp of each other (blue) and between pairs of genes apart by at least 100,000 bp (orange). Neighboring gene pairs displayed significantly more similar evolutionary signals than nonneighboring gene pairs. (B) Ratio between the proportion of gene pairs in intra- and inter-CES clusters, Y axis, occurring within genomic windows, X axis. About 100 window sizes were assessed ranging from 1,000 to 1,000,000 bp.

### Compatible Evolutionary Signal Clusters and Possible Vertical Evolutionary Signals

Phylogenies generated from extended core genomes are generally used as reasonable proxies of the species-tree phylogeny, given the assumption that these genes are less likely to undergo HGT between distantly related groups. However, an extended core phylogeny may not represent the species tree for several reasons, including systematic biases in phylogenetic reconstruction due to shared compositional bias, or strong biases in HGT partners among sets of genes. The extended core tree can still be used as an adequate representation of the consensus evolutionary signal detected in the sampled archaeal genomes, the closest thing we have to the simple “null hypothesis” of a shared history due to vertical inheritance. The 102 genes composing the extended core genome are not equally distributed across CES clusters (fig.

2); cluster#4 contains the greatest number of extended core genes, 44 out of 62 gene families, followed by cluster#3 with 27 out of 111 gene families. The split of the extended core genome into four distinct major CES clusters (fig. 2) suggests differing sets of HGTs among core genes, creating conflicting evolutionary histories between genes from different clusters (fig. 6). Closeness centrality measures ( $\tilde{C} = 0.56$ ) and node strength corrected by cluster size ( $\tilde{S} = 0.19$ ) suggest that cluster#4 gene families share stronger and more cohesive evolutionary trends than gene families from other clusters (supplementary fig. S8, Supplementary Material online). Therefore, cluster#4 contains the set of genes that may be best able to recover a representative evolutionary history across sampled Archaea. Cluster#4’s evolutionary history is also the most similar to that inferred from the extended core genome (figs. 5 and 6). The phylogeny obtained from



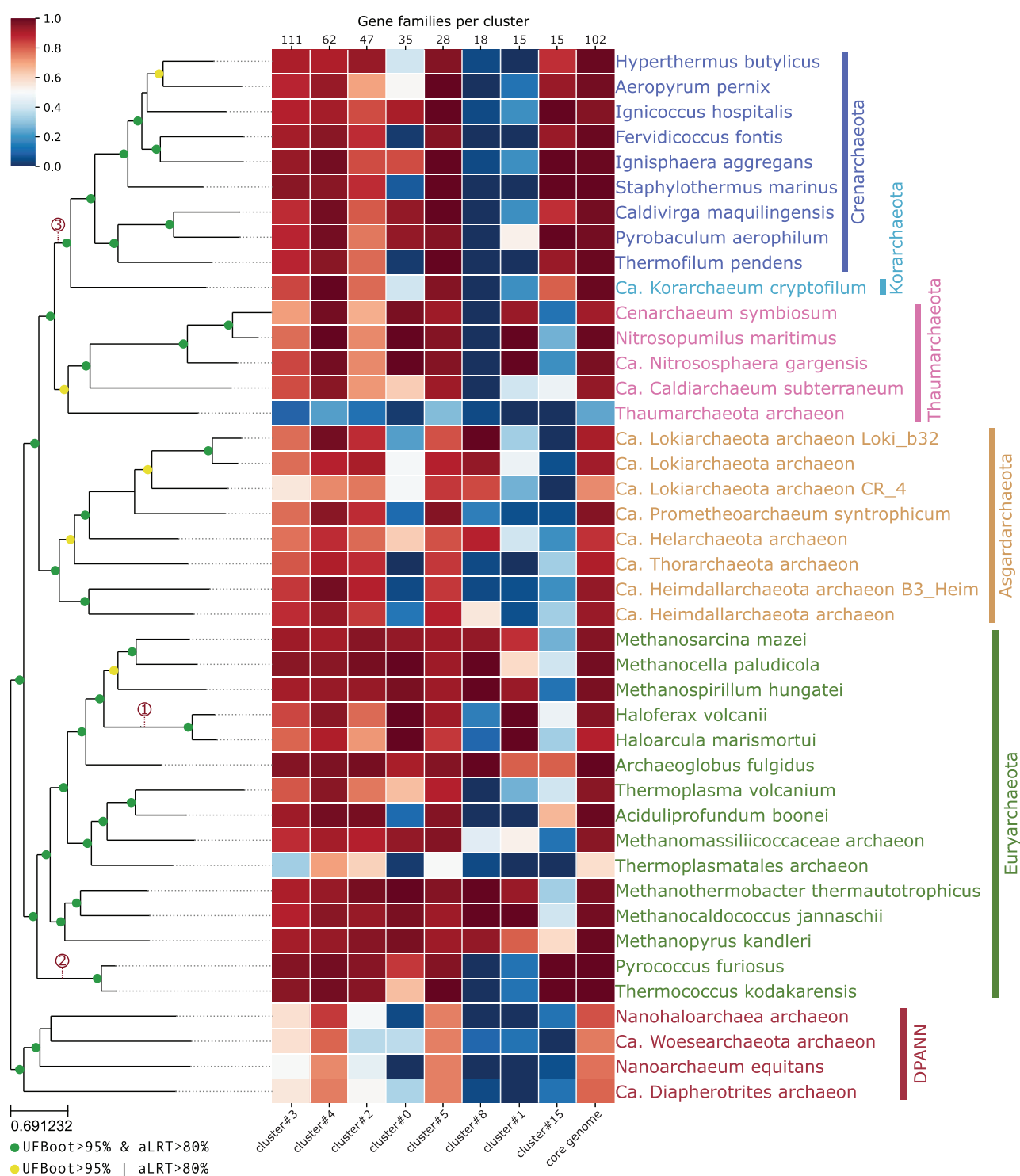
**FIG. 4.**—Heatmap of enriched KEGG Pathways within each CES cluster of gene families. Shades of red represent the proportion of genomes with detected KEGG Pathway enrichment within its homologs of CES gene families. Columns and rows were clustered using complete linkage and correlation coefficients. KEGG Pathways enriched in fewer than 10% of genomes in which CES genes occur are not reported. Cluster#15 did not show significant enrichment of KEGG Pathways.

concatenated cluster#4 genes is highly similar to recently published reconstructions of the archaeal Tree of Life, with virtually identical Euryarchaeota clade structure (Williams et al. 2017, 2020).

Although binning genes with congruent evolutionary histories permits phylogenetic reconstructions less likely to be subject to spurious signals arising from the averaging of conflicting evolutionary signals, the resulting phylogenies remain susceptible to phylogenetic reconstruction artifacts. For example, since  $I_{ES}$  is not estimated from phylogenies but from pairwise evolutionary distances, we do not expect it to be subject to long-branch attraction (LBA) artifacts. Nevertheless, phylogenies reconstructed from sets of genes with high  $I_{ES}$  between each other are as susceptible to LBA as any other data set. Despite robustness to phylogenetic artifacts,  $I_{ES}$  estimates are still affected by sampling biases: the overrepresentation of specific taxonomic groups can lead to underestimating deviations in the evolutionary history of less represented groups.

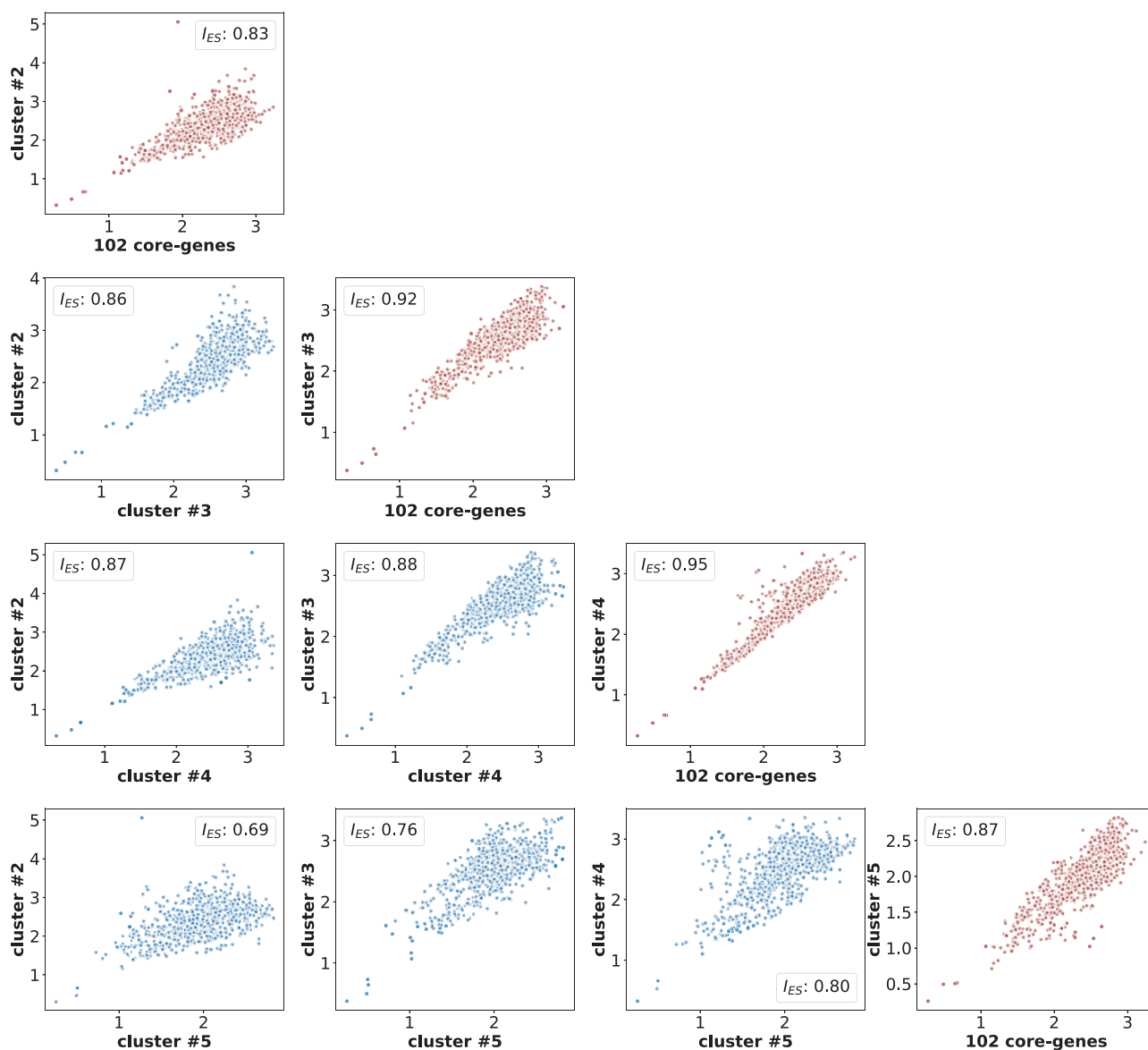
LBA is a frequently invoked in discussions of archaeal phylogeny, specifically with regard to the phyletic status of the DPANN group (Brochier-Armanet et al. 2011; Petitjean et al. 2014; Raymann et al. 2014; Williams et al. 2015; Feng et al. 2021). Regardless of the set of genes used for phylogenetic reconstruction, extended core genome or any of the four CES

clusters of core genes, all resulting trees depicted a well-supported DPANN clade composed of Nanohaloarchaea archaeon, *Ca. Woesearchaeota* archaeon, *Nanoarchaeum equitans*, and *Ca. Diapherotrites* archaeon. To assess the impact of LBA in reconstructing DPANN, we generated phylogenies from the CES clusters of core genes including only a single DPANN taxon at a time. When included individually, each DPANN taxon showed varying placements across trees generated from different CES clusters. This suggests that the initial monophyletic grouping of DPANN for these clusters was, in fact, artificial and that the extended core gene history for these genomes is likely complex (see [Supplementary Material](#) online). Cluster#4 phylogenies individually testing the position of each DPANN taxa placed Nanohaloarchaea sister to Halobacteria, with Nanohaloarchaea+Halobacteria being sister to Methanomicrobia. Cluster#3 also reported Nanohaloarchaea sister to Halobacteria, but with both nested within Methanomicrobia. This placement for Nanohaloarchaea has been previously proposed by Narasingarao et al. (2012), Zhaxybayeva et al. (2013), Petitjean et al. (2014), and Feng et al. (2021). The uncertain placement of Nanoarchaea has also been topic of investigation (Huber et al. 2002; Brochier et al. 2005). Interestingly, each CES cluster recovered a different placement for Nanoarchaea: sister to Euryarchaeota (cluster#2), sister to Korarchaeota+Crenarchaeota (cluster#3),



**Fig. 5.**—Phylogenetic tree of Archaea reconstructed from 62 genes within CES cluster#4. The phylogeny was obtained using LG+F+G+C60 evolutionary model from IQTree and each gene had its parameters independently estimated according to parameter “-sp.” Bipartition supports were estimated using UFBoot and aLRT, each with 1,000 replicates, and bipartitions well supported by both methods are colored in green (UFBoot  $\geq$  95% and aLRT  $\geq$  80%), whereas bipartitions well supported by a single method are colored in yellow. Red dotted lines indicate Nanohaloarchaea (1) and Nanoarchaea (2 and 3) placements reconstructed within phylogenies containing a single DPANN genome at a time. Despite the lack of outgroups to Archaea within our sample the tree is rooted in DPANN for the sake of visualization. The associated heatmap reflects the representation of gene families within CES clusters amongst archaeal genomes.





**FIG. 6.**—Scatterplots of pairwise evolutionary distances between gene families. Pairwise distances between CES clusters are shown in blue, whereas pairwise distances between CES clusters and the 102 core gene data set are shown in red. Similarities between evolutionary histories were estimated by  $I_{ES}$ .

sister to Korarchaeota (cluster#4), and sister to Thermococcales (cluster#5). One of the most accepted Nanoarchaea placements is as sister to Thermococcales (Brochier et al. 2005; Urbonavičius et al. 2008; Dutilh et al. 2014), which in our analyses was recovered only by cluster#5 (see [Supplementary Material](#) online).

Although our tests further support that the monophyly of DPANN is likely due to LBA, we did not detect a significant LBA effect for Woesearchaeota and Diapherotrites. Except for Diapherotrites placing between Class I and II methanogens in cluster#2, phylogenies from all four clusters proposed both taxa grouping together as sister to Euryarchaeota, assuming an archaeal root between TACK+Asgard and Euryarchaeota (see [Supplementary Material](#) online). The disparate placements

of DPANN members within trees from CES clusters also suggests that, in addition to LBA, the DPANN clade from the extended core genome phylogeny is further impacted by the heterogeneity of the phylogenetic signal. This may not only produce a “signal averaging” effect favoring a monophyletic DPANN deeper in the archaeal tree, but may also be a contributing factor to the LBA artifact itself. Heterogeneity among combined phylogenetic signals is likely to increase the estimated branch length, as the incorrect assumption of a single underlying phylogeny will lead to more homoplastic sites.

Assuming that a given set of genomes constitutes a monophyletic clade, it is also reasonable to expect a certain number of gene families to be over represented within the clade and not readily available to genomes outside the clade. Regardless

of the driving force behind the enrichment of gene families within a clade, which may be inheritance from a common ancestor or biased HGT (Andam et al. 2010; Andam and Gogarten 2011), we identified 80 gene families enriched within TACK genomes and 111 within Euryarchaeota ( $q \leq 0.05$ ). In contrast to TACK and Euryarchaeota clades, we did not detect any gene families enriched within the four sampled DPANN genomes, providing phylogenetically independent evidence against its monophyly. Complementarily, and in support of a Nanohaloarchaea+Halobacteria clade, we identified 78 genes present in Nanohaloarchaea archaeon enriched within the three Nanohaloarchaea and Halobacteria genomes. In another contrast, Thaumarchaeota, a well-accepted clade with similar number of sampled genomes in our data set, was found to have 456 gene families enriched within its five genomes. Although the differing degrees of physiological, metabolic, and genetic diversity within these groups certainly influence the number of shared gene families, it remains striking that this particular signal of shared ancestry is conspicuously lacking in DPANN.

#### Common and Distinct Evolutionary Trends between CES Clusters

Among CES clusters of core gene families, cluster#4 and cluster#5 are most evenly represented across archaeal groups, whereas cluster#2 and cluster#3 are poorly distributed among DPANN (fig. 5). All four CES clusters of core gene families have low frequency within Thaumarchaeota archaeon SCGC AB-539-E09, and only cluster#2 and cluster#4 are present in any abundance in Thermoplasmatales archaeon SCGC AB-539-N05, 29 and 44 gene families respectively. All four clusters display very similar overall phylogenies calculated from concatenations of genes within each cluster, varying mainly within the organization of Euryarchaeota (fig. 5 and [Supplementary Material](#) online). All four CES clusters of core genes reconstructed the monophyly of Euryarchaeota, with the exception of cluster#2, which placed *Pyrococcus furiosus*, *Thermococcus kodakarensis*, *Methanocaldococcus jannaschii*, *Methanothermobacter thermautotrophicus*, and *Methanopyruus kandleri* together as sister to Asgardarchaeota+TACK. Only cluster#4 recovered the monophyly of Methanomicrobia as sister to Halobacteria, as supported by Bayesian reconstructions reported in (Martijn et al. 2020; Williams et al. 2020). The other three CES clusters place Halobacteria within Methanomicrobia, as reported by (Williams et al. 2017) and in the RNA polymerase phylogeny reported in (Da Cunha et al. 2017).

All four core CES clusters robustly identified Asgardarchaeota as sister to TACK (fig. 5), with small variation in the Asgardarchaeota phylogeny, and cluster#5 placed Korarchaeota at the base of the TACK superphylum (see [Supplementary Material](#) online) as previously reported in the literature (Williams et al. 2017, 2020); the remaining three

CES clusters of core genes place Korarchaeota sister to Crenarchaeota (fig. 5 and [Supplementary Material](#) online). When assessing all-versus-all  $I_{ES}$  between CES clusters of core genes, the evolutionary signal detected from cluster#4 is the least dissimilar to the other three (fig. 6). This shortest path from cluster#4's evolutionary trajectory to others suggests that cluster#4 best approximates the average archaeal evolutionary history (fig. 6). In general, the overall high  $I_{ES}$  estimates between core CES clusters suggest that despite composing distinct clusters, gene histories between clusters are generally congruent, with deviations reflecting small divergences potentially representing genes with specific sets of reticulate histories.

Phylogenetic trees obtained from accessory gene families in cluster#0, cluster#1, cluster#8, and cluster#15 reconstructed all represented archaeal phyla as monophyletic (except for *P. furiosus* in Euryarchaeota in cluster#0, [supplementary fig. S9, Supplementary Material](#) online), suggesting a shared common origin of accessory genes from each CES cluster by all genomes from the same phylum. Although the monophyly of archaeal phyla within trees of CES clusters of accessory genes does not permit an accurate prediction of the directionality of possible interphyla HGTs, intraphylum distances congruent to the supposed vertical inheritance signal can be used to evaluate interphylum distances under a wODR model ([supplementary figs. S10, S12, S14, and S15, Supplementary Material](#) online). When compared with pairwise distances expected from vertical inheritance, interphylum distances that are significantly shorter than estimates obtained from intraphylum distances may be attributed to HGT acquisition by one of the phyla in question. For each CES cluster of accessory genes, we assessed wODR of its pairwise distances against the inferred vertical evolution signal estimated from cluster#4.

When comparing pairwise distances obtained from cluster#1 against cluster#4, distances between Euryarchaeota and Thaumarchaeota are consistently placed below the estimated regression line ([supplementary figs. S10 and S11, Supplementary Material](#) online). This suggests that cluster#1 genes were horizontally transferred between ancestors of both phyla, causing shorter evolutionary distances between phyla than expected if their homologs diverged exclusively by vertical inheritance.

Interphyla distances between Euryarchaeota and Crenarchaeota obtained from cluster#0 fit the evolutionary rate expected using intraphylum distances for this CES cluster ([supplementary fig. S12, Supplementary Material](#) online), suggesting that homologs from both phyla were vertically inherited from a common ancestor. Different behavior was seen for cluster#0 interphyla distances involving Thaumarchaeota (Crenarchaeota to Thaumarchaeota and Euryarchaeota to Thaumarchaeota), which are shorter than expected from the wODR using intraphylum distances ([supplementary fig. S12, Supplementary Material](#) online) and display significantly

greater residuals than distances between Crenarchaeota and Euryarchaeota (supplementary fig. S13, Supplementary Material online). The absence of cluster#0 genes among Asgardarchaeota and Korarchaeota and the short interphyla distances to Thaumarchaeota homologs suggest an extensive loss among other phyla and horizontal acquisition by the thaumarchaeal ancestor from either crenarchaeal or euryarchaeal donors.

Despite the occurrence of accessory genes from cluster#1 and cluster#8 in methanogenic Euryarchaeota (fig. 5) and the enrichment of methane metabolism pathways (fig. 4), evolutionary histories of both CES clusters are not related (fig. 2). Gene families in CES cluster#8 did not display  $I_{ES} \geq 0.7$  outside its own cluster, constituting a separate connected component in the CES network depicted in figure 2. That said, cluster#8 gene families display shorter Euryarchaeota-Asgardarchaeota distances when compared with cluster#4 distances, but unlike cluster#0 and cluster#1, intra-Asgardarchaeota and intra-Euryarchaeota pairwise distances are not mutually compatible under a single linear regression (supplementary fig. S14, Supplementary Material online). The lack of a strong wODR anchor in the form of intraphyla distances suggests a more complex horizontal exchange history of cluster#8 genes, possibly involving intraphylum HGTs, which we cannot accurately assess with the data set used in this study. CES cluster#15 of accessory genes is well distributed among Crenarchaeota, and its intraphylum pairwise distances are congruent to cluster#4 distances, but their patchy occurrence among Euryarchaeota and Korarchaeota (fig. 5) does not permit a confident assessment of this cluster's interphyla evolutionary history (supplementary fig. S9, Supplementary Material online).

#### Consistency of Duplicated Gene Copies within CES Clusters

Among the eight larger CES clusters, 89 gene families occur in multiple copies among genomes; in order for CES phylogenies be reconstructed a single copy must be selected as the best representative of the evolutionary signal shared by CES genes. These 89 gene families are found in multiples a total of 237 times across the 42 sampled archaeal genomes, 52 times within CES cluster#4. During each wODR between pairs of gene families only the copy yielding the least sum of squared residuals is selected as best representing the shared history by both families. In 71.7% of cases when multiple copies are present in a genome the same copy is supported by more than 70% of similarly evolving gene families (supplementary fig. S16, Supplementary Material online). The significant difference between the observed support of selected copies against a null distribution where each copy has the same probability of yielding the least sum of squared residuals further corroborates  $I_{ES}$ 's effectiveness and robustness to stochastic noise (supplementary fig. S16, Supplementary Material online).

## Conclusions

We have presented  $I_{ES}$ , a new, robust, and efficient method to detect gene families with compatible evolutionary histories, which may predict good candidates to be used in phylogenomic tree reconstructions. The distance regression basis of our proposed method does not require hypotheses regarding evolutionary relationships between taxa represented by the branching pattern of phylogenetic trees. Besides significant gains in accuracy and computing efficiency compared with other tree-based approaches,  $I_{ES}$  introduces a new and robust strategy to pair copies of gene families that best represent shared evolutionary trends. The strong effect of genetic linkage in pairwise  $I_{ES}$  estimates within archaeal gene families constitute independent evidence of  $I_{ES}$ 's ability to recover shared evolutionary histories within empirical data sets.

Despite similar performances of Pearson's  $r$  and wODR  $R^2$  in detecting these trends,  $I_{ES}$  achieves the same result in a more efficient way. The utilization of wODR also imparts more robust statistical support not directly available to previous Pearson's  $r$  implementations, whereas the assessment of pairwise distances between taxa provides robustness in the presence of artifacts associated with phylogenetic inference (e.g., LBA). The ability to assess residuals of each data point independently also allows for evaluations of specific homologs, a useful tool for HGT detection.  $I_{ES}$  can thus be incorporated into phylogenomics pipelines and used to guide the selection of gene families for more accurate and robust species-tree inference, as well as the detection of meaningful clusters of gene families evolved in shared, yet reticulate, patterns. Results obtained from all three simulated scenarios and their replicates corroborate the efficiency of  $I_{ES}$  under multiple conditions, which further supports its application to assess distinct data sets.

By assessing similarities of evolutionary signal between archaeal gene families using  $I_{ES}$  we were able to detect several clusters of shared distinct evolutionary trends. Phylogenetic reconstruction using concatenated sequences from each of the four major CES clusters of core genes provides strong evidence for the existence of four major evolutionary trends. The phylogeny resulting from CES cluster#4, in particular, recovers a species tree hypothesis consistent with that proposed in several other studies, and does so while using a more empirically supported selection of gene families that does not presuppose vertical inheritance.

CES clusters obtained using  $I_{ES}$  also provide key evidence for horizontal exchange of functionally related genes between phyla (supplementary fig. S10, Supplementary Material online). For example, given the almost exclusive occurrence of genes from CES cluster#1 among methanogenic Euryarchaeota and Thaumarchaeota, tree-based approaches are not able to report the potential HGT between these phyla. Separately, intra- and interphyla distances obtained from CES cluster#1 are strongly correlated to distances described in CES cluster#4, however the significant placement of interphyla

distances below the wODR line strongly suggests an HGT between ancestors of both phyla.

The method used to analyze the archaeal gene set is general and can thus be applied to other genome sets. Furthermore, the  $I_{ES}$  implementation described provides a straightforward framework for future improvements and a possible alternative to phylogenetic reconciliation approaches to identify HGT events, as described in [supplementary figures S9–S15, Supplementary Material](#) online.

## Materials and Methods

ODR is an errors-in-variables regression method that accounts for measurement errors in both explanatory and response variables (Boggs et al. 1987), instead of attributing all errors exclusively to the response variable, as performed by OLS. Although OLS regression models seek to minimize the sum of squared residuals of the response variable, ODR minimizes the sum of squared residuals from each data point obtained by the combination of explanatory and response variables. Novichkov et al. (2004) assessed the compatibility between the evolutionary history of genes with a reference genomic evolutionary history using Pearson's  $r$  and estimates of an OLS regression's intercept. The latter extra step, when compared with other implementations using solely Pearson's  $r$  (Ramani and Marcotte 2003; Izarzugaza et al. 2008; Gueudré et al. 2016) is required to infer if data points not fitting a regression line through zero are caused by HGT. The approach proposed by Novichkov et al. requires two key conditions that restrict its usage on empirical data sets: 1) there must exist a reference history to which gene histories are compared; and 2) there are no errors in the reference distances between genomes.

The approach described here is based on ODR. Its modeling of errors within both assessed variables decreases the necessity of comparing gene family pairwise distances against a well-established reference. When evaluating evolutionary histories of two gene families with no clear separation between explanatory and response errors-in-variables approaches (e.g., ODR) are better suited to compare pairwise evolutionary distances. Our implementation uses a wODR model with regression line through the origin ( $\alpha = 0$ , where  $\alpha$  is the Y-axis intercept) to avoid overfitting ODR model to the detriment of coherent evolutionary assumptions. Data points are independently weighed based on residuals from an initial regression line to decrease the model's susceptibility to few homologs with strong signal incompatibility.

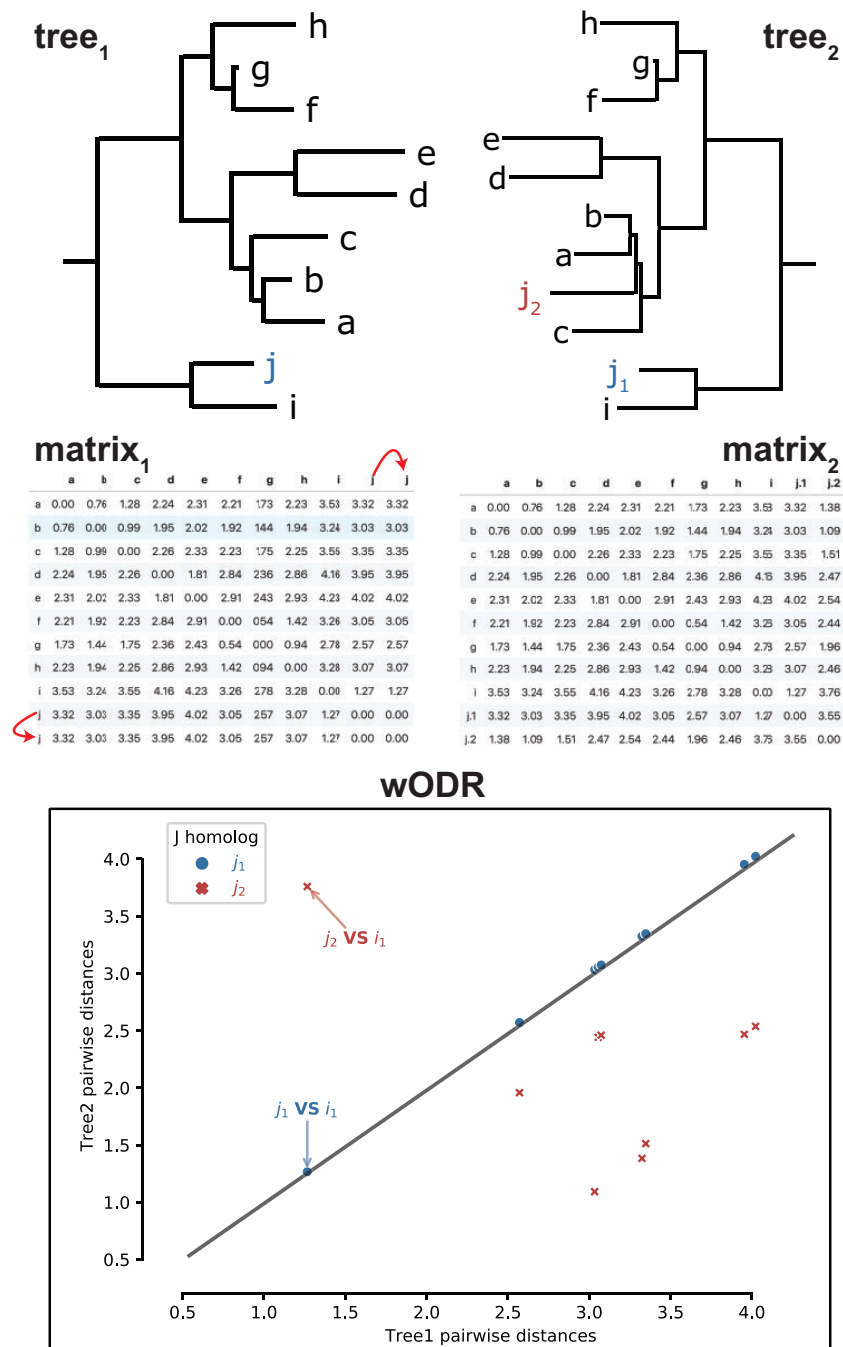
### Algorithm Explanation

In the simplest scenario of two gene families occurring as single copies in the same set of genomes,  $I_{ES}$  is equal to the coefficient of determination ( $R^2$ ) of a wODR between pairwise distance matrices of both gene families. Data points are weighted in regard to their impact on the model fitting.

The weighing step is required to avoid a few outlying sequences preventing the identification of a signal shared by the majority; weights are estimated as the inverse of residuals obtained from a geometric mean regression with intercept equal to zero and slope equal to  $s_Y/s_X$ , where  $s_Y$  and  $s_X$  are the standard deviations of the regressed variables. If two assessed gene families do not occur in the same set of genomes, the wODR  $R^2$  is calculated exclusively using homologs from genomes containing both families; the resulting  $R^2$  is then adjusted by the Bray–Curtis Index ( $I_{BC}$ ).  $I_{BC}$  is defined as  $1 - D_{BC}$ , where  $D_{BC}$  is the Bray–Curtis Dissimilarity (Bray and Curtis 1957) calculated from the copy number of each gene family within genomes. The incorporation of unequal genomic occurrence between gene families prevents possible overestimation of evolutionary signal similarity by the wODR  $R^2$  caused by gene losses and duplications that are not observed by the regression. [Supplementary figure S1, Supplementary Material](#) online, depicts how the decrease in taxa overlap can lead to overestimated shared evolution solely by wODR  $R^2$ , and consequently the importance of  $I_{BC}$  adjustment. We simulated two gene families separated by five Subtree Prune and Regraft (SPR) transformations and measured the evolutionary similarity between both gene families as we randomly removed one taxon from each simulated gene family ([supplementary fig. S1, Supplementary Material](#) online). As the set of taxa used during the regression becomes unrepresentative of underlying evolutionary processes, estimates based only on wODR  $R^2$  tend to artificially increase. We will subsequently refer to the wODR  $R^2 \times I_{BC}$  product as  $I_{ES}$ .

A main advantage of our proposed method over tree-based approaches is its ability to quantify the evolutionary signal shared by gene families with different copy numbers within genomes, as depicted in [figure 7](#). When estimating  $I_{ES}$  between one gene family occurring exclusively as single-copy ( $gene_1$ ) and another gene family ( $gene_2$ ) with two copies within genome  $J$  ( $j_1$  and  $j_2$ ), we initially select which of  $J$ 's copy of  $gene_2$  ( $j_1$  or  $j_2$ ) maximizes  $I_{ES}$  between both gene families. During an initial wODR step,  $gene_1$ 's pairwise distances involving its single  $J$  copy are duplicated in such a way that distances involving both  $j_1$  and  $j_2$  are initially paired with it ([fig. 7](#)); the  $gene_2$  copy in genome  $J$  yielding the least sum of squared residuals will be kept during further steps. The final pairwise  $I_{ES}$  will be estimated using the copy of  $gene_2$  that results in the least sum of residuals when paired with the single  $gene_1$  copy ( $j_1$  in [fig. 7](#)). Whenever both gene families occur in multiples within the same genome, all nonoverlapping products of copies from both families are part of the final  $I_{ES}$  estimation. In our implementation, wODR is performed through the SciPy (Virtanen et al. 2020) API of ODRPACK (Boggs et al. 1989). Our method's capability to automatically select gene copies that optimize evolutionary signal similarity between two gene families vastly expands the scope of data sets fit for general evaluation of evolutionary signal





**Fig. 7.**—Steps for  $I_{ES}$  estimation between gene families containing multiple gene copies.  $tree_1$  and  $tree_2$  are phylogenetic trees of two hypothetical gene families,  $gene_1$  and  $gene_2$ , respectively.  $matrix_1$  and  $matrix_2$  contain pairwise evolutionary distances between taxa from their respective gene families. The red arrows in  $matrix_1$  highlight the duplication of pairwise distances involving the  $j$  homolog of  $gene_1$  necessary to match dimensions of the two matrices. The wODR scatterplot displays the linear relationships between distances from both gene families, and highlights distances related to the  $j_1$  homolog of  $gene_2$  in blue and related to the  $j_2$  homolog in red. Arrows also highlight pairwise distances homologs in genomes  $J$  and  $I$  from both gene families.

compatibilities. The presence of multiple gene copies within a genome constitutes a key bottleneck to methods commonly used to assess the similarity of evolutionary histories. Tree-based evolutionary distance assessment algorithms are not generally capable of pairing genes between two gene families when at least one family contains multiple gene copies within

genomes (Stamatakis 2006; Nguyen et al. 2015; Gori et al. 2016; Huerta-Cepas et al. 2016); Pearson  $r$  implementations either rely on multiple tests (Gertz et al. 2003; Ramani and Marcotte 2003; Izarzugaza et al. 2008) or on predicting structural interaction between gene products (Gueudré et al. 2016).



Despite identical taxonomic occurrence of *gene1* and *gene2*, their copy numbers diverge within genome *J*, which likely arose from a horizontal exchange without replacement of *gene2*. To reflect this difference in evolutionary events between gene family histories in  $I_{ES}$  we adjust the resulting wODR  $R^2 = 1$  with an  $I_{BC} = 0.95$ .

### Statistics and Data Analysis

Pandas Python library (McKinney 2010) was used to perform operations on pairwise distance matrices and for generating condensed versions of the matrices submitted to the wODR model. Effect size ( $f$ ) hypothesis tests of differences between distributions were obtained using Common Language statistics (McGraw and Wong 1992), and  $P$  value correction for multiple tests was performed using False Discovery Rate implementation in StatsModels Python library (Seabold and Perktold 2010).

Network community detection was performed using the Louvain clustering method (Blondel et al. 2008) implementation available in the iGraph (Csardi and Nepusz 2006) Python library (igraph.community\_multilevel).

Enrichment of gene families within sets of genomes were assessed using hypergeometric tests and  $P$  values corrected with Benjamini–Hochberg’s false discovery rate and expressed as  $q$  values (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001).

### Data Simulation

Our simulated data set is composed of three sets of 50 trees generated through random stepwise perturbations from a single starting tree. Each set of 50 trees differs from the other on the intensity of the stepwise perturbations, which were simulated through 49 consecutive random SPR transformations with varying regrafting ranges. Small perturbations were caused by regrafting pruned subtrees to a branch within 10% of the branches closest to the original placement; medium perturbations regrafted within the 50% closest branches; and strong perturbations regrafted the pruned subtree to any branch in the tree. Additionally, each bipartition’s branch length was multiplied by independently drawn gamma distributed random variables ( $\mu = 1$  and  $\sigma = 0.2$ ) after each SPR. These three sets of simulated trees will be referred to as SPR10, SPR50, and SPR100 and were independently replicated ten times.

In a real-world scenario, multiple complex mechanisms can shuffle evolutionary signals without altering gene copy numbers (e.g., hidden paralogy, xenologous gene displacement, and incomplete lineage sorting); whereas these mechanisms tend to cause varying levels of perturbation to the underlying evolutionary signal the topological consequence to the tree is the same for all, an SPR. Through consecutive perturbations of the initial tree in the form of random SPR and branch length transformations, we obtained three sets of simulated gene

families with histories of greatly varying similarities. All trees were simulated with in-house scripts using ETE3 (Huerta-Cepas et al. 2016). All simulated trees are available in [Supplementary Material](#) online.

Simulated gene family phylogenies were used to generate sequence alignments containing insertions and deletions using INDELible (Fletcher and Yang 2009) (see [Supplementary Material](#) online), which outputs perfectly aligned homologous sites of simulated sequences. Sequences from each resulting simulated gene family were also aligned using MAFFT (Kato and Standley 2013) with the `–auto` parameter; both the true alignment provided by INDELible and the empirical alignment generated by MAFFT were used to construct phylogenetic trees and pairwise distance matrices using IQTree (Nguyen et al. 2015). Differences between aligned homologous sites simulated by INDELible and the sequence alignment obtained using MAFFT were assessed using Sum-of-Pairs score (SP-score) calculated by FastSP (Mirarab and Warnow 2011).

### Archaeal Empirical Data Set

Complete genome sequences of 42 Archaea from the Euryarchaeota phylum and from TACK, DPANN, and Asgardarchaeota groups were downloaded from NCBI GenBank ([Supplementary table S1, Supplementary Material](#) online). Other candidate phyla known from metagenomic sequences as well as some remaining members of the DPANN group were not included, as their expected phylogenetic relationships are not as well understood. Archaea was selected as the test data set since the evolutionary relationships between some major groups are well-established, whereas others remain contested. Furthermore, many sets of archaeal metabolic genes have a strong phyletic dependence (e.g., methanogenesis among Euryarchaeota; Borrel et al. 2013), therefore facilitating the assessment of shared evolutionary trends driven by similar ecological and/or metabolic requirements. Clustering of homologous proteins was performed using the orthoMCL (Li et al. 2003) implementation available in the GET\_HOMOLOGUES package (Contreras-Moreira and Vinuesa 2013). Evolutionary similarity comparisons were restricted to gene families present in at least ten genomes.

Pairwise maximum likelihood distances between homologous proteins were generated using IQTree under the LG+G evolutionary model. Phylogenetic trees from clusters of gene families with CES and extended core genome (i.e., single copy and present in at least 35 out of the 42 sampled Archaea genomes) were reconstructed from concatenated multiple sequence alignments using the LG+C60+F+G and individual partitions corresponding to each concatenated gene.

Enrichment of gene functions among CES clusters were performed using StringDB API (Szklarczyk et al. 2019). For each genome, homologs from CES gene families were submitted independently for enrichment assessment. Retrieved

protein annotations are available in the [Supplementary Material](#) online.

### Tree-Based Metrics of Evolutionary Similarity

All four tree-based metrics were used to calculate distances between all pairwise combinations of trees reconstructed from simulated alignments.  $D_{\text{geo}}$  was calculated using the treeCI Python package (Gori et al. 2016), RF was obtained using ETE3, and both  $D_{\text{ms}}$  and  $D_{\text{qt}}$  were calculated using treeCmp (Bogdanowicz et al. 2012).

Given the nonadditive nature of tree-based metrics and uniform probability of simulated SPR moves across all branches, estimates from tree-based methods showed substantial degrees of saturation when compared with the number of perturbations between simulated gene families. In comparisons presented below we used a transformation that applies, if the approach to the steady state follows an exponential decay:  $d_{\text{adj}} = -\ln(1 - d_{\text{norm}})$ , where  $d_{\text{adj}}$  and  $d_{\text{norm}}$  are the adjusted and normalized distance estimates. For sequence divergence, this is known as Poisson Correction (Nei and Zhang 2006).

### Supplementary Material

[Supplementary data](#) are available at *Genome Biology and Evolution* online.

### Acknowledgments

This work was supported by the Simons Foundation Collaboration on the Origins of Life (Award No. 339603) and NSF Integrated Earth Systems Program (Award No. 1615426) to G.P.F. S.M.S. was supported through Geisel School of Medicine at Dartmouth's Center for Quantitative Biology through a grant from the National Institute of General Medical Sciences of the National Institutes of Health (Award No. P20GM130454). J.C.S. was funded in part by a CNPq senior researcher fellowship.

### Data Availability

The data underlying this article are available in the article and in its [supplementary Material](#) online. Configuration files necessary to reproduce simulated data sets as well as detailed characterization of the used empirical data set are available at figshare (10.6084/m9.figshare.12986051). A Python implementation of  $I_{\text{ES}}$  is available at <https://github.com/thiberio/evolSimIndex> and ready to use docker container of the Python implementation is available at <https://hub.docker.com/r/thiberio/evolsimindex>.

### Literature Cited

Andam CP, Gogarten JP. 2011. Biased gene transfer and its implications for the concept of lineage. *Biol Direct*. 6:47.

- Andam CP, Williams D, Gogarten JP. 2010. Biased gene transfer mimics patterns created through shared ancestry. *Proc Natl Acad Sci U S A*. 107(23):10679–10684.
- Baptiste E, et al. 2009. Prokaryotic evolution and the tree of life are two different things. *Biol Direct*. 4:34.
- Barker D, Pagel M. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput Biol*. 1(1):e3.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 57(1):289–300.
- Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 29(4):1165–1188.
- Billera LJ, Holmes SP, Vogtmann K. 2001. Geometry of the space of phylogenetic trees. *Adv Appl Math*. 27(4):733–767.
- Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. *J Stat Mech*. 2008(10):P10008.
- Bogdanowicz D, Giaro K. 2012. Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Trans Comput Biol Bioinform*. 9(1):150–160.
- Bogdanowicz D, Giaro K, Wróbel B. 2012. TreeCmp: comparison of trees in polynomial time. *Evol Bioinform Online*. 8:EBO.S9657.
- Boggs PT, Byrd RH, Schnabel RB. 1987. A stable and efficient algorithm for nonlinear orthogonal distance regression. *SIAM J Sci Stat Comput*. 8(6):1052–1078.
- Boggs PT, Donaldson JR, Byrd R. h, Schnabel RB. 1989. Algorithm 676: ODRPACK: software for weighted orthogonal distance regression. *ACM Trans Math Softw*. 15(4):348–364.
- Borrel G, et al. 2013. Phylogenomic data support a seventh order of methylotrophic methanogens and provide insights into the evolution of methanogenesis. *Genome Biol Evol*. 5(10):1769–1780.
- Bray JR, Curtis JT. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecol Monogr*. 27(4):325–349.
- Brochier-Armanet C, Forterre P, Gribaldo S. 2011. Phylogeny and evolution of the Archaea: one hundred genomes later. *Curr Opin Microbiol*. 14(3):274–281.
- Brochier C, Gribaldo S, Zivanovic Y, Confalonieri F, Forterre P. 2005. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? *Genome Biol*. 6(5):R42.
- Contreras-Moreira B, Vinuesa P. 2013. GET\_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol*. 79(24):7696–7701.
- Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal Complex Sy*. 1695: 1695.
- Da Cunha V, 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:e1006810.
- Dagan T, Martin W. 2006. The tree of one percent. *Genome Biol*. 7(10):118.
- Dutilh BE, et al. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun*. 5:1–11.
- Estabrook GF, McMorris FR, Meacham CA. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Syst Biol*. 34(2):193–200.
- Feng Y, et al. 2021. Reconstructing the evolutionary origins of extreme halophilic archaeal lineages. *Genome Biol Evol*. 13(8):evab166. doi: 10.1093/gbe/evab166.
- Fletcher W, Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*. 26(8):1879–1888.
- Gertz J, et al. 2003. Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics* 19(16):2039–2045.

- Goh C-S, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Co-evolution of proteins with their interaction partners. *J Mol Biol.* 299(2):283–293.
- Gori K, Suchan T, Alvarez N, Goldman N, Dessimoz C. 2016. Clustering genes of common evolutionary history. *Mol Biol Evol.* 33(6):1590–1605.
- Gueudré T, Baldassi C, Zamparo M, Weigt M, Pagnani A. 2016. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc Natl Acad Sci U S A.* 113(43):12186–12191.
- Huber H, et al. 2002. A new phylum of Archaea represented by a nano-sized hyperthermophilic symbiont. *Nature* 417(6884):63–67.
- Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol.* 33(6):1635–1638.
- Izarguzaga JMG, Juan D, Pons C, Pazos F, Valencia A. 2008. Enhancing the prediction of protein pairings between interacting families using orthology information. *BMC Bioinformatics* 9:35.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4):772–780.
- Kimmel R, Sethian J. A. 1998. Computing geodesic paths on manifolds. *Proc Natl Acad Sci U S A.* 95(15):8431–8435.
- Koonin EV, Wolf YI, Puigbò P. 2009. The phylogenetic forest and the quest for the elusive tree of life. *Cold Spring Harb Symp Quant Biol.* 74:205–213.
- Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. 2005. The net of life: reconstructing the microbial phylogenetic network. *Genome Res.* 15(7):954–959.
- Kupczok A, von Haeseler A, Klaere S. 2008. An exact algorithm for the geodesic distance between phylogenetic trees. *J Comput Biol.* 15(6):577–591.
- Leigh JW, Susko E, Baumgartner M, Roger AJ. 2008. Testing congruence in phylogenomic analysis. *Syst Biol.* 57(1):104–115.
- Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13(9):2178–2189.
- Lin Y, Rajan V, Moret BME. 2012. A metric for phylogenetic trees based on matching. *IEEE/ACM Trans Comput Biol Bioinform.* 9(4):1014–1022.
- Liu C, Wright B, Allen-Vercoe E, Gu H, Beiko R. 2018. Phylogenetic clustering of genes reveals shared evolutionary trajectories and putative gene functions. *Genome Biol Evol.* 10(9):2255–2265.
- Martijn J, et al. 2020. Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nat Commun.* 11(1):5490.
- McGraw KO, Wong SP. 1992. A common language effect size statistic. *Psychol Bull.* 111(2):361–365.
- McKinney W. 2010. Data structures for statistical computing in python. In: van der Walt S, Millman J, editors. *Proceedings of the 9th Python in Science Conference; 2010 June 28–July 3; Austin (TX).* p. 56–61.
- Meilă M. 2007. Comparing clusterings—an information based distance. *J Multivar Anal.* 98(5):873–895.
- Mirarab S, Bayzid MS, Bossau B, Warnow T. 2014. Statistical binning improves species tree estimation in the presence of gene tree heterogeneity. *Science* 346(6215):1250463.
- Mirarab S, Warnow T. 2011. FASTSP: linear time calculation of alignment accuracy. *Bioinformatics* 27(23):3250–3258.
- Narasingarao P, et al. 2012. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J.* 6(1):81–93.
- Nei M, Zhang J. 2006. Evolutionary distance: estimation. In: *Encyclopedia of life sciences.* Chichester (United Kingdom): John Wiley & Sons, Ltd. p. 1–4.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 32(1):268–274.
- Novichkov PS, et al. 2004. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. *J Bacteriol.* 186(19):6575–6585.
- Owen M, Provan JS. 2011. A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Trans Comput Biol Bioinform.* 8(1):2–13.
- Papke RT, Gogarten JP. 2012. Ecology. How bacterial lineages emerge. *Science* 336(6077):45–46.
- Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* 14(9):609–614.
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A.* 96(8):4285–4288.
- Petitjean C, Deschamps P, López-García P, Moreira D. 2014. Rooting the domain archaea by phylogenomic analysis supports the foundation of the New Kingdom proteoarchaeota. *Genome Biol Evol.* 7(1):191–204.
- Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol.* 14(5):615–623.
- Puigbò P, Wolf YI, Koonin EV. 2009. Search for a ‘Tree of Life’ in the thicket of the phylogenetic forest. *J Biol.* 8(6):59.
- Ramani AK, Marcotte EM. 2003. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol.* 327(1):273–284.
- Rangel LT, et al. 2019. Identification and characterization of putative *Aeromonas* spp. T3SS effectors. *PLoS One* 14(6):e0214035.
- Raymann K, Forterre P, Brochier-Armanet C, Gribaldo S. 2014. Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea. *Genome Biol Evol.* 6(1):192–212.
- Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci.* 53(1–2):131–147.
- Seabold S, Perktold J. 2010. statsmodels: econometric and statistical modeling with python. In: van der Walt S, Millman J, editors. *Proceedings of the 9th Python in Science Conference; p. 92–96.* doi: 10.25080/Majora-92bf1922-011
- Shapiro BJ, et al. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science* 336(6077):48–51.
- Sorek R, Zhu Y, Creevey C, Francino M. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318:1449–1452.
- Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
- Szklarczyk D, et al. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47(D1):D607–D613.
- Szölösi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013. Efficient exploration of the space of reconciled gene trees. *Syst Biol.* 62(6):901–912.
- Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol.* 3(9):711–721.
- Urbanavicius J, et al. 2008. Acquisition of a bacterial RnaA-type tRNA(uracil-54, C5)-methyltransferase by Archaea through an ancient horizontal gene transfer. *Mol Microbiol.* 67(2):323–335.
- Vert J-P. 2002. A tree kernel to analyse phylogenetic profiles. *Bioinformatics* 18(Suppl 1):S276–S284.
- Virtanen P, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 17(3):261–272.
- Williams TA, Cox CJ, Foster PG, Szölösi GJ, Embley TM. 2020. Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol.* 4(1):138–147.
- Williams TA, et al. 2015. New substitution models for rooting phylogenetic trees. *Philos Trans R Soc Lond B Biol Sci.* 370(1678):20140336.
- Williams TA, et al. 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc Natl Acad Sci U S A.* 114(23):E4602–E4611.
- Zhaxybayeva O, Stepanauskas R, Mohan NR, Papke RT. 2013. Cell sorting analysis of geographically separated hypersaline environments. *Extremophiles* 17(2):265–275.

Associate editor: Mario dos Reis