

1 Submission intended as an Article in the Methods section of MBE

2

3 **Non-phylogenetic identification of co-evolving genes for reconstructing the archaeal Tree of**  
4 **Life**

5

6 L. Thibério Rangel<sup>1</sup>, Shannon M. Soucy<sup>2</sup>, João Carlos Setubal<sup>3</sup>, Johann Peter Gogarten<sup>4</sup>, Gregory  
7 P. Fournier<sup>1</sup>

8 <sup>1</sup> Department of Earth, Atmospheric & Planetary Sciences, Massachusetts Institute of  
9 Technology, Cambridge, MA, USA

10 <sup>2</sup> Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College,  
11 Hanover, NH, USA

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

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

15

16 Corresponding author: Gregory P. Fournier, [g4nier@mit.edu](mailto:g4nier@mit.edu)

17

## 1 Abstract

2 Assessing the phylogenetic compatibility between individual gene families is a crucial and often  
3 computationally demanding step in many phylogenomics analyses. Here we describe the  
4 Evolutionary Similarity Index ( $I_{ES}$ ) to assess shared evolution between gene families using a  
5 weighted Orthogonal Distance Regression applied to sequence distances. This approach allows  
6 for straightforward pairing of paralogs between co-evolving gene families without resorting to  
7 multiple tests, or *a priori* assumptions of molecular interactions between protein products from  
8 assessed genes. The utilization of pairwise distance matrices, while less informative than  
9 phylogenies, circumvents error-prone comparisons between trees whose topologies are  
10 inherently uncertain. Analyses of simulated tree datasets showed that  $I_{ES}$  was more accurate and  
11 less susceptible to phylogenetic noise than existing tree-based methods (Robinson-Foulds and  
12 geodesic distance) for assessing evolutionary signal compatibility. Applying  $I_{ES}$  to a real dataset  
13 of 1,322 genes from 42 archaeal genomes identified eight major clusters of co-evolving gene  
14 families. Four of these clusters included genes with a taxonomic distribution across all archaeal  
15 phyla, while other clusters included a subset of taxa that do not map to generally accepted  
16 archaeal clades, indicating possible shared horizontal transfers by co-evolving gene families. We  
17 identify one strongly connected set of 62 co-evolving genes occurring as both single-copy and  
18 multiple homologs per genome, with compatible evolutionary histories closely matching  
19 previously published species trees for Archaea. An  $I_{ES}$  implementation is available at  
20 <https://github.com/lthiberiol/evolSimIndex>.

## 21 Introduction

22 Phylogenies reconstructed from single genes are known to poorly reflect the underlying  
23 history of whole genomes, as the detectable phylogenetic signal from an isolated locus cannot be  
24 extrapolated to represent whole genomes (Dagan and Martin 2006; Baptiste et al. 2009; Koonin  
25 et al. 2009). To this effect, it has become common practice to estimate species' evolutionary  
26 histories by concatenating multiple sequence alignments of core genes, which greatly increases  
27 the number of sites available for phylogenetic inference. The preference towards concatenating  
28 core genes is due its expected to horizontal gene transfer (HGT) (Thomas and Nielsen 2005;  
29 Sorek et al. 2007; Popa and Dagan 2011); however, despite the lower frequency of HGT among  
30 some gene families, horizontal exchange still takes place within their history. The slow

1 substitution rate and corresponding high sequence conservation of the core genome can become  
2 a liability due to the increase in neutral and nearly-neutral HGT at the genus and species level  
3 (Papke and Gogarten 2012; Shapiro et al. 2012). Biases in horizontal exchange between closely  
4 related genomes may even reinforce the misconception of strong HGT resistance (Andam et al.  
5 2010; Andam and Gogarten 2011).

6         Given these processes, surveying evolutionary compatibility between different gene  
7 families is important to minimize conflicting evolutionary signals combined during  
8 phylogenomic reconstruction. Multiple strategies have been proposed to assess similarities  
9 between the phylogenetic signals found within individual genes- e.g., Robinson-Foulds  
10 bipartition compatibility (RF) (Robinson and Foulds 1981) and geodesic distance (Kimmel and  
11 Sethian 1998; Kupczok et al. 2008; Owen and Provan 2011). The majority of these methods are  
12 based on straightforward comparisons between tree topologies (Kunin et al. 2005; Leigh et al.  
13 2008; Puigbò et al. 2009; Mirarab et al. 2014; Gori et al. 2016). While an intuitive solution,  
14 comparisons between tree topologies require phylogenetic trees of all assessed gene families to  
15 be accurately reconstructed, adding a substantial computational cost to a reliable execution of an  
16 already computationally demanding task. Furthermore, the vastness of tree space, combined with  
17 the inherent uncertainty of phylogenetic reconstruction, constitutes an error-rich layer in tree-  
18 based evolutionary similarity assessments.

19         Accounting for uncertainty-based variations in tree topology (i.e., bipartition support)  
20 further increases the computational burden and decreases the resolution of the evaluated  
21 phylogenetic signal (e.g., collapsing low support bipartitions or weighing them based on  
22 support). A proposed solution to bypass the computational cost of tree similarity assessments is  
23 Pearson's correlation coefficient ( $r$ ) between evolutionary distance matrices (Goh et al. 2000;  
24 Pazos and Valencia 2001; Novichkov et al. 2004; Rangel et al. 2019). Despite its application in  
25 protein-protein interaction studies, the sensitivity of Pearson's  $r$  to phylogenetic noise and the  
26 granularity of its estimates have yet to be compared to those of tree-based metrics. Unlike tree-  
27 based comparisons, methods based on Pearson's  $r$  enable simple implementations to detect co-  
28 evolving gene families with histories complicated by multiple homologs within genomes by  
29 estimating correlation coefficients using all possible pairings of paralogs between gene families  
30 (Gertz et al. 2003; Ramani and Marcotte 2003). Direct Coupling Analysis (DCA) has also been  
31 used to pair gene copies between co-evolving gene families (Gueudré et al. 2016), but despite

1 positive results the assumption that products of co-evolving genes must be structurally associated  
2 exerts extreme burden toward its general applicability.

3

## 4 New approaches

5 Here we propose the Evolutionary Similarity Index ( $I_{ES}$ ) as a metric of evolutionary  
6 history similarities based on weighted Orthogonal Distance Regression (wODR) between  
7 evolutionary pairwise distance matrices. Simulations show that wODR performed very similarly  
8 to Pearson correlation coefficients, with the added advantage of more robust estimated  
9 relationships between gene families. This approach does not require multiple tests where there  
10 are gene duplications. We show that evolutionary similarity estimates from wODR display a  
11 linear relationship with stepwise perturbations in tree topologies. More common estimates of tree  
12 similarity, such as RF and geodesic distances, tend to overestimate the impact of topology  
13 changes, and consequently are significantly more susceptible to phylogenetic noise. We further  
14 assessed evolutionary similarities across 1,322 archaeal gene families and detected significant  
15 evolutionary incompatibilities between conserved single-copy genes, as well as a clear central  
16 evolutionary tendency involving 62 gene families that occur as both single and multiple-copies  
17 across genomes.

18

## 19 Methodology

20 Orthogonal Distance Regression (ODR) is an errors-in-variables regression method that  
21 accounts for measurement errors in both explanatory and response variables (Boggs et al. 1987),  
22 instead of attributing all errors in the expected values exclusively to the response variable, as  
23 performed by Ordinary Least Squares (OLS). While OLS regressions seek to minimize the sum  
24 of squared residuals of the response variable, ODR minimizes the sum of squared residuals from  
25 each data point obtained by the combination of explanatory and response variables. Novichkov et  
26 al. (Novichkov et al. 2004) assessed the compatibility between the evolutionary history of genes  
27 with a reference genomic evolutionary history using Pearson's  $r$  and estimates of an OLS  
28 regression's intercept. This latter extra step when compared to other implementations using  
29 Pearson's  $r$  (Ramani and Marcotte 2003; Izarzugaza et al. 2008; Gueudré et al. 2016) is required

1 for a robust inference given that at a hypothetical time exactly after genome divergence,  
2 distances between homologs in both genomes must be zero. The approach proposed by  
3 Novichkov et al. requires two key assumptions that restrict the general applicability of  
4 evolutionary assessments of empirical datasets: 1) there must exist a reference history to which  
5 gene histories are compared; and 2) there are no errors in reference distances between genomes.

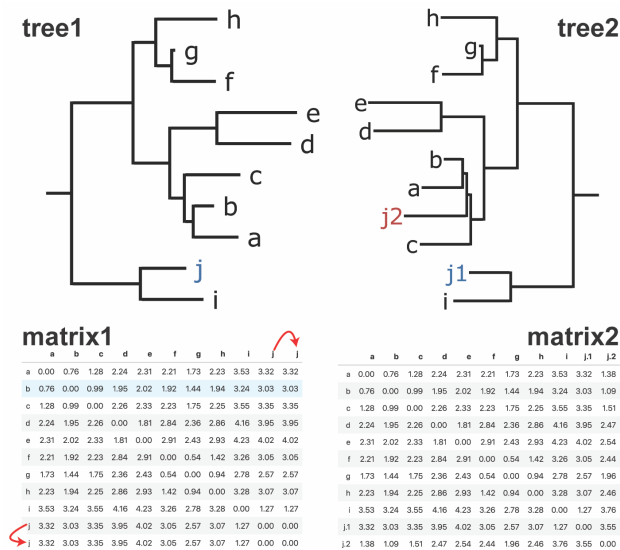
6 The approach described here is based on ODR. Its modelling of errors within both  
7 assessed variables decreases the necessity of a well-established reference distance to compare  
8 gene family pairwise distances against. Consequently, errors-in-variables approaches (e.g., ODR)  
9 are better suited to compare pairwise evolutionary distances between two gene families, where *a*  
10 *priori*, there is no clear separation between explanatory and response variables. Independently  
11 weighing residuals from each data point provides a framework less susceptible to  
12 underestimating overall evolutionary similarities due to few homologs with high incompatibility.  
13 Our implementation also fits wODR while forcing the intercept through the origin, which avoids  
14 overfitting the linear regression model to the detriment of coherent evolutionary assumptions.

15

## 16 Algorithm explanation

17 Tree-based evolutionary distance assessment algorithms are not generally capable of  
18 pairing genes between two gene families when at least one family contains multiple gene copies  
19 (Stamatakis 2006; Nguyen et al. 2015; Gori et al. 2016; Huerta-Cepas et al. 2016). Pearson *r*  
20 implementations either rely on multiple tests (Gertz et al. 2003; Ramani and Marcotte 2003;  
21 Izarzugaza et al. 2008) or on predicting structural interaction between gene products (Gueudré et  
22 al. 2016). Our implementation performs an initial wODR using all pairs of genes within the same  
23 genome, one from each assessed gene family, and reports gene pairs that minimize the sum of  
24 squared residuals. As exemplified in Fig. 1, a hypothetical *gene1* occurs exclusively as single  
25 copy across 10 genomes (Fig. 1, tree1), while *gene2* has an extra copy within genome *J* (Fig. 1,  
26 tree2). In order to identify which copy of *gene2* in *J* (*j1* or *j2*) better represents their shared  
27 evolution we compare *gene1* pairwise distances involving *j* with *gene2* pairwise distances  
28 involving *j1* and *j2*. Consequently, to do that we must duplicate *J*'s rows and columns in matrix1  
29 to match matrix2 dimensions (Fig. 1, matrix1). The scatter plot in Fig. 1 highlights pairwise  
30 distances involving *j1* in blue and *j2* in red, and as shown by the fitted wODR regression, *gene2*  
31 pairwise distances involving *j1* fits better to the expected linear association between matrix1 and

1 matrix2 than pairwise distances involving  $j2$ . The smallest sum of residuals obtained by the  $j1$   
 2 homolog of  $gene2$  correctly pairs it with  $J$ 's  $gene1$  homolog, while  $j2$ 's  $gene2$  homolog is likely  
 3 a product of HGT from a shared common ancestor of  $A$  and  $B$ . When both gene families occur in  
 4 multiples within the same genome, all pairs of unique loci are reported. Once best matching  
 5 genes from each gene family are paired, or if both occur exclusively as single copy, a final  
 6 wODR is performed using paired homologs from each gene family. wODR is performed through  
 7 the SciPy (Virtanen et al. 2020) API of ODRPACK (Boggs et al. 1989). Initial weights of  
 8 pairwise distance are estimated as the inverse of residuals obtained from geometric distance  
 9 regression with no intercept and slope equal to  $s_Y/s_X$ , where  $s_Y$  and  $s_X$  are standard deviations  
 10 from the regressed distance matrices.



11  
 12 *Fig. 1* Steps for  $I_{ES}$  estimation between gene families containing multiple gene copies. *tree1* and *tree2* are phylogenetic trees of  
 13 two hypothetical gene families, *gene1* and *gene2*, respectively. *matrix1* and *matrix2* contain pairwise evolutionary distances  
 14 between taxa from their respective gene families. The red arrows in *matrix1* highlight the duplication of pairwise distances  
 15 involving the  $j$  homolog of *gene1* necessary to match dimensions of the two matrices. The wODR scatterplot fits the linear

1 relationships between distances from both gene families, and highlights distances related to the *j1* homolog of *gene2* in blue and  
2 related to the *j2* homolog in red. Arrows also highlight pairwise distances homologs in genomes *J* and *I* from both gene families.

3         Given that regression models only account for data points equally represented in both  
4 assessed variables, gene losses and duplications are not directly accounted for when comparing  
5 evolutionary histories through wODR. To incorporate unequal genomic occurrence between  
6 gene families to our proposed measurement of evolutionary similarity, the wODR Coefficient of  
7 Determination, i.e.  $R^2$ , is adjusted by the Bray-Curtis Index ( $I_{BC}$ ).  $I_{BC}$  is defined as  $1 - D_{BC}$ ,  
8 where  $D_{BC}$  is Bray-Curtis Dissimilarity (Bray and Curtis 1957) calculated from absolute  
9 genome counts in each gene family. From hereon we will refer to the wODR  $R^2 \times I_{BC}$  product  
10 as  $I_{ES}$ . Continuing with the example depicted in Fig. 1, despite *gene1* and *gene2* identical  
11 genomic occurrence, their copy numbers diverge within genome *J*, which as mentioned before,  
12 arose from a horizontal exchange of *gene2*. To reflect this difference in evolutionary events  
13 within gene family histories in the proposed  $I_{ES}$ , the resulting wODR  $R^2 = 1$  is adjusted using an  
14  $I_{BC} = 0.95$ .

15

## 16 **Statistics and data analysis**

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

23

## 24 **Data Simulation**

25         We constructed ten simulated datasets, each one containing 50 trees generated from  
26 stepwise random Subtree Prune and Regraft (SPR) transformations. Each dataset contains one  
27 initial random rooted tree with 50 taxa generated by ETE3 (Huerta-Cepas et al. 2016). To obtain  
28 the remaining 49 trees, the initial tree (*tree\_1*) undergoes a series of 49 consecutive SPR  
29 transformations in such a way that *tree\_1* differs from *tree\_2*, *tree\_3*, and *tree\_n* by 1, 2, and  $n$   
30 SPR transformations, respectively. At each SPR the branch leading to the regrafted clade  
31 undergoes two transformations to simulate changes in substitution rate after an HGT event. The

1 first transformation multiplies the branch length by a random uniform variable ranging from 0 to  
2 1, simulating at which point during the branch's history the transfer occurred. The second  
3 transformation multiplies by a random gamma distributed variable ( $\alpha = \beta = 100$ ), simulating  
4 changes in substitution rates in the recipient clade after said transfer. All simulated trees are  
5 available in Supplementary Material.

6 All simulated trees were also used to generate sequence simulations using INDELible  
7 (Fletcher and Yang 2009) (Supplementary Material). Phylogenetic trees and pairwise distance  
8 matrices were reconstructed using IQTree (Nguyen et al. 2015) using the LG+G model.

### 10 [Archaeal empirical dataset](#)

11 Complete genome sequences of 42 Archaea were downloaded from NCBI GenBank  
12 (Supplementary Table S1), and clustering of homologous proteins performed using the  
13 orthoMCL (Li et al. 2003) implementation available at GET\_HOMOLOGUES (Contreras-  
14 Moreira and Vinuesa 2013; Vinuesa and Contreras-Moreira 2015). Archaea were selected as the  
15 test dataset since the evolutionary relationships between some major groups are well-established,  
16 while others remain contested. Furthermore, many sets of archaeal metabolic genes have a strong  
17 phyletic dependence (e.g., methanogenesis among Euryarchaeota) more easily permitting tests of  
18 gene co-evolution at different phylogenetic distances. Evolutionary similarity comparisons were  
19 restricted to homologous groups present in at least 10 genomes. Pairwise maximum likelihood  
20 distances between homologous proteins were generated using IQTree and LG+G model.

21 Enrichment of gene functions among co-evolving gene families were performed using  
22 StringDB API (Szklarczyk et al. 2019). For each genome, homologs from co-evolving gene  
23 families were submitted independently for enrichment assessment. Retrieved protein annotations  
24 are available in the Supplementary Material.

### 26 [Geodesic and Robinson-Foulds distance calculations](#)

27 Geodesic distances between single copy gene families (both simulated and real datasets)  
28 were calculated using the treeCl Python package (Gori et al. 2016). RF distances between single  
29 copy gene families were calculated by ETE3.

30



## 1 Results and Discussion

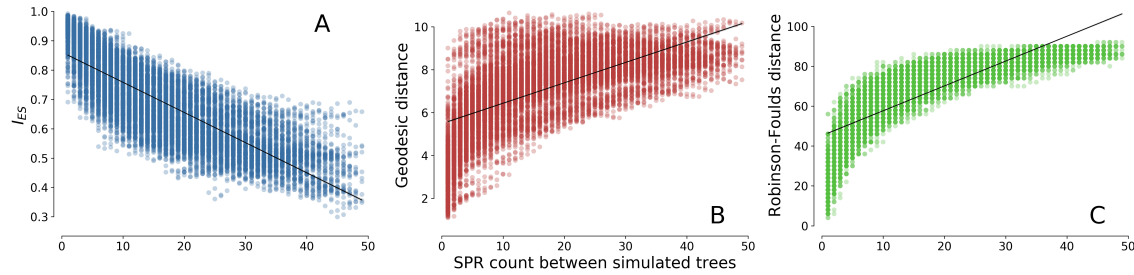
2

### 3 Simulated dataset

4 Evolutionary histories between simulated gene families were compared to each other  
5 using three distinct metrics: RF, geodesic distance, and  $I_{ES}$ . Results reported by all three  
6 approaches successfully identified the monotonic increase in SPR operations from a starting tree  
7 (Fig. 2 and Supplementary Fig. S1). Measurements obtained from RF and geodesic approaches,  
8 however, frequently overestimated the impact of SPR transformations between two gene  
9 families, leading to a fast saturation of dissimilarities between evolutionary histories (Fig. 2B  
10 and Fig. 2C). The dissimilarity saturations detected by RF and geodesic measurements occur as  
11 they fail to identify the decreasing similarity between two trees separated by more than 15 to 20  
12 SPR transformations, or even 10 SPR in some replicates (simulation replicate #1, Supplementary  
13 Fig. S1). Both of these approaches rely on the proportion of compatible bipartitions shared by  
14 two trees, which is very susceptible to small changes at deep bipartitions, where changing one  
15 single leaf can potentially create fully incompatible bipartition tables.

16 In contrast,  $I_{ES}$  displayed a robust linear relationship with the number of SPR  
17 transformations between gene families ( $\bar{r} = -0.87$ , Fig. 2A). The lower level of information  
18 assessed by  $I_{ES}$ , pairwise distance matrices instead of dichotomic trees, is less susceptible to  
19 dissimilarity saturation, corresponding to a more linear relation between expected and observed  
20 changes in evolutionary histories. Furthermore,  $I_{ES}$  is much more efficient, computationally.  
21 Tree reconstruction of the alignment simulated from *tree\_1* of the first simulation replicate (50  
22 taxa and 500 sites with no indels) under LG+G model in a single thread by IQTree took 80.604  
23 seconds, while exclusively calculating pairwise distance matrix for the same alignment took  
24 1.713 seconds. Both computations were performed in a 3 GHz Intel Xeon W. The difference in  
25 computing time of almost 50x, without bipartition support assessment, shows another, practical  
26 advantage for assessing evolutionary similarity through  $I_{ES}$  in large datasets.

27



1  
2 *Fig. 2* Scatter plot of evolutionary similarity metrics against number of SPR transformations between simulated gene families  
3 from all ten replicates. Solid black lines are estimated from OLS regressions between number of SPR transformations and  
4 evolutionary similarity metrics. All three scatter plots display the number of SPR transformations between two trees in the X-  
5 axis, while varying the evolution similarity metric displayed in the Y-axis. A) displays wODR  $R^2$  between distance matrices of  
6 simulated gene families in the Y-axis, B) displays geodesic distances between trees reconstructed from each simulated alignment,  
7 and C) displays RF distances estimated from the same trees.

## 8 9 Robustness assessment between approaches

10 The dichotomic pattern in a cladogram is extremely susceptible to phylogenetic  
11 uncertainty, combined with the vast tree space available for 50 taxa, causing topological  
12 variations from phylogenetic uncertainty to not be directly differentiated from real deviations in  
13 evolutionary history (Szöllösi et al. 2013). The simpler information used to estimate  $I_{ES}$  (i.e.,  
14 pairwise Maximum Likelihood distances) is less prone to such uncertainty as it bypasses forming  
15 hypotheses about the evolutionary relationships between taxa. This assumption is corroborated  
16 by pairwise comparisons within bootstrap replicates, where  $I_{ES}$  correctly detected replicates as  
17 such, i.e., virtually identical to each other, while RF and geodesic measures failed to identify the  
18 common nature of bootstrap replicates. In addition to its accurate predictions,  $I_{ES}$  consistently  
19 displayed very little variance within its estimates between bootstrap replicates.

20 Each sequence simulation alignment was used to generate 10 bootstrap replicates.  
21 Pairwise comparisons between 10 bootstrap replicates summed up to 45 comparisons within a  
22 single alignment, given that we simulated a total of 500 alignments, we assessed 22,500 pairwise  
23 comparisons between bootstrap replicates across all simulated datasets.  $I_{ES}$  values correctly  
24 identified bootstrap replicates as sharing virtually identical evolutionary histories,  $\overline{R^2} = 0.96$ ,  
25 and did so very consistently (CV = 1.39%, where CV stands for Coefficient of Variation).  
26 Despite successfully identifying increasing evolutionary changes between simulated trees, RF  
27 distances inconsistently predicted similarities between histories of bootstrapped trees (CV =

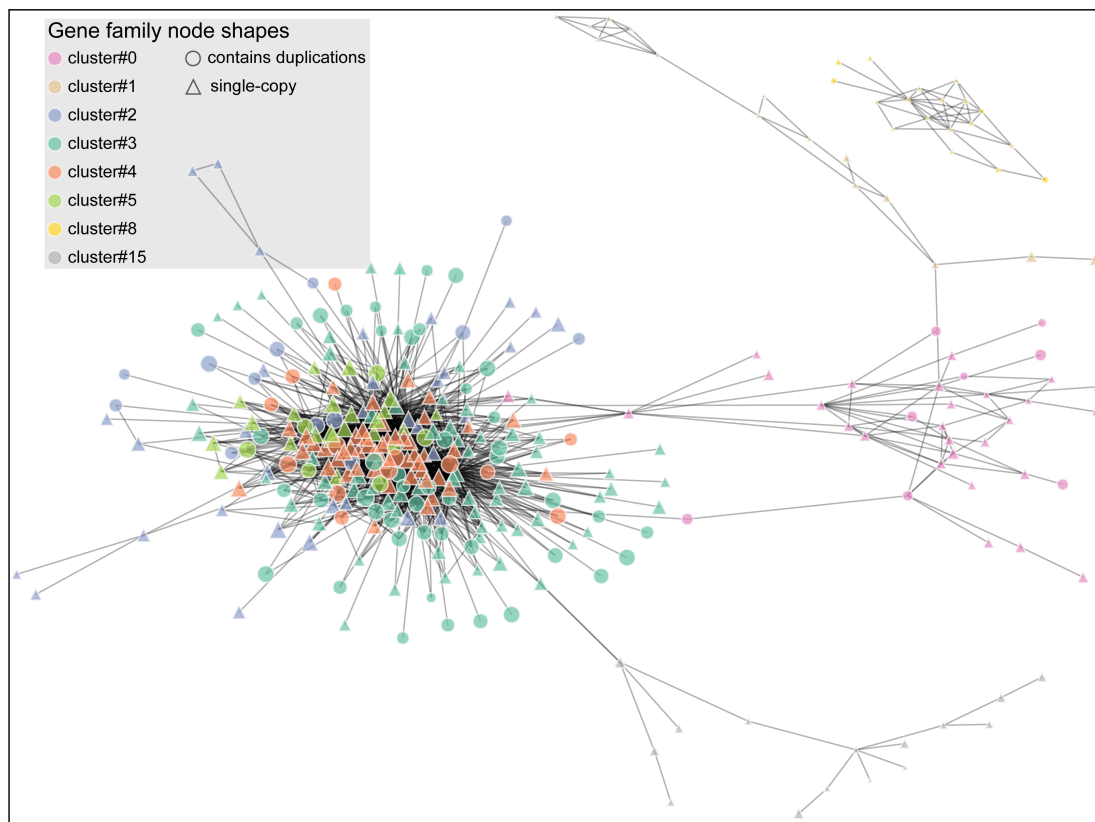
1 57.45%), as variations in bootstrapped alignments caused small perturbations in reconstructed  
2 tree topologies, and subsequent underestimation of evolutionary similarity between bootstrap  
3 replicates (an average of 18% incompatible bipartitions). While more prone to saturation than RF  
4 and more with variance than  $I_{ES}$  estimates, geodesic distance displayed an acceptable degree of  
5 variation within its estimates (CV = 14.97%). Geodesic distances yielded an average of 1.25  
6 between bootstrap replicates, which cannot be directly translated to a proportion of incompatible  
7 bipartitions between trees.

8

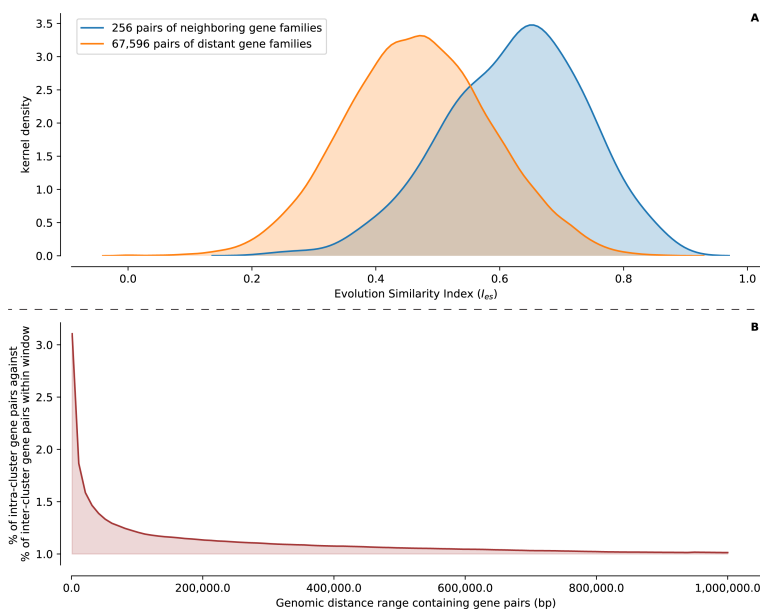
### 9 Evolutionary similarities within archaeal gene families

10 In order to test  $I_{ES}$  performance when estimating shared evolution in an empirical set of  
11 gene families, we evaluated 1,322 families of homologous proteins assembled from annotated  
12 CDSs extracted from 42 archaeal genomes (Supplementary Table S1). This empirical dataset  
13 contains conserved and accessory gene families with different sizes due to gene losses,  
14 duplications, and transfers.

15  $I_{ES}$  was estimated for all pairwise combinations of gene families present in at least 10  
16 genomes, with 2,142 out of 748,712 comparisons having  $I_{ES}$  values of at least 0.7. Pairs of gene  
17 families with an  $I_{ES} \geq 0.7$  were added as nodes to a weighted network with its estimated  $I_{ES}$   
18 value as an edge connecting both gene families. In total 419 unique archaeal gene families were  
19 added to the network, while the remaining 908 gene families did not display any  $I_{ES} \geq 0.7$  with  
20 other gene families. The resulting evolutionary similarity network (Fig. 3) is heavily imbalanced,  
21 with just 11% of nodes involved in 50% of network edges and the majority of gene families,  
22 68%, did not display  $I_{ES}$  above the 0.7 threshold with other gene families, suggesting a general  
23 incompatibility of phylogenetic signal, or lack thereof to detect its compatibility with others.  
24 However, the high edge concentration within just a few nodes suggests a strong central  
25 evolutionary backbone (Puigbò et al. 2009) preserved among a few gene families, from which  
26 the evolutionary trajectories of others have diverged. Similarities between evolutionary histories,  
27 as estimated by  $I_{ES}$ , are strongly associated with genomic linkage ( $p = 6.18e^{-75}$  and  $f = 0.86$ ).  
28 Gene families frequently occurring in each other's genomic vicinity (i.e., fewer than 10,000 bp  
29 apart in at least 21 genomes) displayed significantly greater  $I_{ES}$  relative to pairs of gene families  
30 that were further apart (i.e., more than 100,000 bp apart in at least 21 genomes) (Fig. 4a).



1  
2 Fig. 3 Gene family co-evolution network, each node represents a gene family, and edges connecting nodes represent shared  
3 evolutionary history ( $I_{ES} \geq 0.7$ ). Nodes in the same colors are identified as co-evolving by Louvain community detection.  
4 Triangular nodes represent single-copy genes, and circular ones are gene families containing gene duplications. Clusters of co-  
5 evolving gene families with less than ten members are not represented.



6  
7 Fig. 4 A) distributions of  $I_{ES}$  between pairs of genes within 10,000 bp of each other in blue and between pairs of genes apart by at  
8 least 100,000 bp in orange. Neighboring gene pairs displayed significantly more similar evolutionary histories than non-

1 neighboring gene pairs. B) ratio between the proportion of co-evolving gene pairs and non-co-evolving gene pairs, Y-axis,  
2 occurring within genomic windows, X-axis. 100 window sizes were assessed ranging from 1,000 bp to 1,000,000 bp.

3

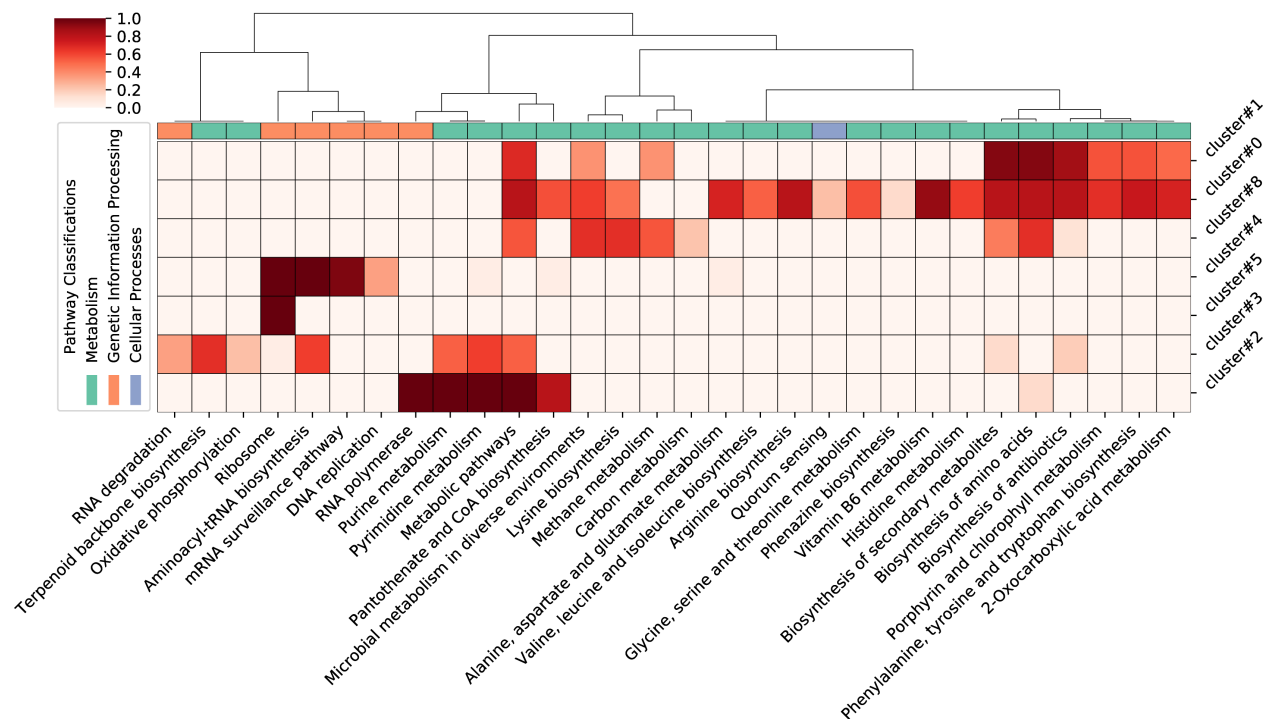
#### 4 Clusters of co-evolving gene families

5 Shared patterns of evolution across gene families were assessed using Louvain  
6 community detection (Blondel et al. 2008), which reported eight major clusters with at least 10  
7 similarly evolving gene families (Fig. 6). Each of these clusters comprise gene families sharing  
8 common evolutionary trends and paths, although the co-evolution assumption based on  $I_{ES}$   
9 estimates and the clustering process is agnostic of specific shared evolutionary events or their  
10 causes. That said, the association between genomic linkage and co-evolution is very pronounced.  
11 Across small nucleotide distances between loci, linkage is a strong predictor of gene co-  
12 evolution, but its predictive power rapidly decreases as the number of nucleotides between two  
13 given loci increase (Fig. 4b), displaying a linear log-log relationship (Supplementary Fig. S2).  
14 Comparisons between intra- and inter-cluster genomic linkage showed that the proportion of co-  
15 evolving genes within 1,000 bp of each other is three times the proportion of non-co-evolving  
16 genes within the same window. Increasing the surveilled genomic window decreases the  
17 difference between proportions; within a 10,000 bp window, the proportion of co-evolving genes  
18 is reduced to 1.8 the proportion of non-co-evolving, and at a 100,000 bp window this difference  
19 in proportions falls to 1.2 (Fig. 4b).

20 Among the eight clusters with ten or more co-evolving gene families four are comprised  
21 of mostly core genes, and four are composed of mostly accessory genes (Fig. 6). The four  
22 clusters of co-evolving core genes (cluster#2, cluster#3, cluster#4, and cluster#5) are promising  
23 candidates for reconstructing the phylogenetic signal of vertical inheritance within Archaea. Core  
24 genes composing these co-evolving clusters are broadly distributed among archaeal clades and  
25 generally occur as single copies within genomes, although can also be found in multiples.  
26 Clusters of co-evolving accessory genes (cluster#0, cluster#1, cluster#8, and cluster#15 in Fig. 6)  
27 do not reflect specific archaeal clades, instead they link polyphyletic clades with biased  
28 distribution caused by HGTs and/or gene losses shared by co-evolving gene families. For  
29 example, cluster#0 is well represented amongst Euryarchaeota and hyperthermophilic TACK;  
30 cluster#15 comprises gene families with shared evolutionary histories mainly occurring within  
31 Crenarchaeota and hyperthermophilic Euryarchaeota; co-evolving accessory gene families in

1 cluster#1 and cluster#8 display congruent signals tying methanogenic Euryarchaeota with  
2 Thaumarchaeota and Asgardarchaeota, respectively.

3



4

5 *Fig. 5* Heatmap of enriched KEGG Pathways, columns, within clusters of co-evolving gene families, rows. Shades of red  
6 represent the proportion of genomes with detected KEGG Pathway enrichment within its homologs of co-evolving gene families.  
7 Columns and rows were clustered using complete linkage and correlation coefficients. KEGG Pathways enriched in less than  
8 10% of genomes in which co-evolving genes occur are not reported. Cluster#15 did not report significant enrichment of KEGG  
9 Pathways.

10 CDSs from 21 out of 42 sampled genomes have functional annotation available in  
11 StringDB (Supplementary Material), and through its API we identified annotated KEGG  
12 Pathways enriched within homologs of co-evolving gene families from each genome. In the  
13 dendrogram and heatmap depicted in Fig. 5 we clearly identify two sets of opposing clusters of  
14 co-evolving gene families: accessories (top three rows) and core (bottom four rows), and their  
15 associations with genetic information processing and metabolism KEGG Pathways (indicated by  
16 column color in the top row). All four clusters of co-evolving core gene families are enriched  
17 with KEGG Pathways related to genetic information processing (e.g., Ribosome, DNA  
18 replication, and Aminoacyl-tRNA biosynthesis). Co-evolving accessory gene families on the  
19 other hand tend to be enriched with KEGG Pathways related to metabolism (e.g., Methane  
20 metabolism, Microbial metabolism in diverse environments, and Biosynthesis of antibiotics in

1 Fig. 5). It is also important to emphasize the opposite pattern of enrichment and depletion of  
2 KEGG Pathways between clusters of core and accessory genes. KEGG Pathways related to  
3 metabolism display minor enrichment signal within core co-evolving gene families, and KEGG  
4 Pathways related to genetic information processing are not enriched within clusters of co-  
5 evolving accessory genes (Fig. 5). Co-evolving accessory genes comprised within cluster#1,  
6 whose occurrence is restricted to methanogenic Euryarchaeota and Thaumarchaeota, are  
7 enriched for methane metabolism within six genomes. Similarly, gene families from in cluster#8,  
8 restricted to methanogenic Euryarchaeota and Asgardarchaeota, are also enriched in methane  
9 metabolism in five genomes (Fig. 5). The biased occurrence of co-evolving gene families within  
10 cluster#1 and cluster#8 towards methanogenic Archaea (Fig. 6) and the enrichment of methane  
11 metabolism within gene families from both clusters (Fig. 5) support a possible origin of these  
12 genes within Euryarchaeota with subsequent independent horizontal transfer to other archaeal  
13 clades.

14 The horizontal exchange of genes commonly accepted as resistant to HGT is exemplified  
15 by the split of extended core-genes in four distinct co-evolving clusters (Fig. 3 and Fig. 7). In  
16 regard to the distribution in extended core-genes among co-evolving clusters, cluster#4 contains  
17 the greatest number of extended core-genes, 44 out of 102. Closeness centrality measures ( $\tilde{C} =$   
18 0.56) and node strength divided by cluster size ( $\tilde{S} = 0.19$ ) also suggest that cluster#4 gene  
19 families have a stronger and cohesive co-evolution signal than gene families from other clusters  
20 (Supplemental Fig S3). In addition, the phylogeny obtained from concatenated cluster#4 gene  
21 families (Fig. 6) is the most similar to the extended core phylogeny among phylogenies from all  
22 co-evolving clusters (Fig. 7).

23 The 44 extended core genes contained within co-evolving cluster#4 are better  
24 representatives of the extended core-genome phylogenetic signal than extended core-genes not  
25 within this cluster ( $p = 3.78e^{-6}$  and  $f = 0.74$ , Supplementary Fig. S4). The great similarity  
26 between cluster#4 and extended core phylogenies (Fig. 6 and Fig. 7) indicate that gene families  
27 comprising cluster#4 are the major contributors to the vertical evolution signal estimated from  
28 the extended core tree. Though cluster#4 comprises only 60.7% of the number gene families  
29 used to reconstruct the extended core tree, it is still able to provide well supported bipartitions  
30 (Fig. 6), corroborating the shared compatible signal within its co-evolving gene families. Such

1 overall compatibility between phylogenetic signals is not likely present within other extended  
2 core genes (Fig. 7), which are scattered across four distinct co-evolving clusters.

3

#### 4 Common and distinct evolutionary trends between co-evolving clusters

5 Among clusters of co-evolving core gene families, cluster#4 and cluster#5 are the most  
6 evenly represented across archaeal groups, while cluster#2 and cluster#3 are poorly distributed  
7 among DPANN (Fig. 6). All four co-evolving clusters have low frequency within  
8 Thaumarchaeota archaeon SCGC AB-539-E09, and only gene families from cluster#2 and  
9 cluster#4 are significantly present in Thermoplasmatales archaeon SCGC AB-539-N05. All four  
10 clusters display very similar overall phylogenies, varying mainly within the organization of  
11 Euryarchaeota (Fig. 6 and Supplementary Material). All four co-evolving clusters reconstructed  
12 the monophyly of Euryarchaeota, with the exception of cluster#2, which placed *Pyrococcus*  
13 *furiosus*, *Thermococcus kodakarensis*, *Methanocaldococcus jannaschii*, *Methanothermobacter*  
14 *thermautotrophicus*, and *Methanopyruus kandleri* together as sister to Asgardarchaeota+TACK.  
15 Only cluster#4 recovered the monophyly of Methanomicrobia as sister to Halobacteria, with the  
16 other three co-evolving clusters placing Halobacteria within Methanomicrobia.

17 All four core co-evolving clusters robustly identified Asgardarchaeota as sister to TACK  
18 (Fig. 6), with small variation in the Asgardarchaeota phylogeny, and cluster#5 placing  
19 Korarchaeota at the base of the TACK super-phylum. When assessing all-versus-all evolutionary  
20 similarities between clusters of co-evolving core genes, the phylogenetic history reconstructed  
21 from cluster#4 is the least dissimilar to the other three (Fig. 7). This shortest path from  
22 cluster#4's evolutionary trajectory to all others corroborates the hypothesis that cluster#4 best  
23 represents the backbone of the vertical inheritance signal, a central point from which others have  
24 diverged (Fig. 7). In general, the overall high  $I_{ES}$  estimates between co-evolving clusters suggest  
25 that despite composing distinct clusters, gene histories between clusters are generally congruent,  
26 with deviations reflecting small divergences potentially representing genes with specific sets of  
27 reticulate histories.

28 Phylogenetic trees obtained from co-evolving accessory gene families in cluster#0,  
29 cluster#1, cluster#8, and cluster#15 reconstructed all represented archaeal phyla as monophyletic  
30 (except for *P. furiosus* in Euryarchaeota in cluster#0, Supplementary Fig. S5), suggesting a  
31 shared common origin of accessory co-evolving genes from each cluster by all genomes from the



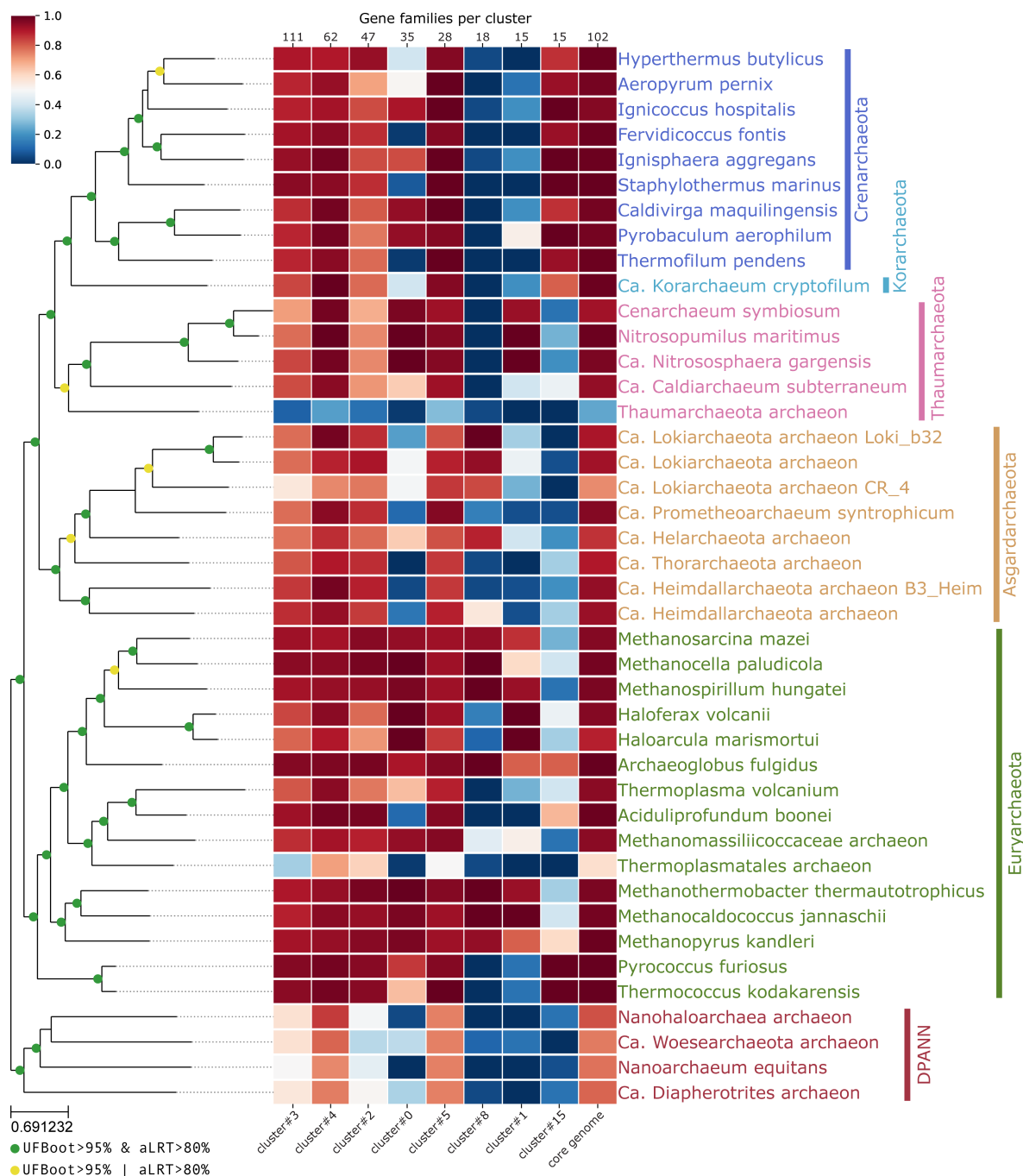
1 same phylum. Although the monophyly of archaeal phyla within trees of co-evolving accessory  
2 genes does not permit an accurate prediction of the directionality of possible inter-phyla HGTs,  
3 intra-phylum distances congruent to the vertical inheritance signal can be used to evaluate the  
4 fitness of inter-phylum distances under a wODR model (Supplemental Fig. S6, S8, S10, and  
5 S11). When compared to pairwise distances expected from vertical inheritance, inter-phylum  
6 distances that significantly differ from estimates obtained by intra-phylum distances may be  
7 attributed to HGT acquisition by one of the phyla in question. For each cluster of co-evolving  
8 accessory genes, we assessed wODR of its pairwise distances against the vertical evolution  
9 estimated from cluster#4.

10         When comparing pairwise distances obtained from cluster#1 against cluster#4, distances  
11 between Euryarchaeota and Thaumarchaeota are consistently placed below the estimated  
12 regression line (Supplementary Fig. S6 and S7). This suggests that cluster#1 genes were  
13 horizontally transferred between ancestors of both phyla, causing shorter evolutionary distances  
14 between phyla than expected if their homologs diverged from the vertical inheritance.

15         Inter-phyla distances between Euryarchaeota and Crenarchaeota obtained from cluster#0  
16 fit the evolutionary rate expected using intra-phylum distances for this co-evolving cluster  
17 (Supplementary Fig. S8), suggesting that homologs from both phyla were vertically inherited  
18 from a common ancestor. On the other hand, cluster#0 inter-phyla distances involving  
19 Thaumarchaeota (Crenarchaeota to Thaumarchaeota and Euryarchaeota to Thaumarchaeota) are  
20 shorter than expected from the wODR using intra-phylum distances (Supplementary Fig. S8) and  
21 display significantly greater residuals than distances between Crenarchaeota and Euryarchaeota  
22 (Supplementary Fig. S9). The absence of cluster#0 genes among Asgardarchaeota and  
23 Korarchaeota and the short inter-phyla distances to Thaumarchaeota homologs suggest an  
24 extensive loss among missing clades and horizontal acquisition by the thaumarchaeal ancestor  
25 from either crenarchaeal or euryarchaeal donors.

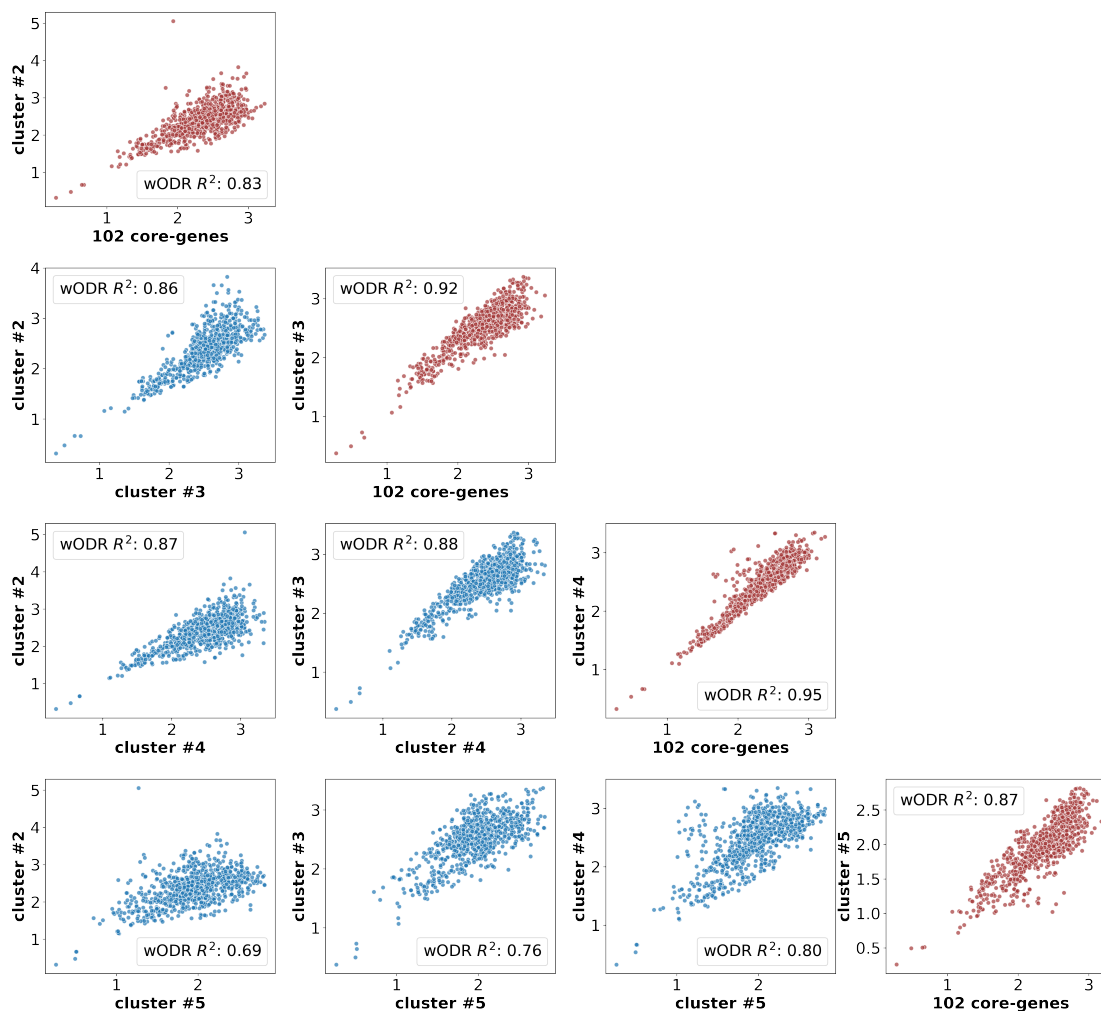
26         Despite the occurrence of accessory genes from cluster#1 and cluster#8 in methanogenic  
27 Euryarchaeota (Fig. 6) and the enrichment of methane metabolism pathways (Fig. 5),  
28 evolutionary histories of both co-evolving clusters are not related (Fig. 3). Co-evolving genes in  
29 cluster#8 did not display  $I_{ES} \geq 0.7$  outside its own cluster, constituting a separate connected  
30 component in the co-evolution network depicted in Fig. 3. That said, cluster#8 gene families  
31 display shorter Euryarchaeota-Asgardarchaeota distances when compared to cluster#4 distances,

1 but unlike cluster#0 and cluster#1, intra-Asgardarchaeota and intra-Euryarchaeota pairwise  
2 distances are not mutually compatible under a single linear regression (Supplemental Fig. S10).  
3 The lack of a strong wODR anchor in the form of intra-phyla distances suggests a more complex  
4 horizontal exchange history of cluster#8 genes, possibly involving intra-phylum HGTs, which  
5 we cannot accurately assess with the dataset used in this study. Cluster#15 co-evolving accessory  
6 genes are well distributed among Crenarchaeota, and its intra-phylum pairwise distances  
7 correspond to cluster#4 distances, but their patchy occurrence among Euryarchaeota and  
8 Korarchaeota (Fig. 6) does not permit a confident assessment of this cluster's evolutionary  
9 history (Supplementary Fig. S5).  
10



1  
2 *Fig. 6* Phylogenetic tree of Archaea reconstructed from 62 co-evolving genes within cluster#4. The phylogeny was obtained  
3 using LG+F+G+C60 evolutionary model from IQTree and each gene had its parameters independently estimated according to  
4 parameter “-sp”. Bipartition supports were estimated using UFBboot and aLRT, each with 1,000 replicates, and bipartitions well  
5 supported by both methods are colored in green (UFBboot  $\geq$  95% and aLRT  $\geq$  80%), while bipartitions well supported by a  
6 single method are colored in yellow. Despite the addition of outgroups to Archaea (e.g. Bacteria) the tree is rooted in DPANN for  
7 the sake of visualization. The associated heatmap reflects the representation of co-evolving gene families amongst archaeal  
8 genomes.

1



2

3 *Fig. 7* Scatter plots of pairwise evolutionary distances reconstructed from each widely distributed gene family versus each other  
4 in blue. And in red, scatter plots of pairwise evolutionary distances reconstructed from each widely distributed gene family versus  
5 pairwise evolutionary distances reconstructed from 102 extended core gene families. Similarities between evolutionary histories  
6 of pairs of co-evolving clusters, and between co-evolving clusters and extended core genes were estimated by  $I_{ES}$ .

## 7 Conclusions

8 The results presented demonstrate the overall accuracy and robustness of  $I_{ES}$  estimates  
9 using both simulated and empirical datasets, as well as comparatively to common tools available  
10 to the community. Among the presented evidence, the clear identification of the strong  
11 association between genomic linkage and gene co-evolution based on  $I_{ES}$  estimates in itself  
12 constitutes an independent evidence of  $I_{ES}$  competence to assess shared evolutionary histories. In  
13 regard to its impact to phylogenomic analysis, reconstructing the vertical inheritance signal from

1 cluster#4 gene families provides a widely applicable improved alternative to analyses limited by  
2 conserved single copy genes.

3           One major consequence of co-evolution between gene families is their co-  
4 occurrence among genomes, given that if genes followed similar evolutionary trajectories, they  
5 are likely to be observed in comparable numbers within the same genomes. Despite similar  
6 performances of Pearson's  $r$  and wODR  $R^2$ , the detection of similar occurrence patterns among  
7 genomes in  $I_{ES}$  constitutes an important step towards more efficient co-evolution inference  
8 between genes. The utilization of wODR also imparts more robust statistical support not directly  
9 available to previous Pearson's  $r$  implementations. The ability to assess residuals of each  
10 datapoint independently also allows for evaluations of specific homologs, a useful tool for HGT  
11 detection. Future applications of  $I_{ES}$  can guide generation of sequence datasets for more accurate  
12 and robust species-tree inference, as well as the detection of significant clusters of gene families  
13 evolving in shared, yet reticulate, patterns.

14           Results presented here support  $I_{ES}$  as a superior proxy for evolutionary similarity between  
15 gene families, outperforming classical tree-based metrics. Comparisons between bootstrapped  
16 replicates within simulated datasets, which should be virtually identical, further corroborates the  
17 robustness of  $I_{ES}$  to phylogenetic uncertainty when compared to RF and geodesic distances.  
18

## 19 Acknowledgements

20 This work was supported by Simons Foundation Collaboration on the Origins of Life Award  
21 #339603 and NSF Integrated Earth Systems Program Award #1615426 to GPF. SMS was  
22 supported through Geisel School of Medicine at Dartmouth's Center for Quantitative Biology  
23 through a grant from the National Institute of General Medical Sciences of the National Institutes  
24 of Health under Award Number P20GM130454.

## 25 References

26 Andam CP, Gogarten JP. 2011. Biased gene transfer and its implications for the concept of  
27 lineage. Biol. Direct [Internet] 6:47. Available from:  
28 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3191353&tool=pmcentrez&ren](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3191353&tool=pmcentrez&rendertype=abstract)  
29 [dertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3191353&tool=pmcentrez&rendertype=abstract)

- 1 Andam CP, Williams D, Gogarten JP. 2010. Biased gene transfer mimics patterns created  
2 through shared ancestry. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 107:10679–10684.  
3 Available from:  
4 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2890805&tool=pmcentrez&ren](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2890805&tool=pmcentrez&rendertype=abstract)  
5 [dertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2890805&tool=pmcentrez&rendertype=abstract)
- 6 Baptiste E, O'Malley M a, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, Lapointe F-  
7 J, Dupré J, Dagan T, Boucher Y, et al. 2009. Prokaryotic evolution and the tree of life are  
8 two different things. *Biol. Direct* [Internet] 4:34. Available from:  
9 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2761302&tool=pmcentrez&ren](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2761302&tool=pmcentrez&rendertype=abstract)  
10 [dertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2761302&tool=pmcentrez&rendertype=abstract)
- 11 Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in  
12 large networks. *J. Stat. Mech. Theory Exp.* [Internet] 2008:P10008. Available from:  
13 [http://stacks.iop.org/1742-](http://stacks.iop.org/1742-5468/2008/i=10/a=P10008?key=crossref.46968f6ec61eb8f907a760be1c5ace52)  
14 [5468/2008/i=10/a=P10008?key=crossref.46968f6ec61eb8f907a760be1c5ace52](http://stacks.iop.org/1742-5468/2008/i=10/a=P10008?key=crossref.46968f6ec61eb8f907a760be1c5ace52)
- 15 Boggs PT, Byrd RH, Schnabel RB. 1987. A Stable and Efficient Algorithm for Nonlinear  
16 Orthogonal Distance Regression. *SIAM J. Sci. Stat. Comput.* [Internet] 8:1052–1078.  
17 Available from: <http://epubs.siam.org/doi/10.1137/0908085>
- 18 Boggs PT, Donaldson JR, Byrd R h., Schnabel RB. 1989. Algorithm 676: ODRPACK: software  
19 for weighted orthogonal distance regression. *ACM Trans. Math. Softw.* [Internet] 15:348–  
20 364. Available from: <http://dl.acm.org/doi/10.1145/76909.76913>
- 21 Bray JR, Curtis JT. 1957. An Ordination of the Upland Forest Communities of Southern  
22 Wisconsin. *Ecol. Monogr.* [Internet] 27:325–349. Available from:  
23 <http://doi.wiley.com/10.2307/1942268>
- 24 Contreras-Moreira B, Vinuesa P. 2013. GET\_HOMOLOGUES, a versatile software package for  
25 scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* [Internet]  
26 79:7696–7701. Available from:  
27 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3837814&tool=pmcentrez&ren](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3837814&tool=pmcentrez&rendertype=abstract)  
28 [dertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3837814&tool=pmcentrez&rendertype=abstract)
- 29 Dagan T, Martin W. 2006. The tree of one percent. *Genome Biol.* 7.
- 30 Fletcher W, Yang Z. 2009. INDELible: A flexible simulator of biological sequence evolution.  
31 *Mol. Biol. Evol.* 26:1879–1888.

- 1 Gertz J, Elfond G, Shustrova A, Weisinger M, Pellegrini M, Cokus S, Rothschild B. 2003.  
2 Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics* 19:2039–  
3 2045.
- 4 Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Co-evolution of proteins with  
5 their interaction partners. *J. Mol. Biol.* 299:283–293.
- 6 Gori K, Suchan T, Alvarez N, Goldman N, Dessimoz C. 2016. Clustering Genes of Common  
7 Evolutionary History. *Mol. Biol. Evol.* [Internet] 33:1590–1605. Available from:  
8 <http://mbe.oxfordjournals.org/content/early/2016/02/17/molbev.msw038.short?rss=1>
- 9 Gueudré T, Baldassi C, Zamparo M, Weigt M, Pagnani A. 2016. Simultaneous identification of  
10 specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc.*  
11 *Natl. Acad. Sci. U. S. A.* [Internet] 113:12186–12191. Available from:  
12 <http://www.ncbi.nlm.nih.gov/pubmed/27729520>
- 13 Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of  
14 Phylogenomic Data. *Mol. Biol. Evol.* [Internet] 33:1635–1638. Available from:  
15 <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw046>
- 16 Izarzugaza JMG, Juan D, Pons C, Pazos F, Valencia A. 2008. Enhancing the prediction of  
17 protein pairings between interacting families using orthology information. *BMC*  
18 *Bioinformatics* [Internet] 9:35. Available from:  
19 <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-35>
- 20 Kimmel R, Sethian J a. 1998. Computing geodesic paths on manifolds. *Proc. Natl. Acad. Sci. U.*  
21 *S. A.* [Internet] 95:8431–8435. Available from:  
22 <http://www.ncbi.nlm.nih.gov/pubmed/9671694>
- 23 Koonin E V, Wolf YI, Puigbò P. 2009. The phylogenetic forest and the quest for the elusive tree  
24 of life. *Cold Spring Harb. Symp. Quant. Biol.* [Internet] 74:205–213. Available from:  
25 <http://www.ncbi.nlm.nih.gov/pubmed/19687142>
- 26 Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. 2005. The net of life: reconstructing the  
27 microbial phylogenetic network. *Genome Res.* [Internet] 15:954–959. Available from:  
28 <http://www.ncbi.nlm.nih.gov/pubmed/15965028>
- 29 Kupczok A, von Haeseler A, Klaere S. 2008. An exact algorithm for the geodesic distance  
30 between phylogenetic trees. *J. Comput. Biol.* [Internet] 15:577–591. Available from:  
31 <http://www.ncbi.nlm.nih.gov/pubmed/18631022>

- 1 Leigh JW, Susko E, Baumgartner M, Roger AJ. 2008. Testing Congruence in Phylogenomic  
2 Analysis. *Syst. Biol.* [Internet] 57:104–115. Available from:  
3 <http://sysbio.oxfordjournals.org/cgi/doi/10.1080/10635150801910436>
- 4 Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic  
5 genomes. *Genome Res.* [Internet] 13:2178–2189. Available from:  
6 <http://www.ncbi.nlm.nih.gov/pubmed/12952885>
- 7 McGraw KO, Wong SP. 1992. A common language effect size statistic. *Psychol. Bull.* [Internet]  
8 111:361–365. Available from: [http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-](http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.111.2.361)  
9 2909.111.2.361
- 10 McKinney W. 2010. Data Structures for Statistical Computing in Python. In: p. 56–61. Available  
11 from: <https://conference.scipy.org/proceedings/scipy2010/mckinney.html>
- 12 Mirarab S, Bayzid MS, Bossau B, Warnow T. 2014. Statistical binning improves species tree  
13 estimation in the presence of gene tree heterogeneity. *Science* (80-. ). [Internet] 346.  
14 Available from: <http://dx.doi.org/10.1126/science.1250463>
- 15 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective  
16 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*  
17 [Internet] 32:268–274. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25371430>
- 18 Novichkov PS, Omelchenko M V, Gelfand MS, Mironov A a, Wolf YI, Koonin E V. 2004.  
19 Genome-wide molecular clock and horizontal gene ransfer in bacterial evolution. *J.*  
20 *Bacteriol.* [Internet] 186:6575–6585. Available from:  
21 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=516599&tool=pmcentrez&rend](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=516599&tool=pmcentrez&rendertype=abstract)  
22 [ertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=516599&tool=pmcentrez&rendertype=abstract)
- 23 Owen M, Provan JS. 2011. A Fast Algorithm for Computing Geodesic Distances in Tree Space.  
24 *IEEE/ACM Trans. Comput. Biol. Bioinforma.* [Internet] 8:2–13. Available from:  
25 <http://ieeexplore.ieee.org/document/5396323/>
- 26 Papke RT, Gogarten JP. 2012. Ecology. How bacterial lineages emerge. *Science* [Internet]  
27 336:45–46. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22491845>
- 28 Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicator of protein-protein  
29 interaction. *Protein Eng.* 14:609–614.
- 30 Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Curr. Opin.*  
31 *Microbiol.* [Internet] 14:615–623. Available from:



- 1 <http://www.ncbi.nlm.nih.gov/pubmed/21856213>
- 2 Puigbò P, Wolf YI, Koonin E V. 2009. Search for a “Tree of Life” in the thicket of the  
3 phylogenetic forest. *J. Biol.* [Internet] 8:59. Available from:  
4 <http://www.ncbi.nlm.nih.gov/pubmed/19594957>
- 5 Ramani AK, Marcotte EM. 2003. Exploiting the co-evolution of interacting proteins to discover  
6 interaction specificity. *J. Mol. Biol.* 327:273–284.
- 7 Rangel LT, Marden J, Colston S, Setubal JC, Graf J, Gogarten JP. 2019. Identification and  
8 characterization of putative *Aeromonas* spp. T3SS effectors. Antunes LCM, editor. *PLoS*  
9 *One* [Internet] 14:e0214035. Available from:  
10 <https://dx.plos.org/10.1371/journal.pone.0214035>
- 11 Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math. Biosci.* [Internet]  
12 53:131–147. Available from: <http://linkinghub.elsevier.com/retrieve/pii/0025556481900432>
- 13 Seabold S, Perktold J. 2010. statsmodels: Econometric and statistical modeling with python. In:  
14 9th Python in Science Conference.
- 15 Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, Alm EJ.  
16 2012. Population genomics of early events in the ecological differentiation of bacteria.  
17 *Science* [Internet] 336:48–51. Available from:  
18 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3337212&tool=pmcentrez&ren](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3337212&tool=pmcentrez&rendertype=abstract)  
19 [dertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3337212&tool=pmcentrez&rendertype=abstract)
- 20 Sorek R, Zhu Y, Creevey C, Francino M. 2007. Genome-wide experimental determination of  
21 barriers to horizontal gene transfer. *Science* (80-. ). [Internet] 318:1449–1452. Available  
22 from: <http://www.sciencemag.org/content/318/5855/1449.short>
- 23 Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with  
24 thousands of taxa and mixed models. *Bioinformatics* [Internet] 22:2688–2690. Available  
25 from: <http://www.ncbi.nlm.nih.gov/pubmed/16928733>
- 26 Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva  
27 NT, Morris JH, Bork P, et al. 2019. STRING v11: protein–protein association networks  
28 with increased coverage, supporting functional discovery in genome-wide experimental  
29 datasets. *Nucleic Acids Res.* [Internet] 47:D607–D613. Available from:  
30 <https://academic.oup.com/nar/article/47/D1/D607/5198476>
- 31 Szöllosi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013. Efficient exploration of the

- 1 space of reconciled gene trees. *Syst. Biol.* 62:901–912.
- 2 Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer  
3 between bacteria. *Nat. Rev. Microbiol.* [Internet] 3:711–721. Available from:  
4 <http://dx.doi.org/10.1038/nrmicro1234>
- 5 Vinuesa P, Contreras-Moreira B. 2015. Robust Identification of Orthologues and Paralogues for  
6 Microbial Pan-Genomics Using GET\_HOMOLOGUES: A Case Study of pIncA/C  
7 Plasmids. In: p. 203–232. Available from: [http://link.springer.com/10.1007/978-1-4939-](http://link.springer.com/10.1007/978-1-4939-1720-4_14)  
8 [1720-4\\_14](http://link.springer.com/10.1007/978-1-4939-1720-4_14)
- 9 Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E,  
10 Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for  
11 scientific computing in Python. *Nat. Methods* [Internet] 17:261–272. Available from:  
12 <http://www.nature.com/articles/s41592-019-0686-2>  
13