Current Biology, Volume *33*

Supplemental Information

Remote homolog detection places insect chemoreceptors in a cryptic protein superfamily spanning the tree of life

Nathaniel J. Himmel, David Moi, and Richard Benton

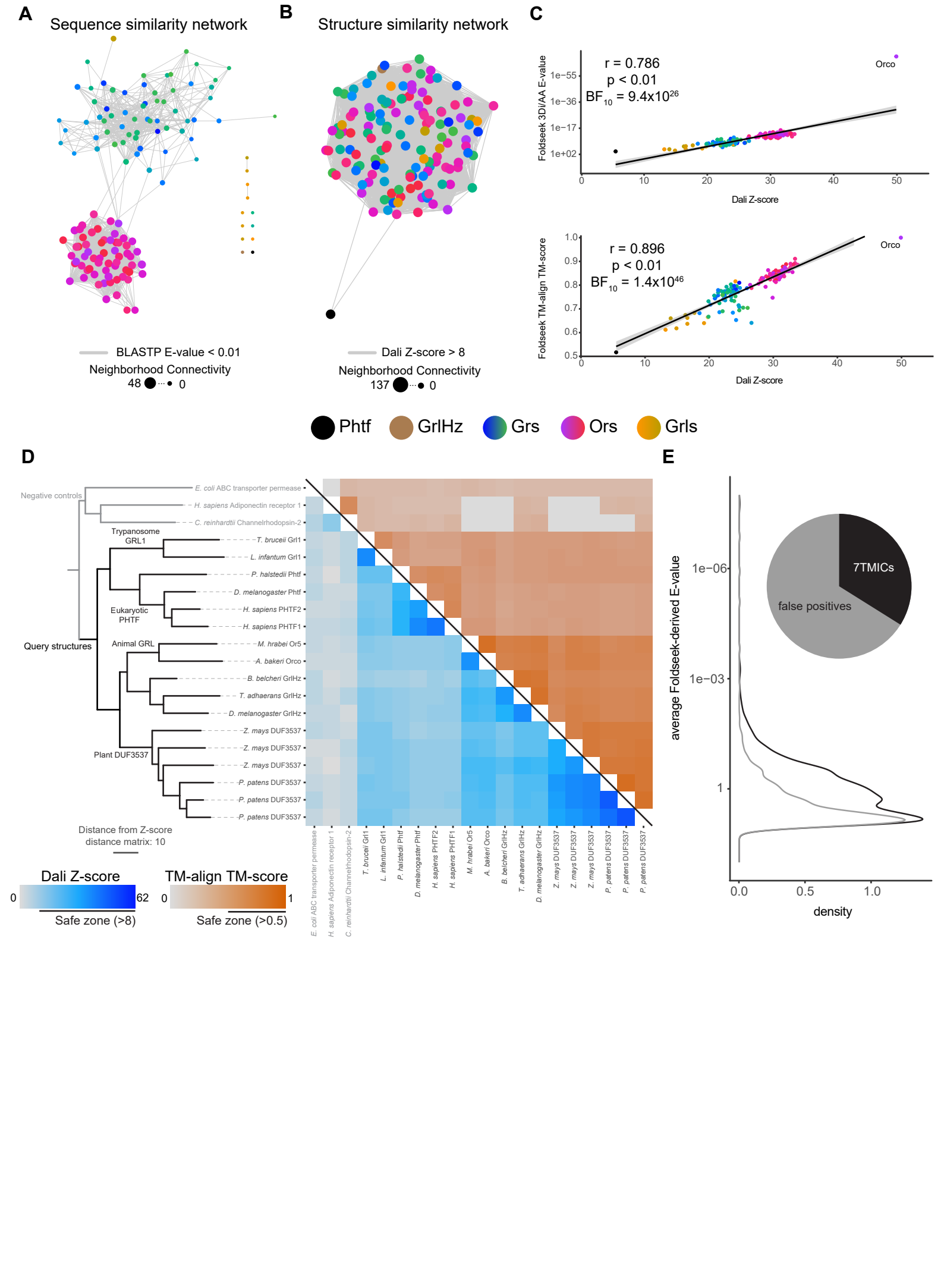
A
Sequence similarity network
BLASTP E-value < 0.01
Neighborhood Connectivity
48 0
B
Structure similarity network
Dali Z-score > 8
Neighborhood Connectivity
137 0
C
r = 0.786
p < 0.01
BF10 = 9.4x10^26
Foldseek 3Di/AA E-value
Dali Z-score
Orco
r = 0.896
p < 0.01
BF10 = 1.4x10^46
Foldseek TM-align TM-score
Dali Z-score
Orco
Phtf
GrlHz
Grs
Ors
Grls
D
Negative controls
E. coli ABC transporter permease
H. sapiens Adiponectin receptor 1
C. reinhardtii Channelrhodopsin-2
Trypanosome GRL1
T. bruceii Grl1
L. infantum Grl1
P. halstedii Phtf
D. melanogaster Phtf
Eukaryotic PHTF
H. sapiens PHTF2
H. sapiens PHTF1
Query structures
Animal GRL
M. hrabei Or5
A. bakeri Orco
B. belcheri GrlHz
T. adhaerans GrlHz
D. melanogaster GrlHz
Plant DUF3537
Z. mays DUF3537
Z. mays DUF3537
Z. mays DUF3537
P. patens DUF3537
P. patens DUF3537
P. patens DUF3537
Distance from Z-score distance matrix: 10
Dali Z-score
0
62
Safe zone (>8)
TM-align TM-score
0
1
Safe zone (>0.5)
E
average Foldseek-derived E-value
1e−06
1e−03
1
7TMICs
false positives
0.0
0.5
1.0
density

Figure S1. All-to-all pairwise protein similarity networks of *D. melanogaster* 7TMICs, Foldseek benchmarking, and summary of the Foldseek screen, related to Figure 1 and Figure 2.

(**A**) All-to-all BLASTP network of *D. melanogaster* 7TMICs; consistent with their low pairwise sequence similarity, this analysis fails to link every 7TMIC to all others. Rather, the major *D. melanogaster* classes (Ors and Grs) are separated into two identifiable community structures, with sparse connectivity among the Grs, and between the Grs and Ors. Other 7TMICs—including Grls, GrlHz, Phtf and two Grs—form singlets, indicating an inability to identify hypothetical homologs using BLASTP. The color key is shown below panels (A-C), and matches that of **Figure 1E**.

(**B**) All-to-all Dali network of *D. melanogaster* 7TMICs. In contrast to (A), structural comparisons result in a "hairball" network, where nearly all proteins are linked to all others, excepting Phtf, which is presumed to be the most distantly related.

(**C**) Plots of structural similarity scores between Orco and other *D. melanogaster* 7TMICs, comparing Dali to Foldseek-derived scores. Foldseek generates Orco-to-all E-values that tightly correlate with the rapidly generated 3Di+AA-derived E-values (top) and the slowly generated TM-align-derived TM-scores (bottom).

(**D**) Protein models used in the Foldseek screen, and negative controls used for subsequent Dali-based validation, with a clustering dendrogram based on all-to-all Dali comparisons between the queries and negative controls. The dendrogram is derived from the Dali Z-score distance matrix. The annotation on the heatmap corresponds to the "groups" described in the methods. The heatmap shows all-to-all Dali Z-scores and TM-scores. All 7TMIC-to-control comparisons are well below thresholds of confidence in fold similarity: the Dali Z-score maximum is 7.1 and averages 4.3; and the TM-align TM-score maximum is 0.31 and averages 0.2. By contrast, for 7TMIC-to-7TMIC comparisons: the Z-score minimum is 8.5 and averages 22.0, and the TM-score minimum is 0.4 and averages 0.6.

(**E**) Stacked density plot showing the frequency distribution of the hits of the Foldseek screen, by E-value, with the inset pie-chart showing the proportion of true positives to false positives. Many true positives had relatively poor E-values, with similar or worse scores than many false positives, demonstrating the need for structural validation in a Foldseek screen.

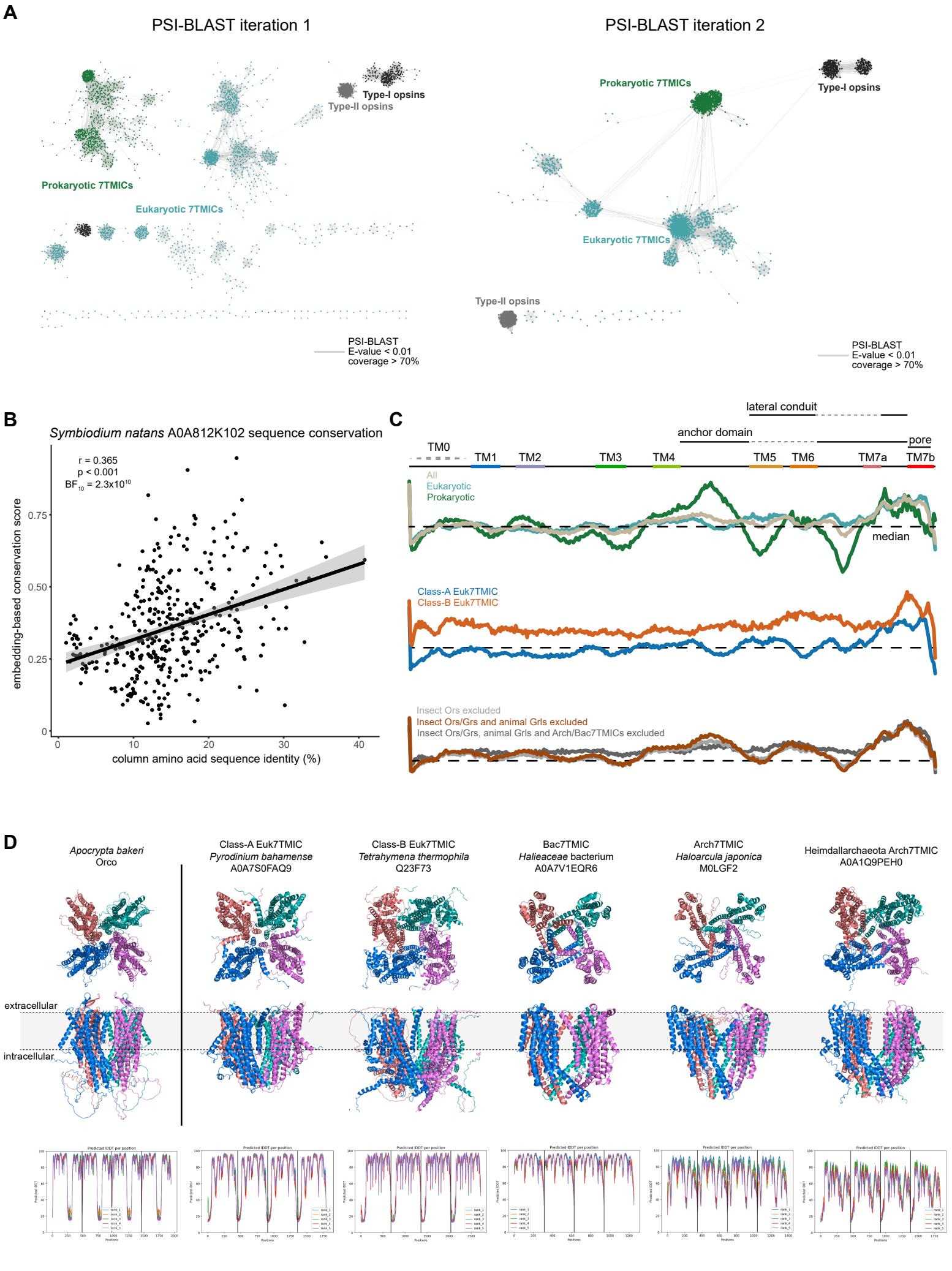
A
PSI-BLAST iteration 1
Type-I opsins
Type-II opsins
Prokaryotic 7TMICs
Eukaryotic 7TMICs
PSI-BLAST
E-value < 0.01
coverage > 70%
PSI-BLAST iteration 2
Prokaryotic 7TMICs
Type-I opsins
Eukaryotic 7TMICs
Type-II opsins
PSI-BLAST
E-value < 0.01
coverage > 70%
B
Symbiodium natans A0A812K102 sequence conservation
r = 0.365
p < 0.001
BF10 = 2.3x10^10
embedding-based conservation score
column amino acid sequence identity (%)
C
lateral conduit
anchor domain
pore
TM0
TM1
TM2
TM3
TM4
TM5
TM6
TM7a
TM7b
All
Eukaryotic
Prokaryotic
median
Class-A Euk7TMIC
Class-B Euk7TMIC
Insect Ors excluded
Insect Ors/Grs and animal Grls excluded
Insect Ors/Grs, animal Grls and Arch/Bac7TMICs excluded
D
Apocrypta bakeri Orco
Class-A Euk7TMIC Pyrodinium bahamense A0A7S0FAQ9
Class-B Euk7TMIC Tetrahymena thermophila Q23F73
Bac7TMIC Halieaceae bacterium A0A7V1EQR6
Arch7TMIC Haloarcula japonica M0LGF2
Heimdallarchaeota Arch7TMIC A0A1Q9PEH0
extracellular
intracellular

Figure S2. Initial iterations of the PSI-BLAST sequence similarity networks, 7TMIC sequence conservation analysis, and predicted quaternary structures of select, newly-identified 7TMICs, related to Figure 3.

(**A**) Sequence similarity networks were generated by all-to-all PSI-BLAST searches of a 50% clustered sequence database of 7TMICs, alongside databases of Type-I and Type-II opsins. Iterations 1 and 2 are visualized here. Subsequent iterations resemble the clustering pattern of iteration 3, as visualized in **Figure 3B**, albeit with strengthening community structures. Left: PSI-BLAST iteration 1. In this network, sequences formed several non-contiguous clusters, and failed to cluster together 7TMICs and Type-I opsins, which is expected given the substantial sequence dissimilarity of 7TMICs. Right: PSI-BLAST iteration 2. Surprisingly, PSI-BLAST networking produced bidirectional linking of the majority of 7TMICs, although presumed spurious linkages to outgroups began to form (which did not greatly multiply in subsequent iterations), and a small number of 7TMICs do not form links to the core 7TMIC cluster(s) (although all join a 7TMIC community structure by iteration 3 [**Figure 3B**]).

(**B**) Embedding-based conservation scores weakly but significantly correlate with column sequence identity from the A0A812K102-centered sequence alignment.

(**C**) Average embedding-based conservation scores for different subsets of 7TMICs, demonstrating that, while family-specific patterns exist, the conservation of anchor domain and pore regions is consistent. The TM and domain labels are derived from A0A812K102, as visualized in **Figure 3**.

(**D**) Predicted tetramers for select 7TMICs. Top: top (presumed extracellular) and side views of the tetrameric arrangement of 7TMICs predicted by AlphaFold-Multimer, showing the formation of a hypothetical pore along TM7b, similar to *A. bakeri* Orco (far-left). Bottom: local Distance Difference Test (lDDT) scores (used to assess model confidence), plotted for each of the 5 replicate models generated. Each color represents a different replicate; vertical black lines separate each of the modelled subunits. Generally, the transmembrane-spanning alpha helices are the most confidently predicted, leading to the similar patten of lDDT peaks and troughs across models.

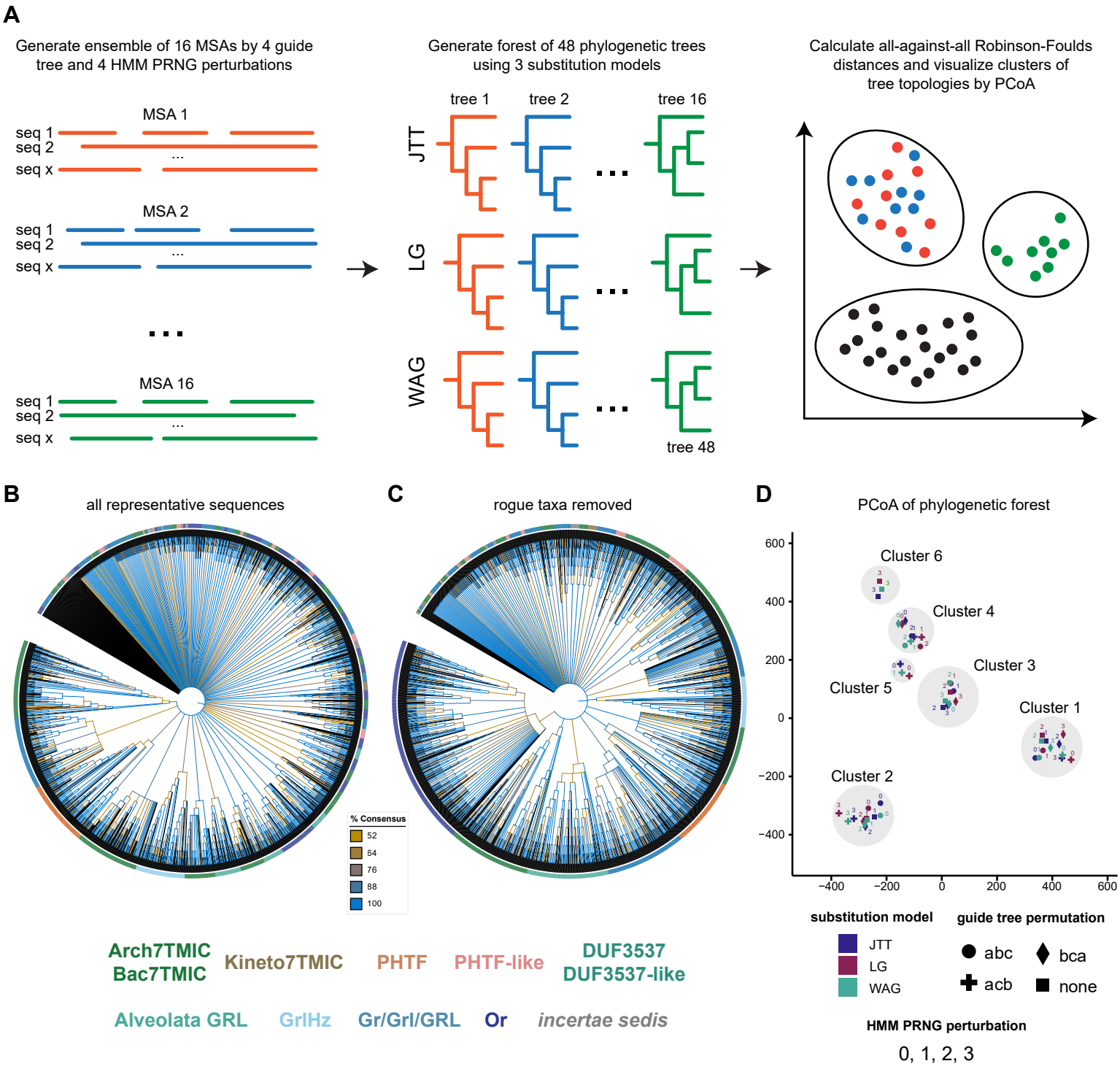
A
Generate ensemble of 16 MSAs by 4 guide tree and 4 HMM PRNG perturbations
MSA 1
seq 1
seq 2
...
seq x
MSA 2
seq 1
seq 2
...
seq x
MSA 16
seq 1
seq 2
...
seq x
Generate forest of 48 phylogenetic trees using 3 substitution models
tree 1
tree 2
tree 16
JTT
LG
WAG
tree 48
Calculate all-against-all Robinson-Foulds distances and visualize clusters of tree topologies by PCoA
B
all representative sequences
C
rogue taxa removed
% Consensus
52
64
76
88
100
D
PCoA of phylogenetic forest
Cluster 6
Cluster 4
Cluster 3
Cluster 5
Cluster 1
Cluster 2
substitution model
JTT
LG
WAG
guide tree permutation
abc
bca
acb
none
HMM PRNG perturbation
0, 1, 2, 3
Arch7TMIC
Bac7TMIC
Kineto7TMIC
PHTF
PHTF-like
DUF3537
DUF3537-like
Alveolata GRL
GrlHz
Gr/Grl/GRL
Or
incertae sedis
E
1
2
3
4
5
6
F
branch length: 0.1
G
branch length: 0.1

Figure S3. Phylogenetic and tree space analysis, related to Figure 4.
(**A**) Pipeline for sequence-based phylogenetic analysis. First, an ensemble of 16 multiple sequence alignments (MSAs) are made by perturbating the guide tree and the Hidden Markov model's pseudorandom number generator (HMM PRNG). Second, phylogenetic trees are generated for each of the MSAs, using 3 different amino acid substitution models, resulting in 48 trees. Finally, differences in the topology of the 48 trees are calculated by pairwise Robinson-Foulds distances; the resulting distance matrix is subsequently visualized in two dimensions by principal coordinate analysis (PCoA).
(**B**) Majority consensus tree for the 48 phylogenetic trees based on alignments of the representative 7TMIC sequences. 7TMIC clades/colors were assigned manually based on visual inspection of a CLANS-based clustering analysis; the color key for 7TMIC subfamilies is shown below panels (B-C). Branch colors indicate the percent consensus. There is essentially no clear consensus among these 48 initial trees; obviously monophyletic clades—such as insect Ors—are not reliably predicted, suggesting substantial alignment/phylogenetic errors (as expected for this highly divergent superfamily).
(**C**) Majority consensus tree for the 48 phylogenetic trees based on alignments of a 7TMIC dataset where rogue taxa (i.e. the most phylogenetically unstable leaves) have been removed. While there is still no greatly informative majority consensus topology, this analysis better recapitulates more obvious monophyletic clades, with higher branch consensus, indicating that errors have been minimized (but not eliminated, which we did not expect to occur at these levels of sequence dissimilarity).
(**D**) PCoA of Robinson-Foulds tree space for trees from (C). Trees form 6 topology clusters.
(**E**) Majority consensus trees for each of the 6 clusters, with colors matching (B) and (C). Five of these clusters agree that Kineto7TMICs branch proximally to prokaryotic 7TMICs, consistent with the hypothesis that kinetoplastids (and allies: Discoba) split early in eukaryotic evolution. Clusters 1 and 4 do not have majority consensus on deep 7TMIC branching. The remaining clusters suggest there are at least two Euk7TMIC families, termed Class-A and Class-B Euk7TMICs, but do not agree on the monophyly of Class-B Euk7TMICs. Clusters 4-6 suggest Class-B monophyly, while clusters 1-3 suggest that many proteins are basally branching (and thus, paraphyly). Given that structure-based phylogenetics suggest a monophyletic Class-B, this discordance may be the result of lingering long branch attraction or other errors resulting from the inclusion of rapidly evolved, horizontally-transferred and/or structurally-convergent proteins.
(**F**) Structural phylogeny derived from pairwise distances used the Foldseek 3Di structural alphabet, with colors matching the panels above. This tree is presented as rooted, but as in Figure 4, the true root is likely within the prokaryotic 7TMICs, at the location of the Last Universal Common Ancestor.
(**G**) Structural phylogeny derived from pairwise lDDT scores, with colors matching the panels above. As in (F), the true root is likely at location of the Last Universal Common Ancestor.

	MSA Perturbations		All Sequences		No Rogue Taxa	
MSA Replicate	**Guide Tree**	**PRNG**	**Columns**	**Column Confidence**	**Columns**	**Column Confidence**
1	abc	0	38154	0.315	36865	0.311
2	abc	1	38357	0.297	34833	0.295
3	abc	2	36970	0.314	35376	0.325
4	abc	3	37064	0.314	31247	0.304
5	acb	0	38375	0.31	35627	0.304
6	acb	1	40342	0.298	34754	0.302
7	acb	2	37080	0.317	34994	0.321
8	acb	3	35735	0.313	31345	0.305
9	bca	0	38391	0.315	35914	0.314
10	bca	1	39831	0.3	35813	0.294
11	bca	2	37647	0.316	34847	0.322
12	bca	3	36515	0.305	31406	0.303
13	none	0	38700	0.3	35842	0.309
14	none	1	40114	0.306	34808	0.293
15	none	2	37443	0.308	34932	0.309
16	none	3	37831	0.309	32689	0.298

Table S1. Muscle5 multiple sequence alignment analysis, related to Figure 4.
Column confidence is a measure of the reproducibility of each column, where 0 indicates the column is never found and 1 indicates it is found across all alignments. Dispersion is measured as the median dispersion of aligned letter pairs over the ensemble (D_LP), and the median dispersion of columns over the ensemble (D_Cols) (Robert Edgar, personal communication, 10 May 2023), where 0 is all the same and 1 is all different. Dispersion was extremely high. For the initial set of alignments: D_LP=0.5836 D_Cols=1.0000. After removal of rogue taxa: D_LP=0.5855 D_Cols=1.0000.