Supporting Material

The deep archaeal roots of eukaryotes

Natalya Yutin, Kira S. Makarova, Sergey L. Mekhedov, Yuri I. Wolf, Eugene V. Koonin*

Supplemental Methods

Identification of the core of a protein cluster

Proteins within a cluster of apparent orthologs often exhibit substantial diversity of domain architectures, especially, in eukaryotes. We designed a procedure to extract the maximum region of shared sequence similarity and construct a "core" alignment representing the cluster. This procedure was applied to all eukaryotic, archaeal and bacterial clusters of putative orthologs analyzed in this work.

Step 1: approximate consensus PSSM. The median length of the sequences comprising a cluster was calculated; sequences shorter than x0.5 or longer than x10 of the median were discarded. The remaining sequences were clustered using the NCBI BLASTCLUST program; BLASTCLUST parameters were iteratively adjusted within the preset limits (from "-S 90 -L 0.85" to "-S 0.3 -L 0.5") aiming to obtain 25 subclusters. Sequences with the lengths closest to the median length of their respective subclusters were selected to represent the subcluster.

The selected representatives were aligned using the MUSCLE program (Edgar 2004). Each alignment column was assigned a homogeneity value by scaling the sum-of-pairs score within the column between those of a homogeneous column (the same residue in all aligned sequences) and a random column (YIW, I. A. Seledtsov IA, KSM, unpublished). Columns with homogeneity of less than

0.2 and/or with more than one-third of gap characters were removed from the alignment. A consensus sequence was created from the alignment. The alignment and its consensus were used as the source of the PSI-BLAST PSSM and the query sequence, respectively (Altschul et al. 1997).

Step 2: mapping the approximate consensus to the original sequences. The alignment and its consensus, obtained at Step 1, were used in a PSI-BLAST search against the database consisting of the original cluster members. Proteins that did not show significant similarity to the query (*e*-value >0.01) were removed. If the unaligned N- and C-terminal fragments of the consensus were shorter than 1/3 of the consensus sequence length, the protein sequence was padded with an appropriate number of residues to match the full length of the consensus sequence. Median length of the protein fragments that showed significant similarity to the core PSSM was computed, and sequences shorter than 1/3 of the median were removed.

Interdomain BLAST searches

For each cluster, core protein sequences were aligned using the MUSCLE program (Edgar 2004). As an additional refining step, all sequences were compared to the alignment consensus in positions with the alignment column homogeneity >0.4. Sequences with similarity to the consensus less than x0.3 of the similarity of consensus to itself were removed from the alignment. After this procedure, columns with homogeneity <0.2 and/or with more than one-third of gap characters were removed. A new consensus sequence was created and added back to the refined alignment.

The alignment with its consensus a the query were used for single-pass PSI-BLAST search (Altschul et al. 1997): eukaryotic PSSMs were run against archaeal or bacterial databases, and the respective reciprocal searches were performed. For each cluster in the target database the score against the query cluster was determined as the average score of the cluster members. Target clusters were ranked according to these scores. Reciprocal best hits between eukaryotic and archaeal and eukaryotic and bacterial clusters were recorded.

Phylogenetic analysis

First-round phylogenetic analysis. Initial phylogenic analysis and the choice of the best (for each alignment) amino-acid substitution matrix were performed using the PhyML software (Guindon and Gascuel 2003). Each alignment was run with 8 substitution models offered by PhyML for amino-acid sequences: JTT (Jones, Taylor, and Thornton 1992), Dayhoff (Dayhoff, Schwartz, and Orcutt 1978), WAG (Whelan and Goldman 2001), DCMut (Kosiol and Goldman 2005), RtREV (Dimmic et al. 2002), CpREV (Adachi et al. 2000) VT (Muller and Vingron 2000), and Blosum62 (Henikoff and Henikoff 1992). Two mitochodrial models (mtREV and MtMam) also available in PhyML were not used. For each of 8 substitution models, a Maximum Likelihood tree was constructed using with following parameters: number of relative substitution rate categories was 4; the proportion of invariable sites and alpha (gamma distribution parameter) were adjustable (estimated). The best tree was chosen by maximum log-likelihood of eight trees. The substitution model used for the best tree was chosen for the next round of phylogenetic analysis.

Selection of representatives for the second-round Maximum Likelihood phylogenetic analysis. Maximum-Likelihood (ML) phylogenetic analysis is a very computationally-intensive procedure with resource usage critically dependent on the number of sequences. Thus, reduction of the number of sequences is desirable. Moreover, with the horizontal gene transfer rampant in the prokaryotic world and, possibly, in unicellular eukaryotes as well (Doolittle 1999), not all sequences are equally suited to represent their taxonomic group (obviously, a gene in a bacterial genome that has recently been acquired from archaea cannot be a proper representative of the respective bacterial branch). Ideally, it is desirable to select a "maximally diverse" subset of "typical" sequences, representing the "native phylogenetic position" of a given clade. Operationally, we aim to approximate these goals using the following procedure.

A PhyML-tree was midpoint-rooted using the RETREE program from the PHYLIP package (Felsenstein 1996). Each terminal node (leaf) of the rooted tree was labeled with one of the four taxonomic labels: Crenarchaeota, Euryarchaeota, Eukaryota, or Bacteria (hereinafter CA, EA, E and B; *Nanoarchaeon equitans* was artificially placed within Euryarchaeota for the purpose of this analysis).

Each leaf was given a weight equal to the square root of the number of individual genomes in the subcluster (as the leaves in the PhyML tree are, generally, representatives of similarity-based subclusters). In a leaves-to-root pass over the tree, each internal node was assigned 4 weights (for CA, EA, E and B) which represented the sum of the corresponding weights of the descendant nodes. Each of these weights was then normalized by the taxon-specific sum across the tree. Note that by design of this procedure the root node acquires the weight of 1.0 for each of the four taxa. Each internal node was then formally labeled with a taxon that had the maximum taxon-specific weight.

For each node of the tree, two indices, "representativeness" (R) and "purity" (P), - were computed for the titular taxon. R is simply the taxon-specific weight, i.e., the fraction of all taxon members that descend from this node. P reflects the number of representatives of other taxa descend from this node and is computed as one minus the average weight of other taxa at this node. Obviously, at the leaves P=1 (because a terminal node is assigned to a single taxon); the same holds for all subtrees where leaves are taxonomically homogeneous. In an ideal tree for each taxon, there exists a node with R=1 and P=1 (i.e. all members of a taxon are monophyletic on a taxonomically homogeneous subtree). In the real trees, each node was assigned a "quality" index $Q=RP^2$, and for each taxon, the node with the maximum value of Q that is not a parent of a selected node for another taxon with a higher Q index was identified. The 4 selected nodes (one for each of CA, EA, E and B) represented the "largest nearly homogeneous" clades for the respective taxa and served as the sources of the representatives.

For the purpose of selection of representatives, the four representative nodes were treated as roots of the corresponding independent (sub)trees. Leaves from the non-titular taxa were removed and root-to-leaf as well as leaf-to-leaf distances were computed for the remaining leaves. Optimal root-to-leaf distance was set to that of the 33rd percentile of all root-to-leaf distances (i.e., favoring relatively slowly evolving genes but not the most slowly evolving ones). Each terminal node was "rewarded" for the minimum deviation of its root-to-leaf distance from the optimum value (computed as the ratio of the smaller of the two values to the larger one) and for the distance to the earlier selected representatives (the shortest distance was used). The leaf with the largest product

of the two rewards was added to the pool of selected representatives; leaf-to-leaf distance rewards were recomputed (as the selected set changed) and the process was repeated until the required number of representatives was taken or the pool of candidates was exhausted.

Second-round Maximum Likelihood trees. Representative sequences were re-aligned using the MUSCLE program (Edgar 2004). Maximum Likelihood trees were constructed using the TreeFinder program (Jobb, von Haeseler, and Strimmer 2004), with the evolutionary model chosen by PhyMLtree log-likelihood comparisons with estimated site rate heterogeneity. Tree topologies were compared using the TreeFinder program according to either their expected likelihood weights (Strimmer and Rambaut 2002) or by the Approximately Unbiased test (AU) *p*-value (Shimodaira 2002).

References

Adachi, J., P. J. Waddell, W. Martin, and M. Hasegawa. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol **50**:348-358.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**:3389-3402.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins *in* M. O. Dayhoff, ed. Atlas of Protein Sequence and Structure Nat. Biomed. Res. Foundation, Washington, DC.

Dimmic, M. W., J. S. Rest, D. P. Mindell, and R. A. Goldstein. 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol **55**:65-73.

Doolittle, W. F. 1999. Lateral genomics. Trends Cell Biol **9**:M5-8.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**:1792-1797.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol **266**:418-427.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52**:696-704.

Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A **89**:10915-10919.

Jobb, G., A. von Haeseler, and K. Strimmer. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol **4**:18.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci **8**:275-282.

Kosiol, C., and N. Goldman. 2005. Different versions of the Dayhoff rate matrix. Mol Biol Evol **22**:193-199.

Muller, T., and M. Vingron. 2000. Modeling amino acid replacement. J Comput Biol **7**:761-776.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol **51**:492-508.

Strimmer, K., and A. Rambaut. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc Biol Sci **269**:137-142.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol **18**:691-699.

Supplemental references

Table S1. The statistics of tested topologies for 136 trees.

Family		Log-Likelihood [a]			AU [b]			ELW [c]			S [d]
		C	CA	EA	C	CA	EA	C	CA	EA	
arCOG00018-KOD03974-COG00063	G; Predicted sugar kinase	-12572	-12582	-12582	0.790	0.212	0.301	0.770	0.086	0.144	
arCOG00027-KOD01383-COG00076	E; Glutamate decarboxylase and related PLP-dependent proteins	-15951	-15956	-15955	0.686	0.304	0.446	0.622	0.133	0.245	
arCOG00029-KOD01377-COG00461	F; Orotate phosphoribosyltransferase	-7557	-7562	-7567	0.791	0.351	0.193	0.676	0.216	0.108	
arCOG00042-KOD02840-COG00037	D; Predicted ATPase of the PP-loop superfamily implicated in cell cycle control	-11939	-11942	-11939	0.657	0.371	0.534	0.376	0.297	0.328	
arCOG00063-KOD02387-COG00504	F; CTP synthase (UTP-ammonia lyase)	-21281	-21296	-21270	0.276	0.000	0.729	0.272	0.001	0.727	
arCOG00086-KOD00026-COG00512	E; Anthranilate/para-aminobenzoate synthases component II	-7505	-7506	-7506	0.744	0.292	0.387	0.532	0.197	0.271	
arCOG00090-KOD03179-COG00518	F; GMP synthase - Glutamine amidotransferase domain	-7330	-7340	-7340	0.853	0.219	0.200	0.835	0.083	0.082	
arCOG00109-KOD03191-COG02890	J; Methylase of polypeptide chain release factors	-8797	-8799	-8795	0.444	0.311	0.729	0.264	0.216	0.520	
arCOG00110-KOD02904-COG02813	J; 16S RNA G1207 methylase RsmC	-7773	-7775	-7774	0.670	0.354	0.472	0.443	0.185	0.372	
arCOG00245-KOD02380-COG00287	E; Prephenate dehydrogenase	-11111	-11110	-11112	0.526	0.609	0.402	0.272	0.522	0.206	
arCOG00312-KOD00854-COG00450	O; Peroxiredoxin	-7593	-7593	-7593	0.644	0.489	0.262	0.457	0.333	0.209	

arCOG00324-KOD01343-COG00123	B; Deacetylases including yeast histone deacetylase and acetoin utilization protein	-13049	-13049	-13048	0.267	0.281	0.807	0.190	0.191	0.619	
arCOG00350-KOD02423-COG01161	R; Predicted GTPases	-10552	-10549	-10548	0.351	0.472	0.708	0.291	0.257	0.452	
arCOG00402-KOD04163-COG00442	J; Prolyl-tRNA synthetase	-20255	-20266	-20265	0.773	0.325	0.322	0.659	0.180	0.161	A
arCOG00403-KOD02509-COG00172	J; Seryl-tRNA synthetase	-15937	-15930	-15935	0.218	0.849	0.116	0.211	0.728	0.061	
arCOG00405-KOD02298-COG00423	J; Glycyl-tRNA synthetase (class II)	-22370	-22364	-22366	0.330	0.614	0.509	0.115	0.434	0.452	
arCOG00406-KOD00556-COG00017	J; Aspartyl/asparaginyl-tRNA synthetases	-19133	-19141	-19142	0.783	0.282	0.095	0.711	0.252	0.037	
arCOG00410-KOD02784-COG00016	J; Phenylalanyl-tRNA synthetase alpha subunit	-18167	-18165	-18165	0.120	0.509	0.578	0.098	0.426	0.476	
arCOG00412-KOD02472-COG00072	J; Phenylalanyl-tRNA synthetase beta subunit	-26797	-26802	-26800	0.603	0.174	0.490	0.579	0.071	0.349	
arCOG00415-KOD01434-COG00468	L; RecA/RadA recombinase	-9651	-9645	-9644	0.241	0.255	0.874	0.228	0.254	0.518	EA
arCOG00469-KOD00991-COG00470	L; ATPase involved in DNA replication	-12123	-12123	-12122	0.499	0.530	0.555	0.387	0.219	0.394	
arCOG00474-KOD04410-COG00212	H; 5-formyltetrahydrofolate cyclo-ligase	-8445	-8446	-8446	0.637	0.391	0.509	0.353	0.207	0.440	
arCOG00476-KOD07663-COG00382	H; 4-hydroxybenzoate polyprenyltransferase and related prenyltransferases	-13049	-13059	-13058	0.940	0.000	0.067	0.917	0.012	0.071	
arCOG00487-KOD04426-COG00018	J; Arginyl-tRNA synthetase	-26893	-26896	-26904	0.557	0.468	0.029	0.553	0.435	0.012	
arCOG00488-KOD01636-	L; DNA polymerase sliding clamp subunit (PCNA homolog)	-12905	-12910	-12913	0.705	0.371	0.169	0.662	0.290	0.048	

COG00592										
arCOG00494-KOD04777-COG00136	E; Aspartate-semialdehyde dehydrogenase	-13687	-13686	-13687	0.449	0.661	0.437	0.241	0.433	0.327
arCOG00618-KOD03055-COG00106	E; Phosphoribosylformimino-5-aminoimidazole carboxamide ribonucleotide (ProFAR) isomerase	-9977	-9977	-9974	0.306	0.392	0.727	0.135	0.311	0.554
arCOG00770-KOD01132-COG01199	K; Rad3-related DNA helicases	-11450	-11452	-11450	0.626	0.342	0.517	0.382	0.154	0.464
arCOG00779-KOD01742-COG00200	J; Ribosomal protein L15	-6115	-6125	-6123	0.824	0.115	0.258	0.783	0.055	0.161
arCOG00807-KOD00434-COG00060	J; Isoleucyl-tRNA synthetase	-40464	-40458	-40480	0.316	0.684	0.000	0.325	0.675	0.000
arCOG00809-KOD00437-COG00495	J; Leucyl-tRNA synthetase	-38538	-38546	-38544	0.756	0.355	0.383	0.476	0.256	0.268
arCOG00810-KOD01247-COG00143	J; Methionyl-tRNA synthetase	-23962	-23933	-23971	0.013	0.994	0.000	0.014	0.974	0.013
arCOG00833-KOD03235-COG00456	R; Acetyltransferases	-5439	-5438	-5439	0.274	0.669	0.462	0.207	0.505	0.288
arCOG00874-KOD01123-COG01061	K; DNA or RNA helicases of superfamily II	-17543	-17545	-17544	0.598	0.477	0.497	0.362	0.233	0.405
arCOG00914-KOD01402-COG04992	E; Ornithine/acetylornithine aminotransferase	-15789	-15794	-15792	0.671	0.418	0.459	0.436	0.276	0.288
arCOG00973-KOD01122-COG00144	J; tRNA and rRNA cytosine-C5-methylases	-9298	-9298	-9284	0.077	0.079	0.947	0.048	0.025	0.928
arCOG00976-KOD01661-COG02518	O; Protein-L-isoaspartate carboxylmethyltransferase	-7840	-7837	-7838	0.194	0.832	0.342	0.178	0.546	0.276
arCOG00987-KOD02529-COG00130	J; Pseudouridine synthase	-7490	-7491	-7484	0.139	0.051	0.902	0.128	0.042	0.830

arCOG01001-KOD02775-COG00024	J; Methionine aminopeptidase	-11938	-11938	-11939	0.534	0.618	0.345	0.363	0.431	0.206
arCOG01122-KOD03075-COG00120	G; Ribose 5-phosphate isomerase	-9185	-9183	-9185	0.400	0.654	0.478	0.223	0.522	0.255
arCOG01136-KOD02241-COG00073	R; EMAP domain	-4701	-4707	-4707	0.786	0.288	0.299	0.750	0.125	0.125
arCOG01141-KOD03325-COG00622	R; Predicted phosphoesterase	-6959	-6956	-6955	0.314	0.567	0.618	0.194	0.365	0.441
arCOG01143-KOD00372-COG00639	T; Diadenosine tetraphosphatase and related serine/threonine protein phosphatases	-8321	-8320	-8322	0.558	0.582	0.357	0.335	0.499	0.166
arCOG01163-KOD00785-COG00473	C; Isocitrate/isopropylmalate dehydrogenase	-14990	-14976	-14992	0.082	0.931	0.116	0.040	0.863	0.097
arCOG01169-KOD02670-COG00148	G; Enolase	-15059	-15065	-15075	0.685	0.341	0.029	0.669	0.321	0.010
arCOG01179-KOD03403-COG00361	J; Translation initiation factor 1 (IF-1)	-2806	-2808	-2806	0.548	0.147	0.547	0.432	0.089	0.479
arCOG01183-KOD02708-COG00533	O; Metal-dependent proteases with possible chaperone activity	-10700	-10699	-10701	0.505	0.652	0.407	0.268	0.407	0.325
arCOG01225-KOD01532-COG01100	R; GTPase SAR1 and related small G proteins	-10445	-10448	-10448	0.889	0.172	0.161	0.753	0.123	0.124
arCOG01227-KOD00781-COG00552	U; Signal recognition particle GTPase	-13092	-13090	-13090	0.088	0.497	0.566	0.061	0.458	0.481
arCOG01228-KOD00780-COG00541	U; Signal recognition particle GTPase	-19259	-19262	-19261	0.754	0.349	0.375	0.489	0.310	0.201
arCOG01257-KOD00357-COG00459	O; Chaperonin GroEL (HSP60 family)	-20007	-20009	-20009	0.802	0.289	0.184	0.662	0.213	0.125
arCOG01292-KOD01800-	E; NADPH-dependent glutamate synthase beta chain and related oxidoreductases	-14745	-14750	-14750	0.874	0.205	0.159	0.748	0.160	0.092

COG00493										
arCOG01307-KOD00739-COG00464	O; ATPases of the AAA+ class	-12427	-12427	-12428	0.560	0.536	0.185	0.441	0.414	0.145
arCOG01351-KOD00455-COG00460	E; Homoserine dehydrogenase	-13530	-13516	-13526	0.160	0.891	0.066	0.157	0.805	0.038
arCOG01352-KOD02250-COG00334	E; Glutamate dehydrogenase/leucine dehydrogenase	-16014	-16024	-16024	0.981	0.033	0.036	0.947	0.028	0.024
arCOG01358-KOD04355-COG00621	J; 2-methylthioadenine synthetase	-20219	-20219	-20216	0.395	0.344	0.727	0.334	0.171	0.496
arCOG01371-KOD00747-COG01088	M; dTDP-D-glucose 4 6-dehydratase	-11623	-11621	-11624	0.512	0.593	0.466	0.227	0.460	0.313
arCOG01482-KOD03008-COG00157	H; Nicotinate-nucleotide pyrophosphorylase	-10997	-11013	-11013	0.914	0.111	0.140	0.898	0.048	0.054
arCOG01527-KOD01957-COG00550	L; Topoisomerase IA	-28055	-28060	-28058	0.665	0.243	0.448	0.522	0.087	0.391
arCOG01529-KOD01175-COG00365	I; Acyl-coenzyme A synthetases/AMP-(fatty) acid ligases	-26159	-26151	-26168	0.229	0.867	0.096	0.185	0.753	0.062
arCOG01532-KOD01602-COG00020	I; Undecaprenyl pyrophosphate synthase	-10667	-10663	-10661	0.365	0.483	0.686	0.268	0.272	0.460
arCOG01559-KOD00469-COG00480	J; Translation elongation factors (GTPases)	-30537	-30538	-30539	0.636	0.478	0.207	0.445	0.452	0.103
arCOG01560-KOD01144-COG00532	J; Translation initiation factor 2 (IF-2; GTPase)	-21751	-21743	-21751	0.415	0.674	0.415	0.234	0.603	0.163
arCOG01575-KOD01068-COG00689	J; RNase PH	-9020	-9031	-9032	0.865	0.201	0.204	0.794	0.080	0.126
arCOG01594-KOD00370-COG00458	E; Carbamoylphosphate synthase large subunit (split gene in MJ)	-38565	-38374	-38564	0.000	0.235	0.000	0.000	1.000	0.000

arCOG01695-KOD02030-COG01293	K; Predicted RNA-binding protein homologous to eukaryotic snRNP	-28872	-28869	-28875	0.284	0.817	0.279	0.152	0.615	0.233	
arCOG01704-KOD00672-COG00452	H; Phosphopantothenoylcysteine synthetase/decarboxylase	-13586	-13612	-13612	0.990	0.018	0.016	0.978	0.008	0.014	
arCOG01706-KOD00558-COG00508	C; Pyruvate/2-oxoglutarate dehydrogenase complex dihydrolipoamide acyltransferase (E2) component and related enzymes	-14333	-14331	-14334	0.437	0.667	0.180	0.319	0.583	0.098	
arCOG01722-KOD03311-COG00099	J; Ribosomal protein S13	-5409	-5409	-5406	0.213	0.208	0.861	0.161	0.161	0.678	
arCOG01739-KOD03342-COG00681	U; Signal peptidase I	-5763	-5764	-5764	0.747	0.396	0.283	0.520	0.253	0.227	
arCOG01751-KOD03387-COG01358	J; Ribosomal protein HS6-type (S12/L30/L7a)	-4729	-4730	-4730	0.736	0.256	0.380	0.523	0.215	0.263	
arCOG01758-KOD00900-COG00051	J; Ribosomal protein S10	-3755	-3753	-3751	0.360	0.329	0.766	0.325	0.185	0.489	
arCOG01887-KOD02145-COG00180	J; Tryptophanyl-tRNA synthetase	-14127	-14089	-14127	0.000	0.805	0.000	0.000	1.000	0.000	
arCOG01891-KOD03327-COG00125	F; Thymidylate kinase	-8255	-8245	-8253	0.066	0.866	0.172	0.023	0.821	0.157	
arCOG01920-KOD01999-COG00250	K; Transcription antiterminator	-7163	-7159	-7161	0.363	0.600	0.535	0.154	0.525	0.321	
arCOG01924-KOD00503-COG00252	E; L-asparaginase/archaeal Glu-tRNAGln amidotransferase subunit D	-14581	-14596	-14596	0.903	0.144	0.150	0.888	0.056	0.056	A
arCOG02014-KOD01223-COG00147	E; Anthranilate/para-aminobenzoate synthases component I	-16263	-16275	-16269	0.737	0.000	0.265	0.718	0.002	0.280	
arCOG02208-KOD02831-COG00040	E; ATP phosphoribosyltransferase	-11049	-11050	-11049	0.563	0.147	0.521	0.464	0.119	0.416	
arCOG02297-KOD00975-	E; Branched-chain amino acid aminotransferase/4-amino-4-	-11545	-11540	-11547	0.242	0.802	0.091	0.197	0.767	0.037	

COG00115	deoxychorismate lyase									
arCOG02303-KOD06476-COG00496	R; Predicted acid phosphatase	-10274	-10274	-10278	0.674	0.597	0.180	0.417	0.408	0.176
arCOG02431-KOD05042-COG01611	R; Predicted Rossmann fold nucleotide-binding protein	-7053	-7058	-7058	0.815	0.261	0.256	0.767	0.116	0.116
arCOG02833-KOD09166-COG00265	O; Trypsin-like serine proteases typically periplasmic contain C-terminal PDZ domain	-7497	-7498	-7498	0.691	0.324	0.389	0.424	0.273	0.303
arCOG02969-KOD01046-COG00308	E; Aminopeptidase N	-15523	-15524	-15497	0.042	0.063	0.961	0.018	0.056	0.926
arCOG03199-KOD02788-COG00472	M; UDP-N-acetylmuramyl pentapeptide phosphotransferase/UDP-N-acetylglucosamine-1-phosphate transferase	-14145	-14145	-14135	0.119	0.203	0.847	0.053	0.121	0.825
arCOG04050-KOD02519-COG00258	L; 5'-3' exonuclease (including N-terminal domain of PolI)	-12077	-12083	-12078	0.611	0.151	0.511	0.514	0.073	0.413
arCOG04064-KOD02921-COG00750	M; Predicted membrane-associated Zn-dependent proteases 1	-8068	-8068	-8069	0.611	0.534	0.288	0.372	0.487	0.141
arCOG04067-KOD02309-COG00090	J; Ribosomal protein L2	-9148	-9153	-9147	0.592	0.392	0.584	0.304	0.284	0.412
arCOG04070-KOD00746-COG00087	J; Ribosomal protein L3	-13471	-13475	-13476	0.696	0.428	0.246	0.647	0.236	0.116
arCOG04071-KOD01475-COG00088	J; Ribosomal protein L4	-11194	-11194	-11198	0.572	0.581	0.246	0.449	0.437	0.114
arCOG04072-KOD01751-COG00089	J; Ribosomal protein L23	-3474	-3471	-3474	0.304	0.794	0.280	0.209	0.652	0.139
arCOG04086-KOD03184-COG01841	J; Ribosomal protein L30/L7E	-5159	-5158	-5159	0.349	0.743	0.339	0.247	0.506	0.247
arCOG04087-KOD00877-COG00098	J; Ribosomal protein S5	-6732	-6730	-6736	0.432	0.620	0.105	0.411	0.557	0.032

arCOG04088-KOD00875-COG00256	J; Ribosomal protein L18	-6275	-6277	-6275	0.536	0.397	0.573	0.474	0.150	0.376
arCOG04090-KOD03255-COG00097	J; Ribosomal protein L6P/L9E	-7957	-7947	-7953	0.175	0.811	0.316	0.092	0.726	0.182
arCOG04091-KOD01754-COG00096	J; Ribosomal protein S8	-5250	-5245	-5244	0.155	0.571	0.578	0.090	0.397	0.513
arCOG04092-KOD00397-COG00094	J; Ribosomal protein L5	-6405	-6402	-6404	0.296	0.583	0.580	0.156	0.561	0.282
arCOG04094-KOD03401-COG00198	J; Ribosomal protein L24	-4126	-4126	-4125	0.521	0.469	0.591	0.389	0.288	0.323
arCOG04095-KOD00901-COG00093	J; Ribosomal protein L14	-4138	-4190	-4190	0.636	0.000	0.000	1.000	0.000	0.000
arCOG04096-KOD01728-COG00186	J; Ribosomal protein S17	-4013	-3998	-3997	0.099	0.526	0.579	0.051	0.479	0.470
arCOG04097-KOD03181-COG00092	J; Ribosomal protein S3	-8927	-8928	-8928	0.746	0.341	0.338	0.533	0.233	0.233
arCOG04098-KOD03353-COG00091	J; Ribosomal protein L22	-5818	-5818	-5818	0.493	0.662	0.374	0.472	0.314	0.214
arCOG04099-KOD00898-COG00185	J; Ribosomal protein S19	-4566	-4566	-4565	0.191	0.291	0.816	0.172	0.194	0.633
arCOG04113-KOD00857-COG00197	J; Ribosomal protein L16/L10E	-5917	-5918	-5918	0.602	0.459	0.520	0.361	0.369	0.270
arCOG04121-KOD02299-COG00164	L; Ribonuclease HII	-8051	-8051	-8053	0.573	0.490	0.089	0.493	0.424	0.083
arCOG04131-KOD00820-COG00030	J; Dimethyladenosine transferase (rRNA methylation)	-11072	-11070	-11065	0.243	0.369	0.761	0.115	0.222	0.664
arCOG04133-KOD04492-	E; Chorismate synthase	-13667	-13695	-13693	0.977	0.019	0.035	0.963	0.008	0.029

COG00082											
arCOG04147-KOD00876-COG00605	P; Superoxide dismutase	-7235	-7258	-7254	0.947	0.040	0.092	0.892	0.026	0.082	
arCOG04157-KOD05009-COG00270	L; Site-specific DNA methylase	-10745	-10767	-10767	0.968	0.052	0.051	0.959	0.020	0.021	
arCOG04169-KOD01373-COG00201	U; Preprotein translocase subunit SecY	-20163	-20161	-20176	0.419	0.784	0.207	0.294	0.528	0.178	CA
arCOG04184-KOD03222-COG00127	F; Xanthosine triphosphate pyrophosphatase	-8093	-8095	-8097	0.681	0.391	0.196	0.634	0.305	0.061	
arCOG04185-KOD00400-COG00184	J; Ribosomal protein S15P/S13E	-5502	-5502	-5502	0.392	0.564	0.558	0.295	0.345	0.360	
arCOG04223-KOD01770-COG00023	J; Translation initiation factor 1 (eIF-1/SUI1) and related proteins	-3962	-3965	-3965	0.713	0.375	0.368	0.644	0.178	0.178	
arCOG04231-KOD03338-COG01324	P; Uncharacterized protein involved in tolerance to divalent cations	-4304	-4303	-4302	0.446	0.510	0.615	0.279	0.351	0.370	
arCOG04239-KOD03301-COG00522	J; Ribosomal protein S4 and related proteins	-7667	-7667	-7667	0.360	0.373	0.732	0.276	0.276	0.448	
arCOG04240-KOD00407-COG00100	J; Ribosomal protein S11	-3993	-3992	-3989	0.367	0.399	0.722	0.279	0.171	0.549	
arCOG04241-KOD01522-COG00202	K; DNA-directed RNA polymerase alpha subunit/40 kD subunit	-12243	-12245	-12245	0.862	0.208	0.203	0.713	0.143	0.143	
arCOG04242-KOD03204-COG00102	J; Ribosomal protein L13	-5841	-5834	-5834	0.261	0.645	0.572	0.193	0.400	0.407	
arCOG04243-KOD01753-COG00103	J; Ribosomal protein S9	-5134	-5127	-5127	0.099	0.340	0.812	0.089	0.323	0.589	
arCOG04245-KOD00830-COG00052	J; Ribosomal protein S2	-7369	-7371	-7371	0.731	0.382	0.256	0.540	0.310	0.149	

arCOG04248-KOD02683-COG00846	K; NAD-dependent protein deacetylases SIR2 family	-8015	-8020	-8020	0.915	0.122	0.133	0.842	0.075	0.083	
arCOG04254-KOD03291-COG00049	J; Ribosomal protein S7	-6328	-6337	-6335	0.744	0.163	0.356	0.726	0.073	0.201	
arCOG04255-KOD01749-COG00048	J; Ribosomal protein S12	-6236	-6244	-6237	0.811	0.259	0.370	0.491	0.253	0.257	A
arCOG04257-KOD00260-COG00086	K; DNA-directed RNA polymerase beta' subunit/160 kD subunit	-35415	-35408	-35410	0.265	0.622	0.515	0.106	0.490	0.404	
arCOG04277-KOD03271-COG00231	J; Translation elongation factor P (EF-P)/translation initiation factor 5A (eIF-5A)	-5987	-5990	-5994	0.688	0.490	0.328	0.543	0.240	0.217	
arCOG04288-KOD00815-COG00244	J; Ribosomal protein L10	-11790	-11792	-11792	0.621	0.479	0.436	0.568	0.216	0.216	
arCOG04289-KOD01570-COG00081	J; Ribosomal protein L1	-9573	-9579	-9579	0.915	0.130	0.089	0.855	0.084	0.061	
arCOG04302-KOD01147-COG00008	J; Glutamyl- and glutaminyl-tRNA synthetases	-20153	-20158	-20159	0.724	0.401	0.360	0.577	0.193	0.230	
arCOG04372-KOD00886-COG00080	J; Ribosomal protein L11	-6362	-6358	-6360	0.318	0.793	0.217	0.299	0.571	0.130	
arCOG04398-KOD03143-COG00131	E; Imidazoleglycerol-phosphate dehydratase	-8103	-8101	-8107	0.468	0.603	0.129	0.437	0.519	0.043	
arCOG04449-KOD02554-COG00101	J; Pseudouridylate synthase	-9362	-9360	-9365	0.452	0.606	0.065	0.418	0.544	0.039	
arCOG04465-KOD05645-COG00059	E; Ketol-acid reductoisomerase	-11753	-11742	-11754	0.371	0.736	0.287	0.238	0.662	0.100	
arCOG05412-KOD00626-COG02723	G; Beta-glucosidase/6-phospho-beta-glucosidase/beta-galactosidase	-17714	-17964	-17964	0.235	0.000	0.000	1.000	0.000	0.000	A
1arCOG00404-KOD01936-	J; Histidyl-tRNA synthetase	-17612	-17597	-17614	0.028	0.978	0.000	0.029	0.966	0.005	

COG00124										
2arCOG00404-KOD01936-COG00124	J; Histidyl-tRNA synthetase	-16012	-16003	-16013	0.211	0.861	0.157	0.131	0.813	0.056
1arCOG01886-KOD02144-COG00162	J; Tyrosyl-tRNA synthetase	-13963	-13967	-13959	0.368	0.271	0.780	0.194	0.176	0.630
2arCOG01886-KOD02144-COG00162	J; Tyrosyl-tRNA synthetase	-14510	-14497	-14510	0.117	0.925	0.099	0.055	0.913	0.033
Total		**-1668665**	***-1668865***	***-1669217***						

[a]- Log-likelihood, reported by TreeFinder for the three competing constrained topologies: C, "classical"; CA, "crenarchaeal" and EA, "euryarchaeal".

[b]- AU test *p*-values, reported by TreeFinder for the three competing constrained topologies.

[c]- ELW values, reported by TreeFinder for the three competing constrained topologies.

[d]- Synapomorphies found in the alignments.

Table S2. The 67 analyzed eukaryotic genomes

Species	Taxon	Source
Cryptosporidium parvum	Alveolata, Apicomplexa	GB
Plasmodium berghei	Alveolata, Apicomplexa	GB
Plasmodium chabaudi	Alveolata, Apicomplexa	GB
Plasmodium falciparum	Alveolata, Apicomplexa	GB
Plasmodium yoelii yoelii	Alveolata, Apicomplexa	GB
Theileria annulata	Alveolata, Apicomplexa	GB
Theileria parva	Alveolata, Apicomplexa	GB
Tetrahymena thermophila	Alveolata, Ciliophora	GB
Giardia lamblia	Diplomonadida group, Diplomonadida	GB
Entamoeba histolytica	Entamoebidae, Entamoeba	TIGR
Leishmania major	Euglenozoa, Kinetoplastida	TIGR
Trypanosoma brucei	Euglenozoa, Kinetoplastida	GB
Trypanosoma cruzi	Euglenozoa, Kinetoplastida	GB
Dictyostelium discoideum	Mycetozoa, Dictyosteliida	GB
Cyanidioschyzon merolae	Rhodophyta, Bangiophyceae	University of Tokyo
Phaeodactylum tricornutum	stramenopiles, Bacillariophyta	JGI
Thalassiosira pseudonana	stramenopiles, Bacillariophyta	JGI
Phytophthora ramorum	stramenopiles, Oomycetes	JGI
Phytophthora sojae	stramenopiles, Oomycetes	JGI
Chlamydomonas reinhardtii	Viridiplantae, Chlorophyta	JGI
Ostreococcus lucimarinus	Viridiplantae, Chlorophyta	JGI
Ostreococcus tauri	Viridiplantae, Chlorophyta	EMBL
Arabidopsis thaliana	Viridiplantae, Streptophyta	GB
Oryza sativa (japonica)	Viridiplantae, Streptophyta	GB
Populus trichocarpa	Viridiplantae, Streptophyta	JGI
Aspergillus fumigatus	Fungi, Ascomycota	GB
Aspergillus oryzae	Fungi, Ascomycota	GB
Candida albicans	Fungi, Ascomycota	GB
Candida glabrata	Fungi, Ascomycota	GB
Debaryomyces hansenii	Fungi, Ascomycota	GB
Aspergillus nidulans	Fungi, Ascomycota	GB
Eremothecium gossypii	Fungi, Ascomycota	GB
Gibberella zeae	Fungi, Ascomycota	GB
Trichoderma reesei	Fungi, Ascomycota	JGI
Kluyveromyces lactis	Fungi, Ascomycota	GB
Kluyveromyces waltii	Fungi, Ascomycota	MIT

Magnaporthe grisea	Fungi, Ascomycota	GB
Neurospora crassa	Fungi, Ascomycota	GB
Saccharomyces cerevisiae	Fungi, Ascomycota	GB
Schizosaccharomyces pombe	Fungi, Ascomycota	GB
Yarrowia lipolytica	Fungi, Ascomycota	GB
Cryptococcus neoformans	Fungi, Basidiomycota	GB
Laccaria bicolor	Fungi, Basidiomycota	JGI
Phanerochaete chrysosporium	Fungi, Basidiomycota	JGI
Ustilago maydis	Fungi, Basidiomycota	GB
Encephalitozoon cuniculi	Fungi, Microsporidia	GB
Anopheles gambiae	Metazoa, Arthropoda	GB
Apis mellifera	Metazoa, Arthropoda	EMBL
Drosophila melanogaster	Metazoa, Arthropoda	GB
Bos taurus	Metazoa, Chordata	Ensembl
Canis lupus familiaris	Metazoa, Chordata	GB
Ciona intestinalis	Metazoa, Chordata	JGI
Danio rerio	Metazoa, Chordata	GB
Gallus gallus	Metazoa, Chordata	EMBL
Monodelphis domestica	Metazoa, Chordata	Ensembl
Homo sapiens	Metazoa, Chordata	GB
Macaca mulatta	Metazoa, Chordata	Ensembl
Mus musculus	Metazoa, Chordata	GB
Pan troglodytes	Metazoa, Chordata	EMBL
Rattus norvegicus	Metazoa, Chordata	Ensembl
Takifugu rubripes	Metazoa, Chordata	EMBL
Tetraodon nigroviridis	Metazoa, Chordata	GB
Nematostella vectensis	Metazoa, Cnidaria	JGI
Strongylocentrotus purpuratus	Metazoa, Echinodermata	GB
Caenorhabditis briggsae	Metazoa, Nematoda	GB
Caenorhabditis elegans	Metazoa, Nematoda	GB
Monosiga brevicollis	Choanoflagellida, Codonosigidae	JGI

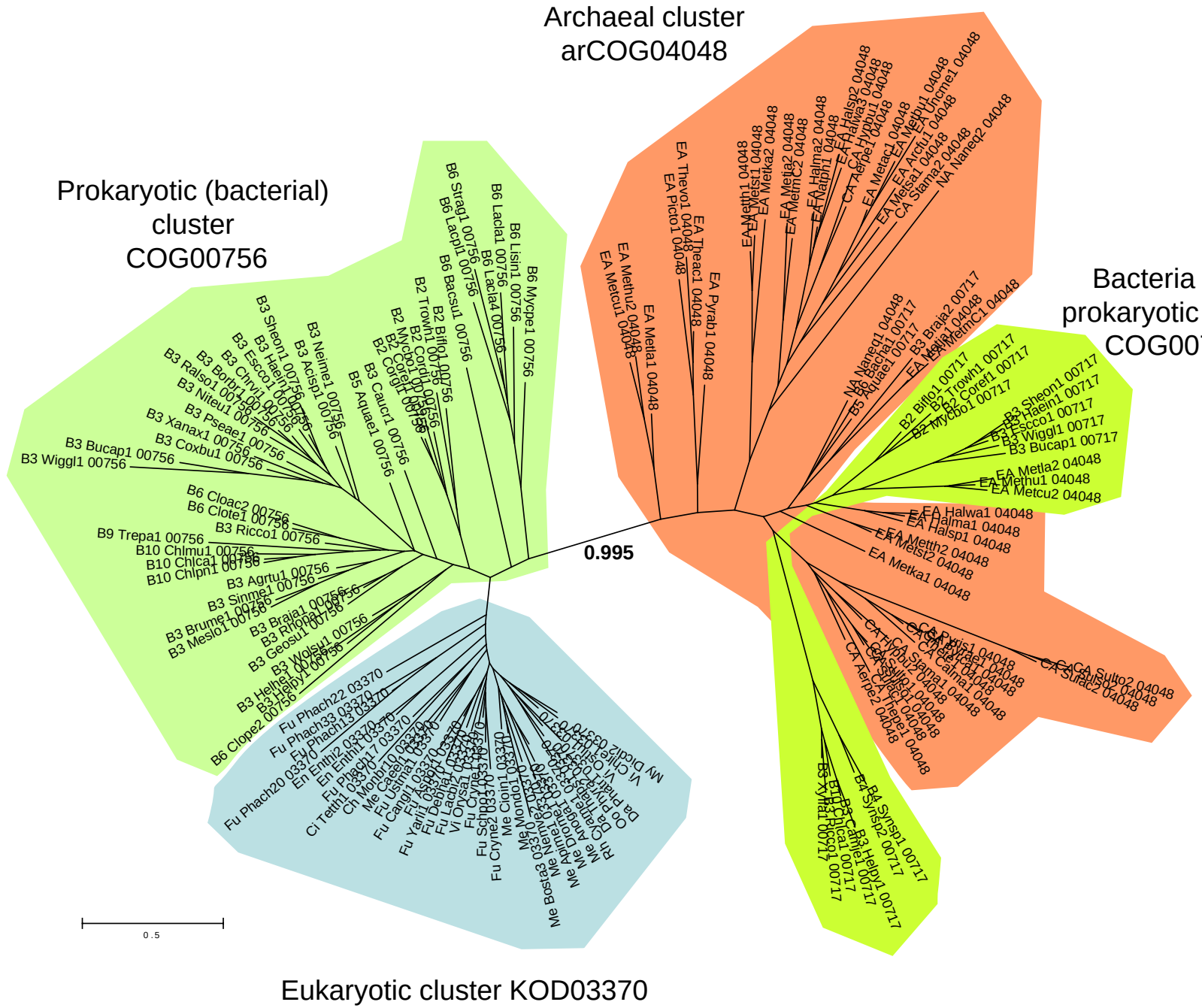

Archaeal cluster
arCOG04048
Prokaryotic (bacterial)
cluster
COG00756
Bacteria from
prokaryotic cluster
COG00717
0.995
0.5
Eukaryotic cluster KOD03370

Fig. S1

arCOG00402

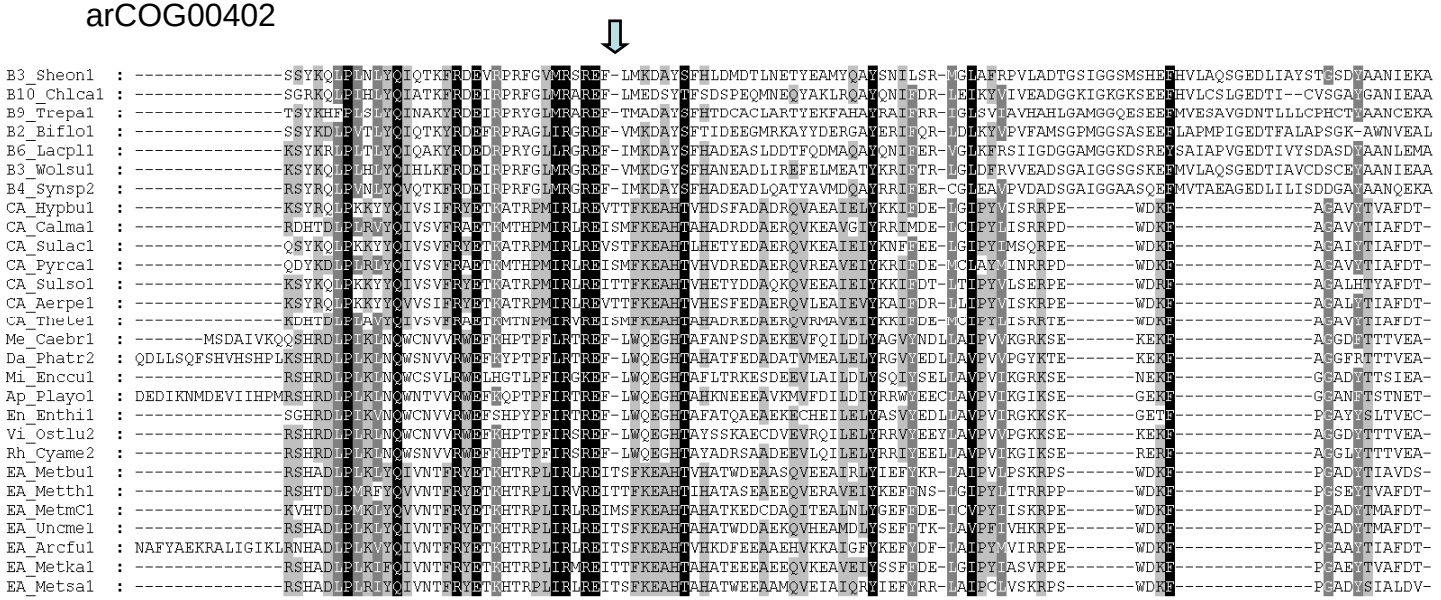


Fig. S2a

arCOG00415 – EA synapomorphy

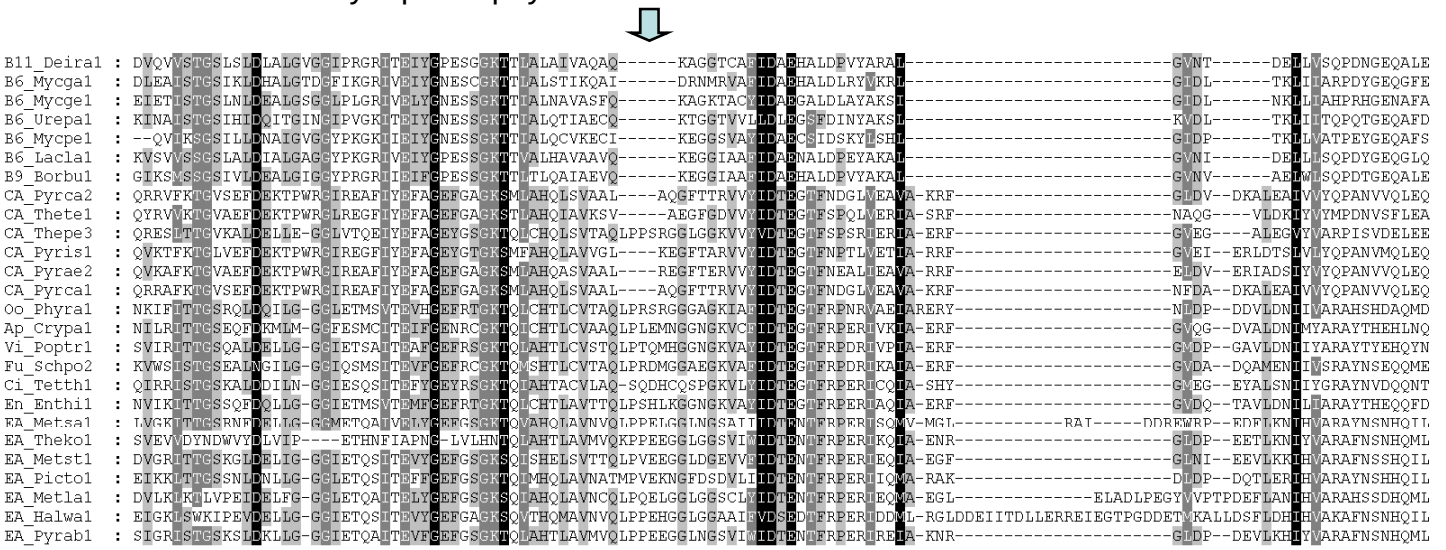


Fig. S2b

arCOG01924

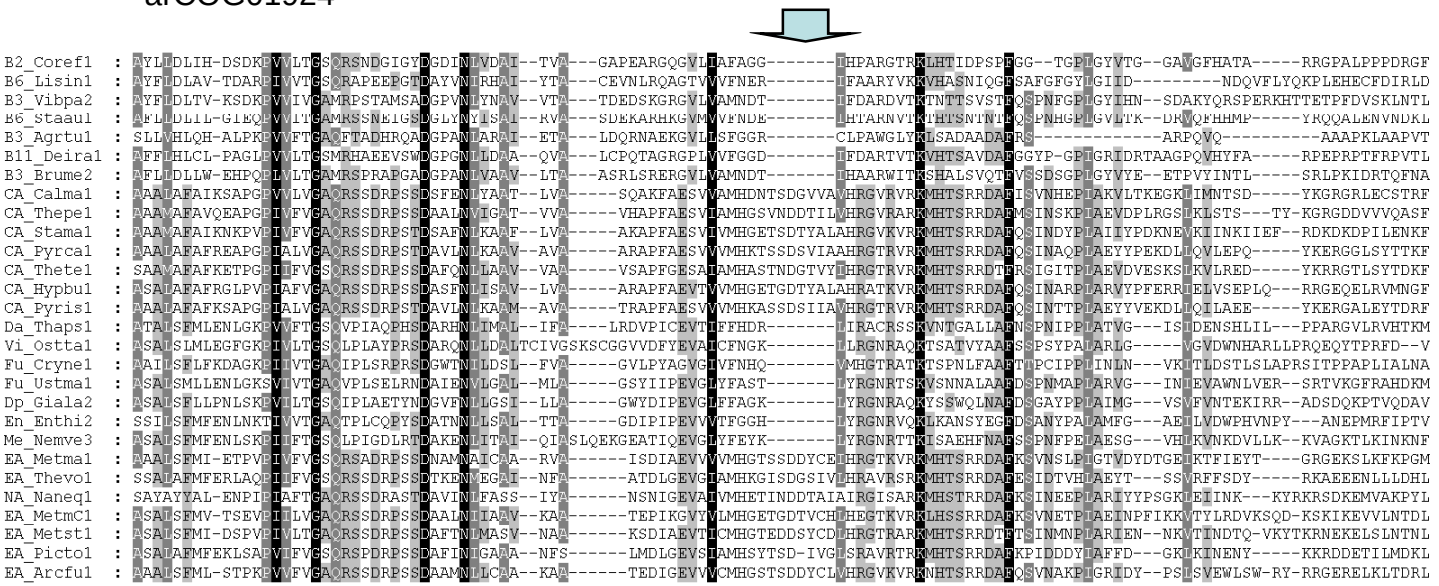


Fig. S2c

arCOG04169 – CA synapomorphy

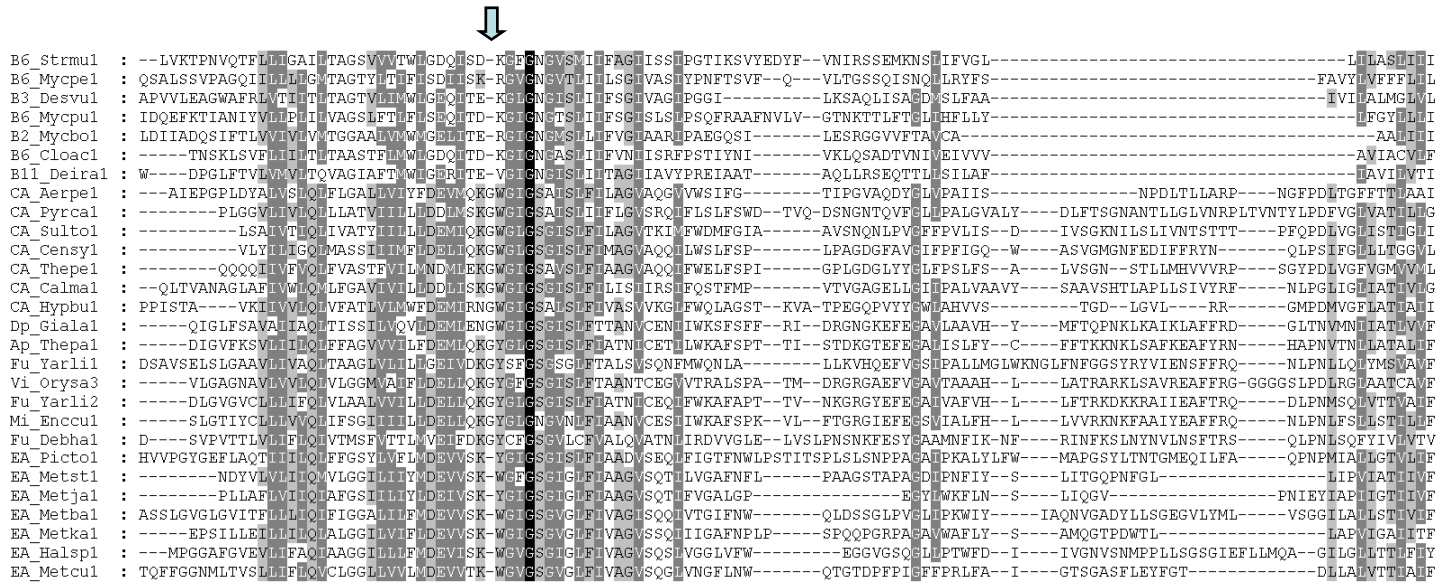


Fig. S2d

Fig. S2e

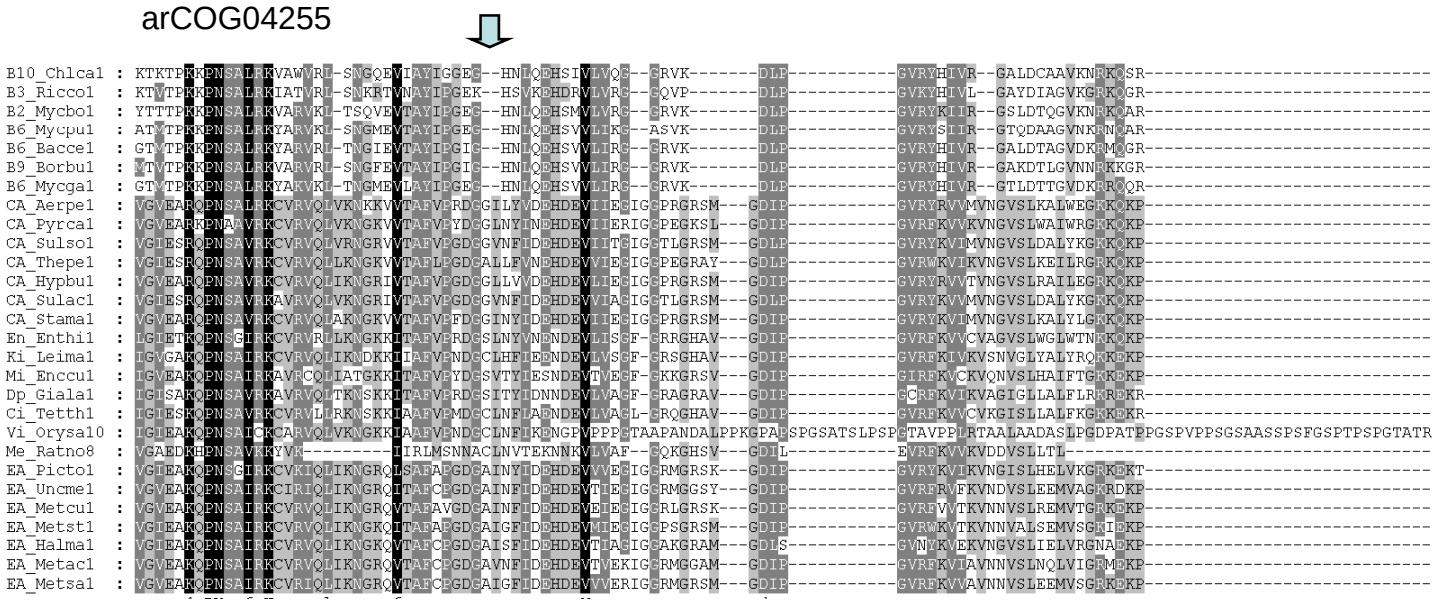
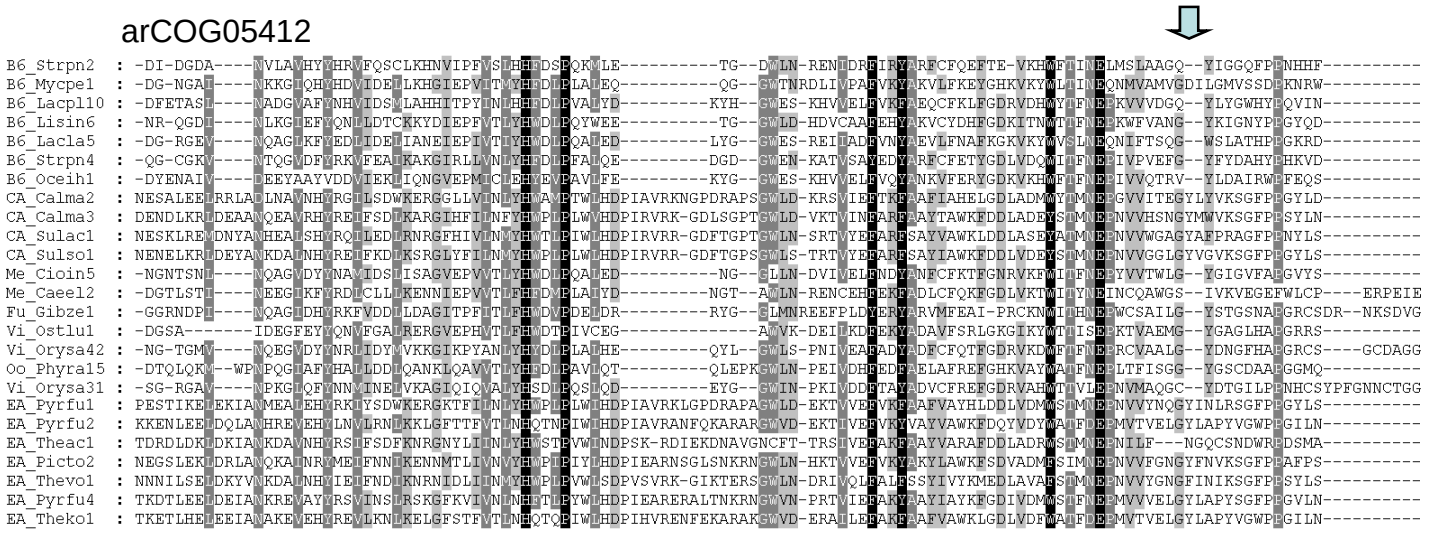


Fig. S2f



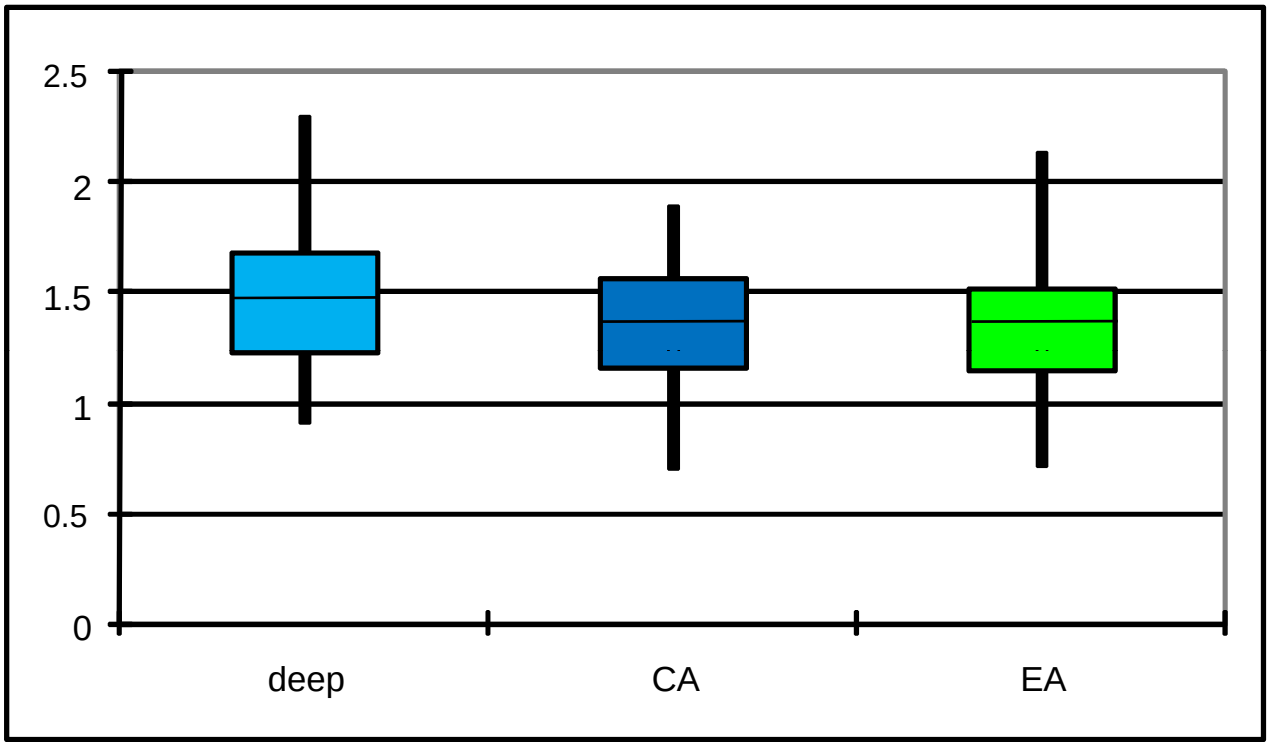


Fig. S3

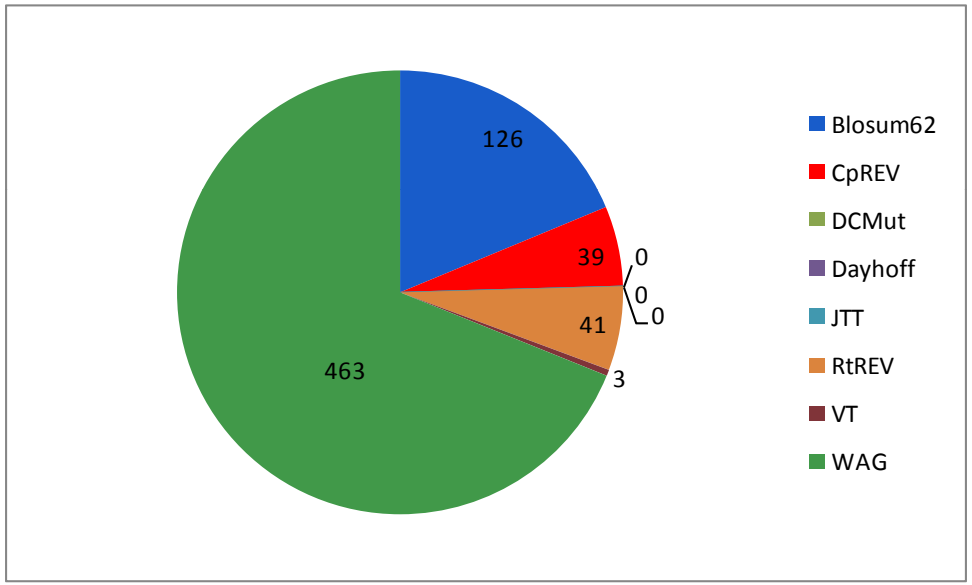

126
39
0
0
0
41
3
463
Blosum62
CpREV
DCMut
Dayhoff
JTT
RtREV
VT
WAG

Fig. S4