nature research

https://doi.org/10.1038/s41559-019-1040-x

Supplementary information

Phylogenomics provides robust support for a two-domains tree of life

In the format provided by the authors and unedited

Supplementary Tables

Supplementary Table 1: Root-to-tip distances for the 104 taxa in the Spang et al. (2015) alignment[1]. Trees were sampled under the CAT+GTR+G4 model in PhyloBayes. Sampled trees were rooted between Bacteria and Archaea, and root-to-tip distances calculated for each taxon. The mean, 2.5% and 97.5% values of these distributions are provided for each taxon. Taxa in bold were considered fast-evolving in Da Cunha et al.[2]

Taxon	Mean distance	2.5%	97.5%
Tetrahymena_thermophila	5.849104179	5.784589589	5.913618769
Leishmania_infantum	5.815973501	5.751396689	5.880550312
Saccharomyces_cerevisiae	5.807909231	5.743241311	5.87257715
Trichomonas_vaginalis	5.76846435	5.704008931	5.832919769
Entamoeba_histolytica_HM_1_IM	5.742424353	5.677797009	5.807051698
Plasmodium_falciparum	5.683036253	5.618404659	5.747667847
Dictyostelium_discoideum	5.623113785	5.558490627	5.687736943
Thalassiosira_pseudonana_CCMP	5.615031865	5.550506334	5.679557396
Homo_sapiens	5.580569228	5.515971968	5.645166488
Arabidopsis_thaliana	5.512460017	5.44786312	5.577056913
Ca_Parvarchaeum_acidophilus_A	5.113603908	5.048267033	5.178940783
Nanoarchaeote_Nst1	4.96108144	4.895721215	5.026441665
Cenarchaeum_symbiosum_A	4.928175449	4.86353693	4.992813968
Ca_Nanosalinarum_sp_J07AB56	4.922883574	4.857794847	4.987972302
Nitrosoarchaeum_limnia_SFB1	4.847062395	4.78249339	4.911631399
Nitrosoarchaeum_koreensis_MY1	4.843375152	4.778792439	4.907957865
Thaum_AAA007_O23	4.836958245	4.772426261	4.901490228
Nitrosopumilus_maritimus_SCM1	4.832059819	4.76749806	4.896621577
Ca_Korarchaeum_cryptofilum_OP	4.808877779	4.744025221	4.873730336
Nano_AAA011_D5	4.784601765	4.719367042	4.849836489
Ca_Micrarchaeum_acidiphilum_A	4.756183772	4.691136667	4.821230877
Metallosphaera_cuprina_Ar_4	4.721618254	4.657338637	4.78589787
Acidilobus_saccharovorans_345	4.706643061	4.64220692	4.771079203
Nanoarchaeum_equitans_Kin4_M	4.697011803	4.631762396	4.762261209
Metallosphaera_sedula_DSM_534	4.690183261	4.625893337	4.754473184
Caldivirga_maquilingensis_IC	4.684141772	4.619707338	4.748576206
Pyrobaculum_aerophilum_IM2	4.679487472	4.61487939	4.744095554
Pyrobaculum_calidifontis_JCM	4.666310065	4.601714933	4.730905198
Aiga_AAA471_F17	4.649056693	4.584523508	4.713589878
Ca_Nanosalina_sp_J07AB43	4.643910073	4.578817264	4.709002882
DeepDSAG_highGC	4.630909004	4.566287925	4.695530083
Geo_AAA471_B05	4.62566644	4.560944688	4.690388192
Aiga_0000106_J15	4.621385329	4.556818591	4.685952068
Thermoproteus_uzoniensis_768	4.618376427	4.553857783	4.682895071
Desulfurococcus_kamchatkensis	4.6027529	4.538157224	4.667348577
Ferroplasma_acidarmanus_fer1	4.594983824	4.530625847	4.659341802
Sulfolobus_acidocaldarius_DSM	4.594836079	4.530496556	4.659175601
Acidianus_hospitalis_W1	4.583640559	4.519347431	4.647933686
Sulfolobus_islandicus_M_16_4	4.580439321	4.516144706	4.644733935
Aeropyrum_pernix_K1	4.579376496	4.514890282	4.643862709
Thermosphaera_aggregans_DSM_1	4.57824046	4.513614428	4.642866492
Sulfolobus_solfataricus_P2	4.575375657	4.511082133	4.639669181
Uncultured_Marine_Group_II_Eu	4.570508424	4.506167395	4.634849453
Ca_Nitrososphaera_gargensis_G	4.567908093	4.503389471	4.632426715
Geoarchaeon_NAG1	4.564255468	4.499576228	4.628934709

Caldiarchaeum_subterraneum	4.548385936	4.483863508	4.612908365
Halobacterium_NRC_1	4.533690802	4.469469674	4.597911929
Sulfolobus_tokodaii_7	4.527604274	4.463287447	4.591921102
Ignisphaera_aggregans_DSM_172	4.520461992	4.456187027	4.584736957
Aiga_AAA471_G05	4.509281494	4.444727341	4.573835646
Fervidicoccus_fontis_Kam940	4.507516541	4.443142518	4.571890564
DeepDSAG_lowGC	4.507404274	4.442740837	4.572067712
Picrophilus_torridus_DSM_9790	4.506392982	4.442017017	4.570768946
Haloarcula_marismortui_ATCC_4	4.506085973	4.44190514	4.570266806
Vulcanisaeta_distributa_DSM_1	4.504333121	4.439890896	4.568775345
DSAG_LKC1BA_M01C	4.501508608	4.436947813	4.566069403
Diapher_AAA011_E11	4.481404729	4.416386105	4.546423352
Halalkalicoccus_jeotgali_B3	4.46840933	4.404229985	4.532588675
Ignicoccus_hospitalis_KIN4_I	4.461666857	4.396992169	4.526341546
Thermoplasma_acidophilum_DSM	4.43702895	4.372687746	4.501370154
MCG_SCGC_AB539E09	4.402904284	4.338238794	4.467569774
Staphylothermus_marinus_F1	4.401637464	4.337023592	4.466251337
Haloferax_volcanii_DS2	4.393409427	4.329276549	4.457542305
Methanospirillum_hungatei_JF	4.392056516	4.327813513	4.45629952
Pyrolobus_fumarii_1A	4.383721137	4.319294066	4.448148208
Thaum_AB_179_E04	4.382389076	4.317816029	4.446962123
Thermofilum_pendens_Hrk_5	4.363760386	4.29929671	4.428224061
Hyperthermus_butylicus_DSM_54	4.35286254	4.288432736	4.417292344
Nano_AAA011_G17	4.302108754	4.236979208	4.367238299
Methanosphaerula_palustris_E1	4.292580021	4.22838623	4.356773811
Methanocorpusculum_labreanum	4.261340186	4.197167761	4.325512611
Methanoculleus_marisnigri_JR1	4.236598121	4.172391993	4.30080425
Methanoplanus_petrolearius_DS	4.228224446	4.16400666	4.292442233
Aenigma_AAA011_O16	4.225873156	4.160770972	4.29097534
Eury_AAA252_I15	4.198301017	4.133650955	4.262951079
Methanosphaera_stadtmanae_DSM	4.010176975	3.945961854	4.074392096
MBGD_SCGC_AB539N05	4.010036202	3.945646449	4.074425954
Aciduliprofundum_boonei_T469	3.975129087	3.910831492	4.039426683
Methanococcus_maripaludis_C6	3.970455466	3.90623709	4.034673842
Methanocella_paludicola_SANAE	3.968585374	3.904368296	4.032802452
Methanohalobium_evestigatum_Z	3.960775392	3.896567587	4.024983198
Methanomassiliicoccus_lu_B10	3.941021402	3.87686274	4.005180063
Methanosaeta_thermophila_PT	3.93389748	3.869684929	3.99811003
Methanosarcina_acetivorans_C2	3.929562361	3.865297976	3.993826746
Archaeoglobus_fulgidus_DSM_43	3.90258109	3.838285054	3.966877126
Ferroglobus_placidus_DSM_1064	3.869291637	3.805013509	3.933569766
Methanopyrus_kandleri_AV19	3.852615574	3.787386554	3.917844594
Methanobacterium_AL_21	3.834192995	3.76994872	3.89843727
Methanocaldococcus_jannaschii	3.721031113	3.656804245	3.78525798
Methanothermobacter_thermauto	3.689897834	3.625702353	3.754093314
Methanotorris_igneus_Kol_5	3.677571139	3.613348888	3.741793389
Borrelia_burgdorferi_B31	3.671234905	3.61098911	3.7314807
Campylobacter_jejuni_NCTC_111	3.66504641	3.603337332	3.726755488
Pyrococcus_furiosus_DSM_3638	3.651422603	3.586930806	3.715914401
Thermococcus_kodakarensis_KOD	3.644041972	3.579548999	3.708534946
Rhodopirellula_baltica_SH_1	3.640456903	3.580024063	3.700889743
Methanothermus_fervidus_DSM_2	3.628864909	3.564645987	3.693083831
Chlamydia_trachomatis_D_UW_3	3.623215462	3.561654452	3.684776472
Rickettsia_prowazekii_Madrid	3.601694444	3.539547004	3.663841883
Bacteroides_thetaiotaomicron	3.491738522	3.430141891	3.553335152

Escherichia_coli_K_12_substr	3.347266041	3.285222585	3.409309496
Synechocystis_PCC_6803	3.345612131	3.287912548	3.403311713
Bacillus_subtilis_168	3.059990207	3.002201819	3.117778595
Thermotoga_maritima_MSB8	3.02929411	2.979040825	3.079547395

Supplementary Table 2: Log likelihoods of the two-domains and three-domains trees computed under the LG+G4+F model on the 35-gene alignment[2] in three maximum likelihood phylogeny programs. Tree searches in RAxML and IQ-Tree resulted in the eocyte tree (identical topology in both cases), while tree searches in PhyML resulted in a three-domains tree. The difference in likelihoods is not statistically significant according to the AU test (AU = 0.233 for three-domains tree, AU = 0.767 for two-domains/eocyte tree).

Program	**Eocyte tree**	**Three-domains tree**
PhyML 3.1	-684701.197745	-684745.835747 (Free search under 3D constraint); -684716.068548 (constrained to optimal 3D tree obtained by RAxML and IQ-Tree)
RAxML 8.2.4	-684701.197240	-684716.067999
IQ-Tree 1.6.2	-684701.201	-684716.068

Supplementary Table 3: Inference of the tree of life using maximum likelihood for a series of models that are more complex than LG+G4+F. While Bayesian methods are required to fit the most flexible and parameter-rich substitution models, a number of models that relax some of the simplifying assumptions of the LG+G4+F model can be fit by maximum likelihood; these models show improved fit (as measured by the standard Bayesian Information Criterion score; lower is better) and recover two-domains trees. +C60: accommodates compositional variation across the sites of the alignment using a mixture of 60 empirically-derived equilibrium frequency profiles[3]; +R: free-rates model (across-site rate variation does not have to be gamma-distributed); +PMSF: a computationally efficient approximation to C60-type profile mixture models[4]. Analyses were performed in IQ-Tree 1.6[5].

Dataset	**Model**	**BIC score**	**Most likely tree**
35 gene[2]	LG+G4+F	1371218.083	2D
35 gene	LG+R8	1368056.447	2D
35 gene	LG+R8+PMSF	n/a	2D
35 gene	LG+C60+G4	1355157.707	2D
35 gene	LG+C60+G4+F	1344206.094	2D
35 gene, SR4-recoded	GTR+G4+F	547821.478	2D
35 gene, SR4-recoded	GTR+R7+F	510402.495	2D

Supplementary Table 4: Convergence and mixing diagnostics for two chains run under the CAT+GTR+G4 model, based on the 35-gene dataset. These chains recovered 2D trees with maximal support. The PhyloBayes manual (v.1.8; 3/6/2019) suggests that maximum discrepancies between chains <0.3 and effective sample sizes >50 are acceptable; <0.1 and ESS >300 are good. The values for all of the statistics were within acceptable limits for all of the analyses. We ran four additional chains, all of which also returned 2D consensus trees (not shown).

Statistic	Value: maximum discrepancy between chains [effective sample size, where relevant]

Maximum split discrepancy between chains (bpcomp maxdiff)	0.0843386
Mean split discrepancy between chains (bpcomp meandiff)	0.00331858
Log likelihood	0.043462 [616]
Tree length	0.180278 [195]
Alpha	0.0422014 [490]
Nmode	0.0184889 [659]
Statent	0.0470897 [177]
Statalpha	0.0192547 [83]
Rrent	0.0914187 [899]
Rrmean	0.0162773 [19120]

Supplementary Table 5: Posterior predictive simulations to assess model adequacy on the 35-gene alignment of Da Cunha et al.[2]. Neither the LG+G4 nor CAT+GTR+G4 model are adequate on the amino acid data, although CAT+GTR+G4 performs very substantially better in all tests (smaller Z-scores) except for that comparing mean-squared across-branch compositional heterogeneity, where both models are inadequate. SR4 recoding substantially improved model fit, as judged by the magnitude of the Z-scores.

Test	Observed value	Predicted (simulated) value	*P*-value	Z-score
LG+G4+F				
Diversity (amino acids/site)	7.44	9.23 +/- 0.027	0	64.2
Empirical convergence probability	0.46	0.421 +/- 0.002	0	19.7
Across-site compositional heterogeneity	0.198	0.0179 +/- 0.0001	0	17.9
Maximum across-branch compositional heterogeneity	0.00356	0.000227 +/- 0.0000549	0	60.7
Mean-squared across-branch compositional heterogeneity	0.000878	0.0000801 +/- 0.00000467	0	170.6
CAT+GTR+G4				
Diversity (amino acids/site)	7.44	7.61+/- 0.02	0	6.9
Empirical convergence probability	0.46	0.444 +/- 0.002	0	7.62

Across-site compositional heterogeneity	0.198	0.190 +/- 0.0001	0	7.51
Maximum across-branch compositional heterogeneity	0.00356	0.000242 +/- 0.0000587	0	56.5
Mean-squared across-branch compositional heterogeneity	0.000878	0.0000792 +/- 0.00000496	0	160.8
CAT+GTR+G4, SR4-recoded alignment				
Diversity (amino acids/site)	2.73391	2.72411 +/- 0.00954356	0.826087	-1.02721
Empirical convergence probability	0.721743	0.714385 +/- 0.00179893	0	4.09023
Across-site compositional heterogeneity	0.105847	0.104195 +/- 0.00048824	0	3.38491
Maximum across-branch compositional heterogeneity	0.0017359	0.000404566 +/- 0.000182094	0	7.31124
Mean-squared across-branch compositional heterogeneity	0.000393838	5.11273e-05 +/- 8.99183e-06	0	38.1135

Supplementary Table 6: Tests for compositional homogeneity on the 35 genes of the Da Cunha et al. concatenation. For each gene, data were simulated on the ML tree under the homogeneous LG+G4+F model, and the value of the chi-square statistic for compositional homogeneity on the observed data was compared to the distribution from the simulated data. “AU-test preference” denotes genes that significantly ($P < 0.05$) favoured either a two-domain or three-domain tree in Da Cunha et al; 5/6 genes favouring the three-domains tree, and 6/10 genes favouring the two-domains tree showed significant compositional heterogeneity at $P < 0.05$.

Gene	**P-value**	AU-test preference (Da Cunha et al.)
arCOG00412	0	Three-domains
arCOG00415	0.045	
arCOG00785	0.001	
arCOG00987	0	
arCOG01183	0	
arCOG01227	0	Two-domains

arCOG01559	0	
arCOG01560	0	Two-domains
arCOG01722	0.003	
arCOG04064	0	Two-domains
arCOG04090	0.001	
arCOG04091	0.35	Two-domains
arCOG04092	0.017	Two-domains
arCOG04094	0	
arCOG04095	0.303	
arCOG04096	0.283	
arCOG04097	0	
arCOG04098	0.076	
arCOG04099	0.189	
arCOG04113	0.112	Two-domains
arCOG04121	0	Two-domains
arCOG04169	0	
arCOG04239	0.071	Two-domains
arCOG04240	0.273	Two-domains
arCOG04241	0.001	Three-domains
arCOG04242	0.03	
arCOG04243	0.004	
arCOG04245	0.137	
arCOG04254	0.257	Three-domains
arCOG04255	0.72	
arCOG04256	0	Three-domains
arCOG04257	0	Three-domains
arCOG04289	0	Three-domains
arCOG1228	0	Two-domains
arCOG1758	0.742	
arCOG1762	0	

Supplementary Table 7: Posterior predictive simulations to assess model adequacy for the LG+G4+F and CAT+GTR+G4 model on the concatenation of 6 genes that supported the 3D tree by AU-tests under the LG+G4+F model. Neither model provides an adequate fit to the data, although the fit of the model was improved by data recoding, as judged by the magnitude of Z-scores.

Test	Observed value	Predicted (simulated) value	*P*-value	Z-score
LG+G4+F				
Diversity (amino acids/site)	7.73613	9.35365 +/- 0.0555767	0	29.1042
Empirical convergence probability	0.45767	0.4158 +/- 0.0039834	0	10.5112
Across-site compositional heterogeneity	0.0197141	0.0176705 +/- 0.000205769	0	9.93191
Maximum across-branch compositional heterogeneity	0.00350825	0.00138755 +/- 0.000479882	0	4.41921
Mean-squared across-branch compositional heterogeneity	0.00113682	0.000344241 +/- 2.28481e-05	0	34.6892
CAT+GTR+G4				
Diversity (amino acids/site)	7.73613	7.92194 +/- 0.050484	0	3.6806
Empirical convergence probability	0.45767	0.434329 +/- 0.00427655	0	5.45811
Across-site compositional heterogeneity	0.0197141	0.0185872 +/- 0.000218861	0	5.14885
Maximum across-branch compositional heterogeneity	0.00350825	0.00131525 +/- 0.000431892	0	5.07766
Mean-squared across-branch compositional heterogeneity	0.00113682	0.000333159 +/- 2.13148e-05	0	37.7046
SR4-recoded alignment, CAT+GTR+G4				
Diversity (amino acids/site)	2.80506	2.79327 +/- 0.0187914	0.735979	-0.627658
Empirical	0.714808	0.703498 +/-	0	3.21478

convergence probability		0.00351809		
Across-site compositional heterogeneity	0.103907	0.101822 +/- 0.000928771	0.0069025	2.24455
Maximum across-branch compositional heterogeneity	0.0022102	0.00104565 +/- 0.000479991	0.0301984	2.42619
Mean-squared across-branch compositional heterogeneity	0.000492719	0.000179406 +/- 3.04399e-05	0	10.2928

Supplementary Table 8: Posterior predictive simulations to assess model adequacy for the LG+G4+F and CAT+GTR+G4 models on the RNA polymerase alignment, and for CAT+GTR+G4 on the SR4-recoded version of the alignment. None of the models provide adequate model fit. The CAT+GTR+G4 fit to the SR4-recoded alignment performs best, as judged by the magnitude of the Z-scores, providing an adequate fit for all tests except for mean across-branch compositional heterogeneity. This analysis recovered a weakly-supported 2D tree.

Test	Observed value	Predicted (simulated) value	*P*-value	Z-score
LG+Gamma(4)+F				
Diversity (amino acids/site)	7.80618	9.60455 +/- 0.0575006	0	31.2757
Empirical convergence probability	0.453555	0.423213 +/- 0.00457267	0	6.63543
Across-site compositional heterogeneity	0.0196181	0.0181525 +/- 0.000232338	0	6.30772
Maximum across-branch compositional heterogeneity	0.00215123	0.000778035 +/- 0.000120288	0	11.4159
Mean-squared across-branch compositional heterogeneity	0.00070898	0.000346737 +/- 2.97392e-05	0	12.1807
CAT+GTR+G4				
Diversity (amino acids/site)	7.80618	8.02098 +/- 0.0576918	0	3.72317
Empirical convergence	0.453555	0.454086 +/- 0.00545943	0.540778	-0.0972681

probability				
Across-site compositional heterogeneity	0.0196181	0.0195244 +/- 0.00027283	0.357591	0.343145
Maximum across-branch compositional heterogeneity	0.00215123	0.000754315 +/- 0.000120968	0	11.5478
Mean-squared across-branch compositional heterogeneity	0.00070898	0.000328024 +/- 2.92218e-05	0	13.0367
CAT+GTR+G4, SR4-recoded alignment				
Diversity (amino acids/site)	2.75743	2.75731 +/- 0.0197719	0.511287	-0.0063465
Empirical convergence probability	0.713136	0.720579 +/- 0.00461626	0.936795	-1.6124
Across-site compositional heterogeneity	0.104389	0.106106 +/- 0.00114295	0.936795	-1.50265
Maximum across-branch compositional heterogeneity	0.000986868	0.000748139 +/- 0.000220275	0.13544	1.08378
Mean-squared across-branch compositional heterogeneity	0.00033653	0.000167961 +/- 4.01539e-05	0.00225734	4.19807

Supplementary Table 9: Posterior predictive simulations to evaluate model fit on the 43-gene archaea and eukaryotes dataset. Neither model provides adequate fit, although the model fit is improved on the recoded dataset, as judged by the magnitude of the Z-scores. Both analyses place eukaryotes within the Asgard archaea, consistent with a 2D tree.

Test	**Observed value**	**Predicted (simulated) value**	***P*-value**	**Z-score**
CAT+GTR+G4				
Diversity (amino acids/site)	6.66709	6.73003 +/- 0.017717	0	3.55293
Empirical convergence probability	0.508746	0.51224 +/- 0.00165878	0.982301	-2.10634
Across-site compositional heterogeneity	0.0223834	0.0224932 +/- 8.21361e-05	0.902655	-1.33706

Maximum across-branch compositional heterogeneity	0.00312705	0.00100232 +/- 0.000234698	0	9.05305
Mean-squared across-branch compositional heterogeneity	0.000715297	9.65774e-05 +/- 5.93711e-06	0	104.212
4-state Susko-Roger recoding, CAT+GTR+G4				
Diversity (amino acids/site)	2.57089	2.56442 +/- 0.007	0.830046	-0.921615
Empirical convergence probability	0.745193	0.74834 +/- 0.00146901	0.978756	-2.14254
Across-site compositional heterogeneity	0.113623	0.113991 +/- 0.000359605	0.854325	-1.0221
Maximum across-branch compositional heterogeneity	0.00311862	0.00157559 +/- 0.000503644	0.0106222	3.06372
Mean-squared across-branch compositional heterogeneity	0.000390961	9.2275e-05 +/- 1.23945e-05	0	24.0982

Supplementary Table 10: Comparison of universal marker gene sets. OMA 21: A 21-gene dataset inferred using the OMA orthology method in this study. Spang 2015: 35 genes from Spang et al. (2015), also analyzed elsewhere in this paper. Cox 2008: 29 gene dataset originally assembled by Cox et al.[6] but used and updated in a series of follow-up studies[7–9]. IDs are arbitrary labels specific to this study. Family 4 is EF2, and was manually removed from the analyses prior to concatenation.

ID	OMA 21	Spang 2015	Cox 2008	Annotation
1				Ribosomal protein S16 (40S subunit)
2				Ribosomal protein L11 (60S)
3				Ribosomal protein S3 (40S)
4				Elongation factor 2 (**Excluded**)
5				RNA-binding signal recognition particle subunit SRP54
6				Ribosomal protein S23 (40S)
7				DNA-directed RNA polymerase 1 core subunit RPA135
8				Translocon subunit Sec61
9				Ribosomal protein S22 (40S)
10				tRNA N6-adenosine threonylcarbamoyltransferase (Kae1p)
11				Ribosomal protein S15 (40S)
12				Ribosomal protein L23

13				Ribosomal protein L16
14				Ribosomal protein S20
15				FeS cluster-binding ribosome biosynthesis/translation initiation protein Rli1
16				Signal recognition particle receptor subunit alpha
17				Ribosomal protein S14
18				DNA-directed RNA polymerase I core subunit RPA190
19				Ribosomal protein S5
20				V-type ATP synthase subunit B (H+ transporting)
21				Ribosomal protein S0
22				Translation elongation factor EF-1 subunit alpha
23				Ribosomal protein S18
24				Ribosomal protein S2
25				Phenylalanine-tRNA ligase subunit beta
26				Translation initiation factor eIF5B
27				Ribosomal protein L1
28				Recombinase Rad51
29				Ribosomal protein L2
30				Aspartate-tRNA ligase
31				Replication factor C subunit 2
32				Methionine aminopeptidase
33				Ribosomal protein L12
34				Chaperonin CCT5
35				Ribosomal protein L35A
36				YDL140Cp-like protein; RNA polymerase II subuniit
37				Cbf5
38				Ribosomal protein L17A
39				Ribosomal protein L26A
40				Ribosomal protein S11B
41				Ribosomal protein L10
42				RNA polymerase II core subunit RPB3
43				Ribosomal protein L9
44				Ribosomal protein S9B
45				ribonuclease H2 catalytic subunit RNH201
46				Peptidase M50 family protein
47				V-type ATPase subunit A
48				Translation initiation factor 2 subunit gamma
49				Tryptophan-tRNA ligase
50				Leucine-tRNA ligase
51				Glutamine-tRNA ligase

Supplementary Table 11: Phylogenomics provides a consistent signal for the 2D tree. This table summarises the analyses of protein concatenations described in the main text. ML: maximum likelihood; PP: Bayesian posterior probability; LBA: long-branch attraction; SR4: 4-state data recoding, as described in the main text.

Dataset	Model	Rationale	Tree	Support
35 genes, 81 taxa[2]	LG+G4+F	Model used previously[22]	-	ML tree is 2D (74% bootstrap), but no significant difference between 2D and 3D trees (AU = 0.229 for 3D)
35 genes, 81 taxa	LG+C60+G4+F	Best-fitting ML model; approximation to CAT+GTR that can be fit using ML	2D	2D (97% bootstrap); 3D tree rejected by AU test (AU = 0.0036)
35 genes, 81 taxa	CAT+GTR+G4	Bayesian method that accounts for site-specific compositions, less susceptible to LBA	2D	PP = 1 for eukaryotes+Asgards
35 genes, 81 taxa	CAT+GTR+G4, SR4 recoded data	SR4 recoding improved model fit	2D	PP = 0.98 for eukaryotes+Heimdallarchaeota
35 genes, 81 taxa	NDCH2, SR4-recoded data	Accounts for branch-specific sequence compositions, another potential source of LBA	2D	PP = 0.88 for eukaryotes+Asgards, PP = 0.85 for eukaryotes+Heimdallarchaeota
12 compositionally-homogeneous genes	Partitioned, LG+G4+F per partition	An analysis of the least heterogeneous subset of the 35-gene dataset	2D	PP = 1 for eukaryotes+TACK Archaea/Asgard; PP = 0.67 for eukaryotes+*Lokiarchaeum*
12 compositionally-homogeneous genes	CAT+GTR+G4	An analysis of the least heterogeneous subset of the 35-gene dataset	2D	PP = 1 for eukaryotes+TACK Archaea/Asgard; PP = 0.91 for eukaryotes+Asgard; PP = 0.77 for eukaryotes+*Lokiarchaeum*
35 genes, 81 taxa	GHOST (LG+G4+F mixture)	Accounts for heterotachy	2D	58% bootstrap support for eukaryotes+Asgards
35 genes, 81 taxa	GHOST (GTR+G4 mixture on SR4-recoded data)	SR4 recoding improved model fit	2D	97% bootstrap support for eukaryotes+TACK Archaea/Asgards; 95% bootstrap support for eukaryotes+Heimdallarchaeota

6 genes favouring 3D tree under LG+G4+F according to AU-test[2]	CAT+GTR+G4	Test of multiple histories in the core gene set	3D	PP = 0.87 for archaeal monophyly
6 genes favouring 3D tree under LG+G4+F according to AU-test	CAT+GTR+G4, SR4-recoded data	SR4 recoding improved model fit	2D	PP = 0.8 for eukaryotes+*Heimdallarchaeota LC_2*
21 genes, 125 taxa	CAT+GTR+G4	Expanded sampling of microbial diversity	2D	PP = 1 for eukaryotes+Asgards; PP = 0.81 for eukaryotes+Heimdallarchaeota
21 genes, 125 taxa	CAT+GTR+G4, SR4-recoded data	SR4 recoding improved model fit	2D	PP = 1 for eukaryotes+Asgards; PP = 0.79 for eukaryotes+Heimdallarchaeota
43 genes, 92 taxa	CAT+GTR+G4	Subset of the 125-taxon dataset focusing on eukaryotes and archaea	N/A	PP = 1 for eukaryotes+Heimdallarchaeota
43 genes, 92 taxa	CAT+GTR+G4, SR4-recoded data	SR4 recoding improved model fit	N/A	PP = 1 for eukaryotes+Heimdallarchaeota; PP = 0.95 for eukaryotes+*Heimdallarchaeota LC_3*
7 genes, 125 taxa	LG+C60+G4+F	An alignment comprised only of genes shared between three different “universal gene” sets	2D	77% bootstrap support for eukaryotes+Asgards
27 genes, 125 taxa	LG+C60+G4+F	An alignment comprised of genes found in at least ⅔ different “universal gene” sets	2D	99% bootstrap support for eukaryotes+Asgards; 76% bootstrap for eukaryotes+Heimdallarchaeota
35 genes, 125 taxa	LG+C60+G4+F	An alignment comprised of the genes found in the 35-gene dataset described above	2D	88% bootstrap support for eukaryotes+TACK; 76% bootstrap support for eukaryotes+Asgards
50 genes, 125 taxa	LG+C60+G4+F	An alignment of all genes present in three different “universal gene” sets	2D	100% bootstrap support for eukaryotes+Heimdallarchaeota

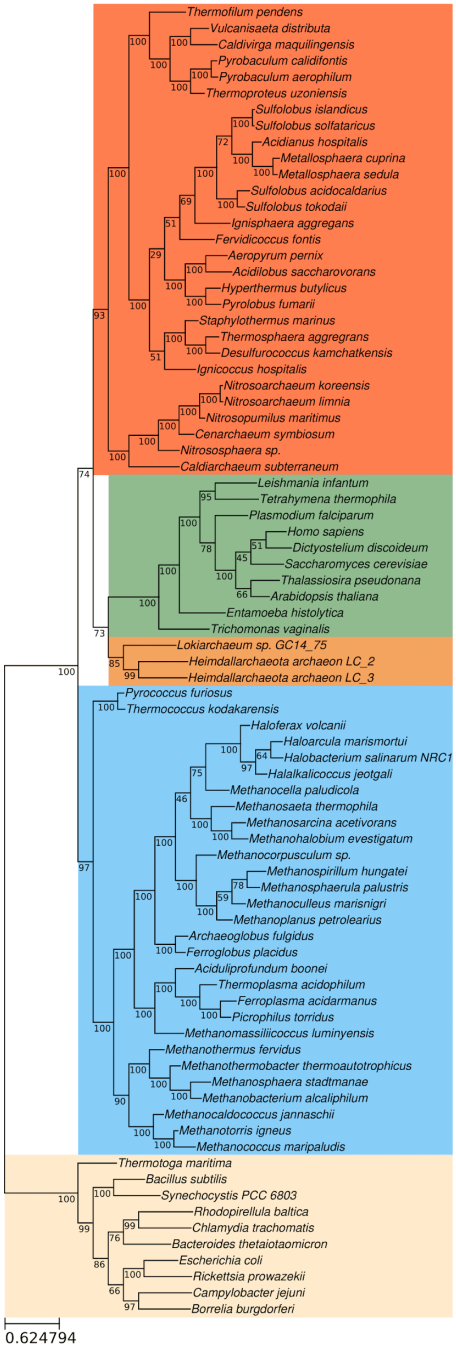
Supplementary Table 12: Dissection of the phylogenetic signal in 3199 single-copy gene trees in archaea and eukaryotes. The phylogenetic signal in these trees is summarised in the supertree analyses presented in the main text, but we also provide a summary of the support for different hypotheses of the archaeal sister group of eukaryotes here. Numbers refer to numbers of ML single gene trees; consider the first line of the table as an example: for 126 out of 242 trees where the distinction can be made based on inspection of the topology alone, eukaryotes form a clan with TACKL Archaea (either as their sister, or within the TACKL). Out of those 242 trees, the eukaryotes formed a clan in 124, and the TACKL group formed a clan in 64. “TACKL” is a term for the clan including Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota and Asgardarchaeota. “Usable trees” denotes the number of trees for which it was possible to distinguish, on the basis of the ML topology, a sister group relationship between eukaryotes and a given archaeal lineage from competing alternatives. This requires (i) at least two members of the candidate archaeal sister group, (ii) at least two eukaryotes and one archaeon from outside the candidate sister group or, alternatively, one eukaryote and at least two archaea from outside the candidate sister group. In each case, it requires assuming that the root of the tree is outside the clan comprising eukaryotes and the candidate sister group.

Archaeal group	Usable trees	Eukaryote+archaeal group clan	Eukaryote clan	Archaeal group clan
TACKL	242	126	124	64
Asgardarchaeota	218	70	66	53
Heimdallarchaeota	164	47	44	39
Thaumarchaeota	172	40	38	34
Aigarchaeota	99	11	11	11
Crenarchaeota	222	49	47	40
Korarchaeota	103	13	13	13
Euryarchaeota	243	81	78	61

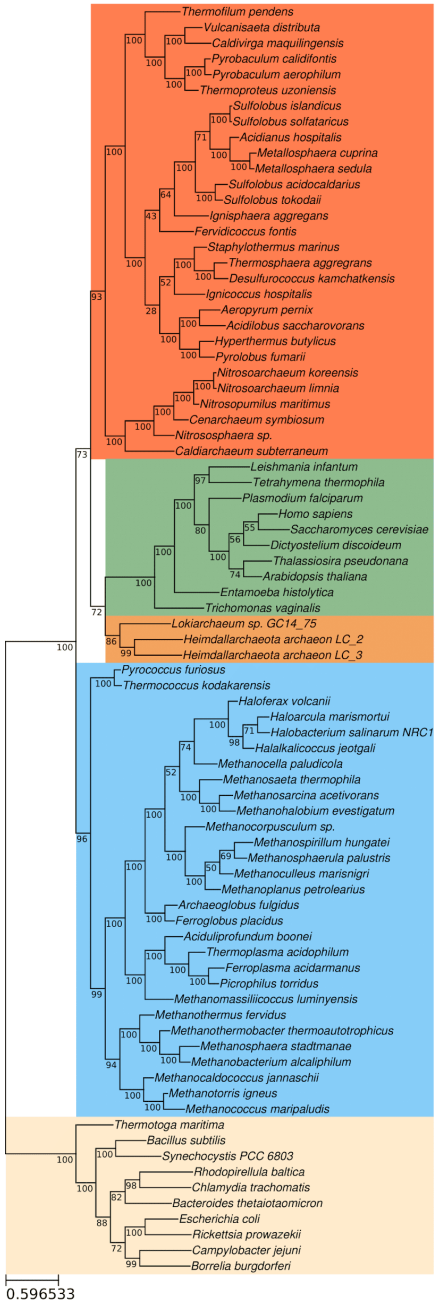
Supplementary Figures

Supplementary Figure 1: Maximum likelihood phylogenies inferred from the 35-gene concatenation of Da Cunha et al.[2] ML trees inferred under the (a) LG+G4+F, (b) LG+R8+F; (c) LG+C60+G4+F; (d) LG+PMSF+R8 models. Analyses were performed in IQ-Tree 1.6.2. The model in (a) was that used by Da Cunha et al.; (b) was the optimal single-matrix model according to the BIC criterion; (c) and (d) are more parameter-rich models that account for across-site compositional heterogeneity in an ML context. In (d), posterior mean site frequency patterns were estimated from the C60 empirical mixture model. Support for the 2D tree increases with improved model fit, whether judged by bootstrap support or AU-tests (see main text). Branch supports are rapid bootstraps computed with the UFBoot2 algorithm. Eukaryotes in green, TACK Archaea in dark orange, Asgard archaea in orange, Euryarchaeota in blue, Bacteria in beige. Branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar. To increase the legibility of support values very short branch lengths have been increased slightly in these supplemental visualisations; the Newick representations with the original branch lengths are included in the data supplement.

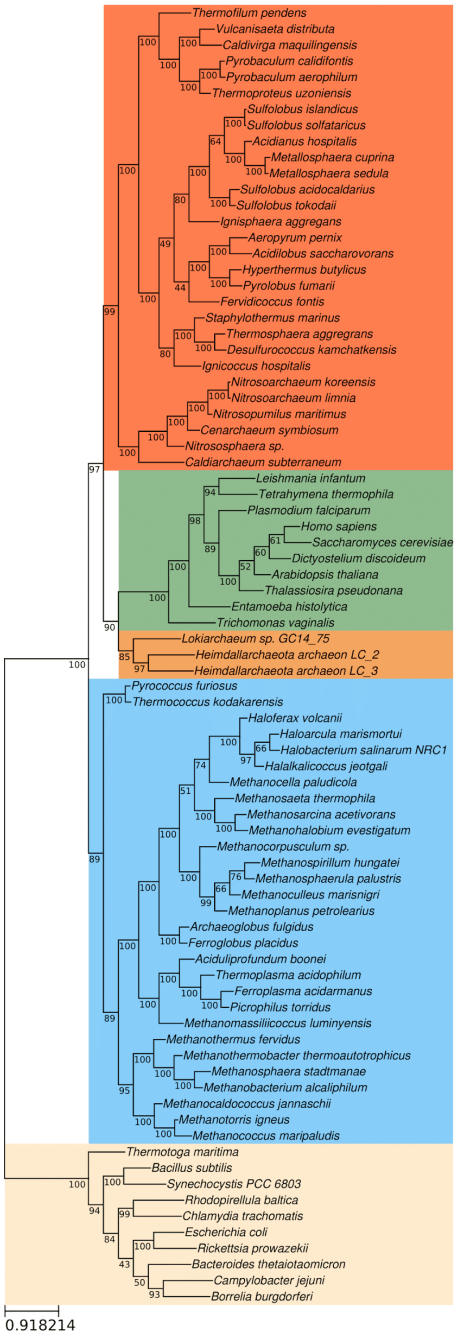
(a)



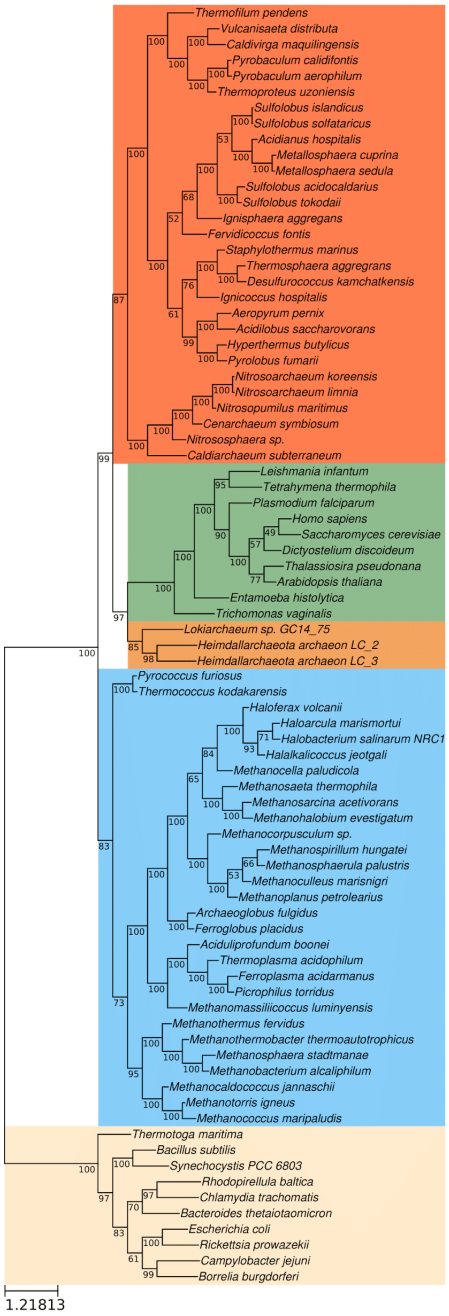
(b)



(c)

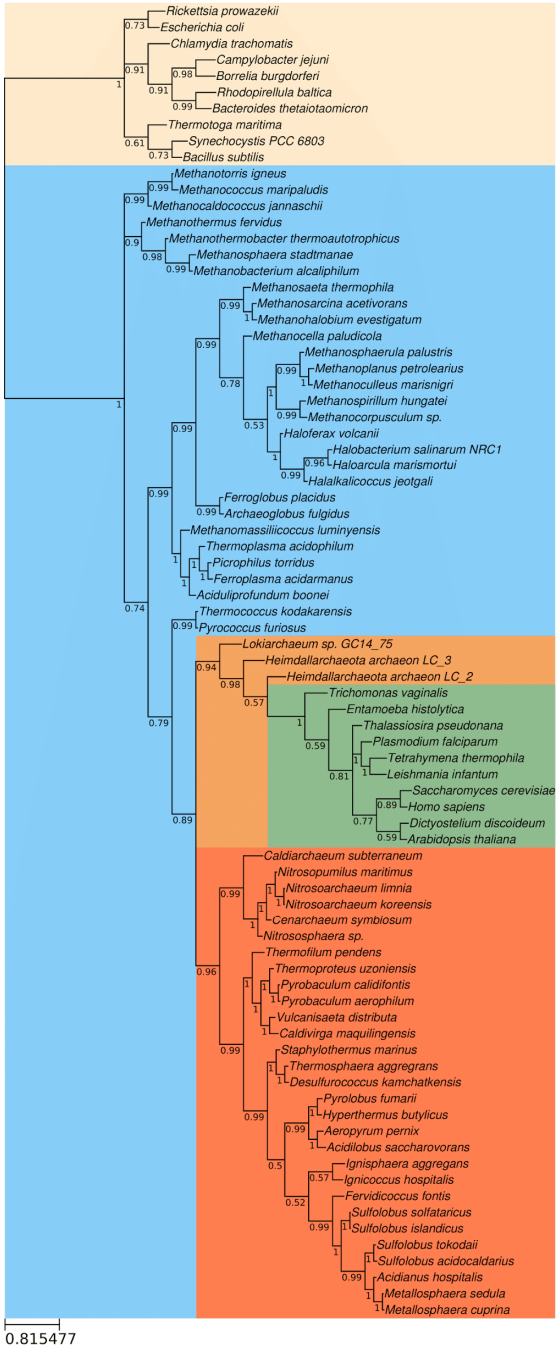

Thermofilum pendens
Vulcanisaeta distributa
Caldivirga maquilingensis
Pyrobaculum calidifontis
Pyrobaculum aerophilum
Thermoproteus uzoniensis
Sulfolobus islandicus
Sulfolobus solfataricus
Acidianus hospitalis
Metallosphaera cuprina
Metallosphaera sedula
Sulfolobus acidocaldarius
Sulfolobus tokodaii
Ignisphaera aggregans
Aeropyrum pernix
Acidilobus saccharovorans
Hyperthermus butylicus
Pyrolobus fumarii
Fervidicoccus fontis
Staphylothermus marinus
Thermosphaera aggregrans
Desulfurococcus kamchatkensis
Ignicoccus hospitalis
Nitrosoarchaeum koreensis
Nitrosoarchaeum limnia
Nitrosopumilus maritimus
Cenarchaeum symbiosum
Nitrososphaera sp.
Caldiarchaeum subterraneum
Leishmania infantum
Tetrahymena thermophila
Plasmodium falciparum
Homo sapiens
Saccharomyces cerevisiae
Dictyostelium discoideum
Arabidopsis thaliana
Thalassiosira pseudonana
Entamoeba histolytica
Trichomonas vaginalis
Lokiarchaeum sp. GC14_75
Heimdallarchaeota archaeon LC_2
Heimdallarchaeota archaeon LC_3
Pyrococcus furiosus
Thermococcus kodakarensis
Haloferax volcanii
Haloarcula marismortui
Halobacterium salinarum NRC1
Halalkalicoccus jeotgali
Methanocella paludicola
Methanosaeta thermophila
Methanosarcina acetivorans
Methanohalobium evestigatum
Methanocorpusculum sp.
Methanospirillum hungatei
Methanosphaerula palustris
Methanoculleus marisnigri
Methanoplanus petrolearius
Archaeoglobus fulgidus
Ferroglobus placidus
Aciduliprofundum boonei
Thermoplasma acidophilum
Ferroplasma acidarmanus
Picrophilus torridus
Methanomassiliicoccus luminyensis
Methanothermus fervidus
Methanothermobacter thermoautotrophicus
Methanosphaera stadtmanae
Methanobacterium alcaliphilum
Methanocaldococcus jannaschii
Methanotorris igneus
Methanococcus maripaludis
Thermotoga maritima
Bacillus subtilis
Synechocystis PCC 6803
Rhodopirellula baltica
Chlamydia trachomatis
Escherichia coli
Rickettsia prowazekii
Bacteroides thetaiotaomicron
Campylobacter jejuni
Borrelia burgdorferi
0.918214

(d)

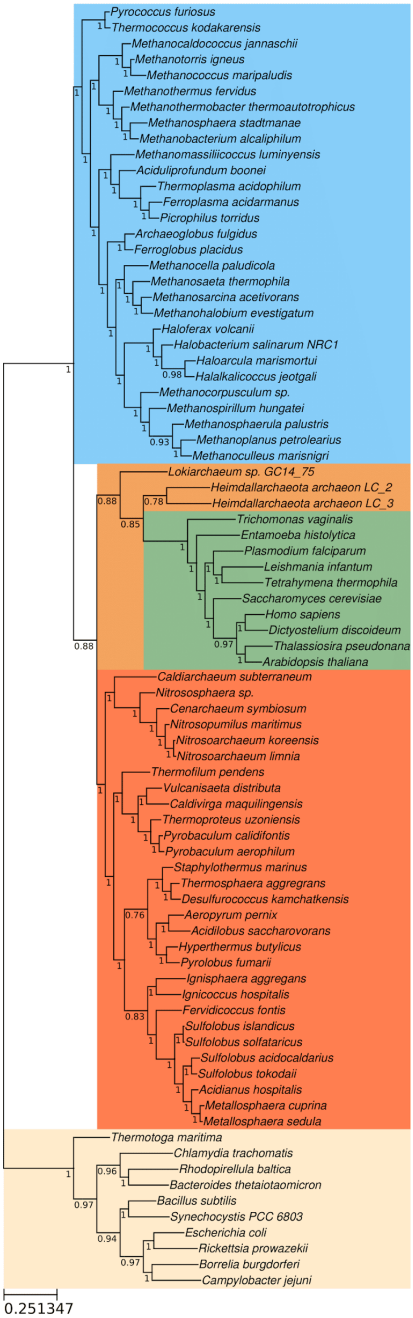

Thermofilum pendens
Vulcanisaeta distributa
Caldivirga maquilingensis
Pyrobaculum calidifontis
Pyrobaculum aerophilum
Thermoproteus uzoniensis
Sulfolobus islandicus
Sulfolobus solfataricus
Acidianus hospitalis
Metallosphaera cuprina
Metallosphaera sedula
Sulfolobus acidocaldarius
Sulfolobus tokodaii
Ignisphaera aggregans
Fervidicoccus fontis
Staphylothermus marinus
Thermosphaera aggregrans
Desulfurococcus kamchatkensis
Ignicoccus hospitalis
Aeropyrum pernix
Acidilobus saccharovorans
Hyperthermus butylicus
Pyrolobus fumarii
Nitrosoarchaeum koreensis
Nitrosoarchaeum limnia
Nitrosopumilus maritimus
Cenarchaeum symbiosum
Nitrososphaera sp.
Caldiarchaeum subterraneum
Leishmania infantum
Tetrahymena thermophila
Plasmodium falciparum
Homo sapiens
Saccharomyces cerevisiae
Dictyostelium discoideum
Thalassiosira pseudonana
Arabidopsis thaliana
Entamoeba histolytica
Trichomonas vaginalis
Lokiarchaeum sp. GC14_75
Heimdallarchaeota archaeon LC_2
Heimdallarchaeota archaeon LC_3
Pyrococcus furiosus
Thermococcus kodakarensis
Haloferax volcanii
Haloarcula marismortui
Halobacterium salinarum NRC1
Halalkalicoccus jeotgali
Methanocella paludicola
Methanosaeta thermophila
Methanosarcina acetivorans
Methanohalobium evestigatum
Methanocorpusculum sp.
Methanospirillum hungatei
Methanosphaerula palustris
Methanoculleus marisnigri
Methanoplanus petrolearius
Archaeoglobus fulgidus
Ferroglobus placidus
Aciduliprofundum boonei
Thermoplasma acidophilum
Ferroplasma acidarmanus
Picrophilus torridus
Methanomassiliicoccus luminyensis
Methanothermus fervidus
Methanothermobacter thermoautotrophicus
Methanosphaera stadtmanae
Methanobacterium alcaliphilum
Methanocaldococcus jannaschii
Methanotorris igneus
Methanococcus maripaludis
Thermotoga maritima
Bacillus subtilis
Synechocystis PCC 6803
Rhodopirellula baltica
Chlamydia trachomatis
Bacteroides thetaiotaomicron
Escherichia coli
Rickettsia prowazekii
Campylobacter jejuni
Borrelia burgdorferi
1.21813

Supplementary Figure 2: Bayesian phylogenies of the SR4-recoded 35-gene concatenation of Da Cunha et al.[2]. (a) Inference under CAT+GTR+G4 in PhyloBayes-MPI; (b) Inference under NDCH2 in p4. Branch supports are posterior probabilities. Eukaryotes in green, TACK Archaea in dark orange, Asgard archaea in orange, Euryarchaeota in blue, Bacteria in beige. Branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar. To increase the legibility of support values very short branch lengths have been increased slightly in these supplemental visualisations; the Newick representations with the original branch lengths are included in the data supplement.

(a)

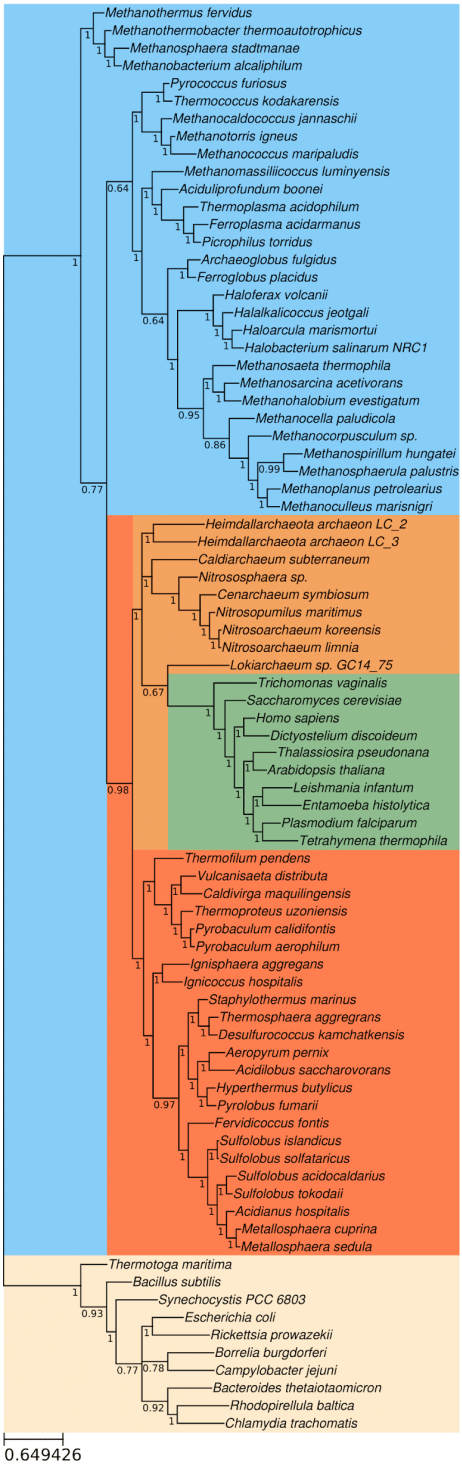


(b)

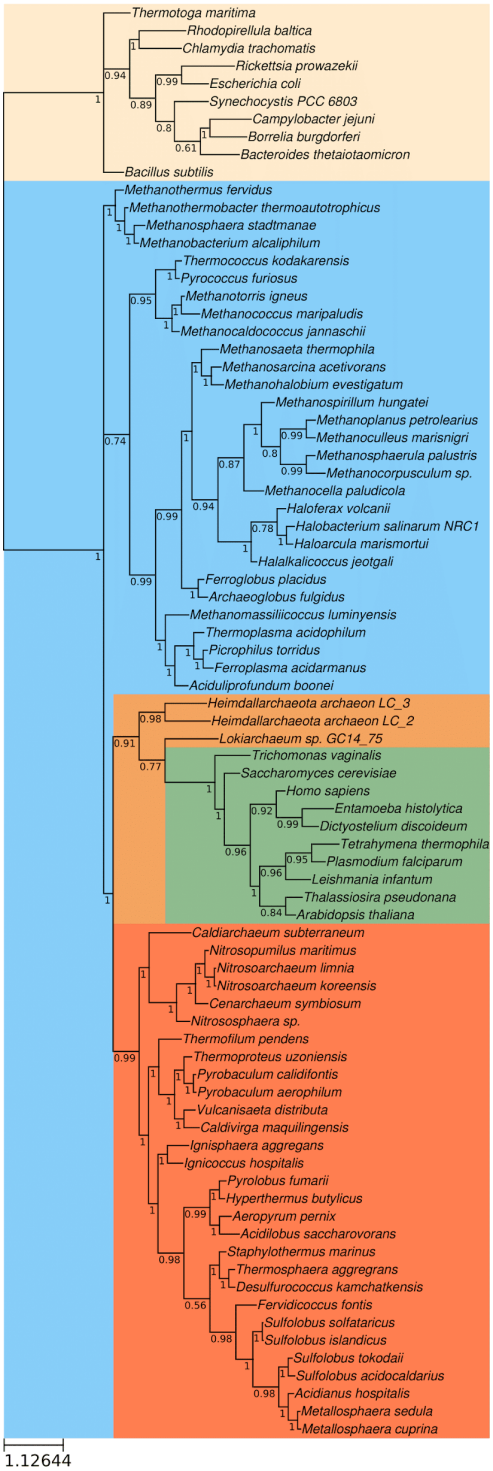


Supplementary Figure 3: Bayesian phylogenies of the 12 composition-homogeneous genes in the 35-gene concatenation. (a) Consensus tree inferred under a partition model, in which each gene was fit with an LG+G4+F model; alpha shape parameters and empirical frequencies were sampled independently for each partition. (b) Consensus tree inferred under the CAT+GTR+G4 model. Branch supports are posterior probabilities. Eukaryotes in green, TACK Archaea in dark orange, Asgard archaea in orange, Euryarchaeota in blue, Bacteria in beige. Branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar. To increase the legibility of support values very short branch lengths have been increased slightly in these supplemental visualisations; the Newick representations with the original branch lengths are included in the data supplement.

(a)

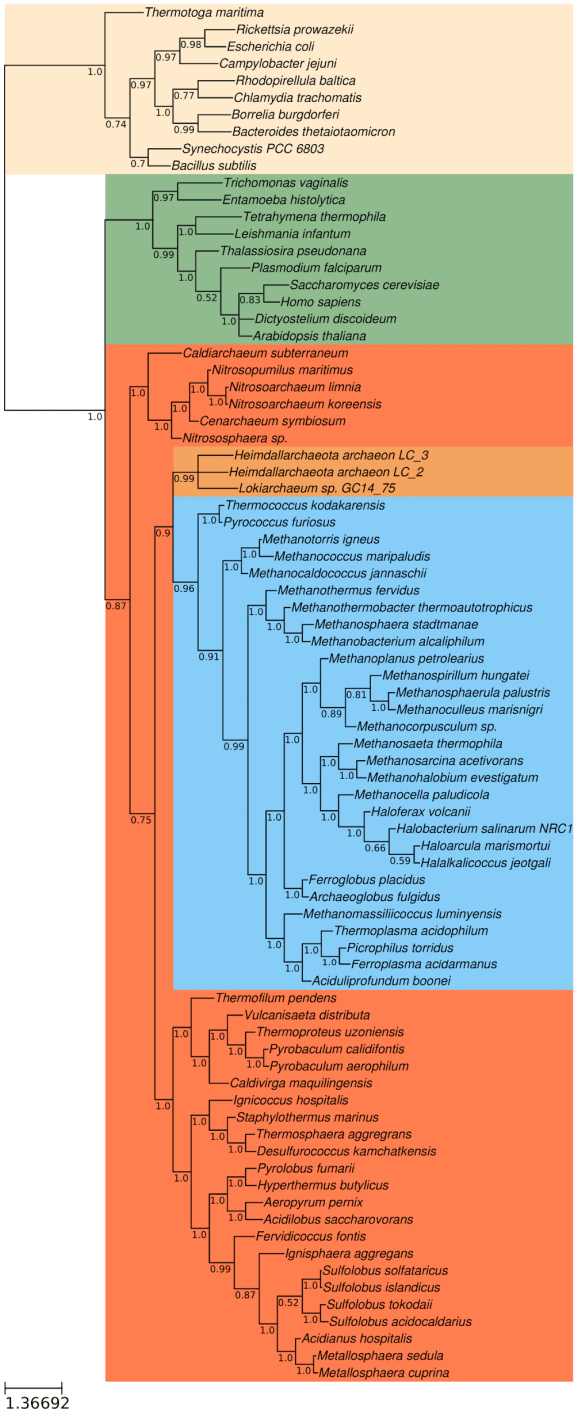


(b)

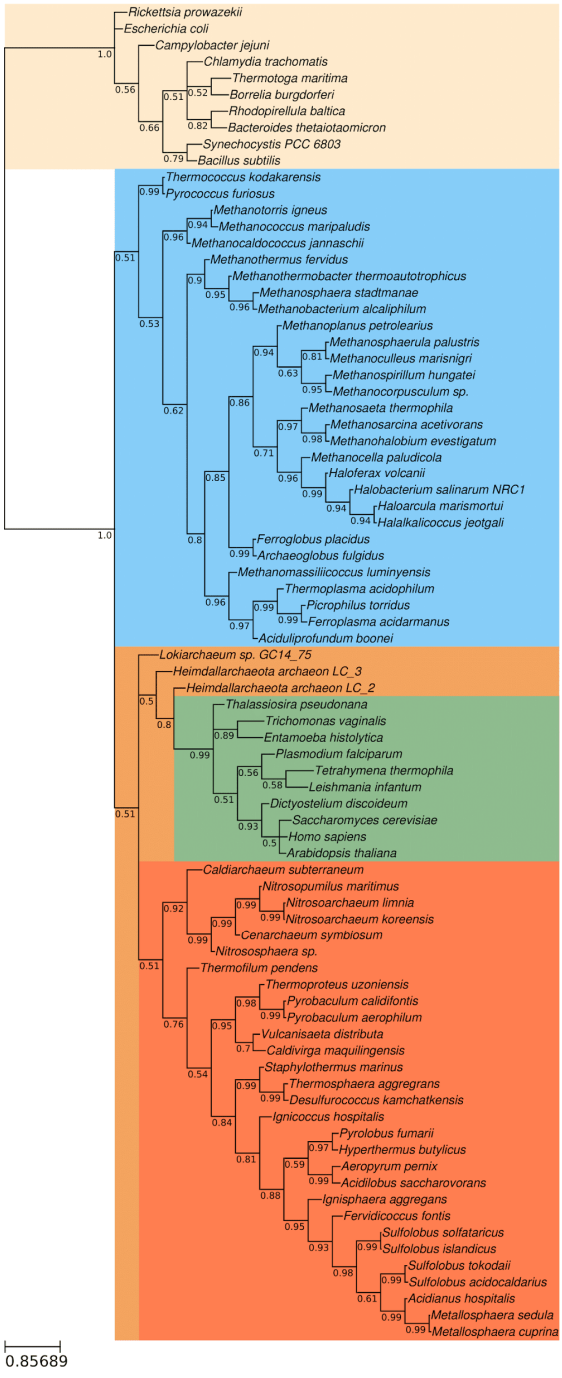


Supplementary Figure 4: Bayesian phylogeny of a concatenated alignment of 6 genes that provided significant support for the three-domains tree when analysed under LG+G4+F in^2. The tree was inferred under the CAT+GTR+G4 model. (a) A moderately supported 3D tree was inferred from the original amino acid alignment; (b) A weakly supported 2D tree was inferred from an SR4-recoded alignment. Eukaryotes in green, TACK Archaea in dark orange, Asgard archaea in orange, Euryarchaeota in blue, Bacteria in beige. Branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar. To increase the legibility of support values very short branch lengths have been increased slightly in these supplemental visualisations; the Newick representations with the original branch lengths are included in the data supplement.

(a)

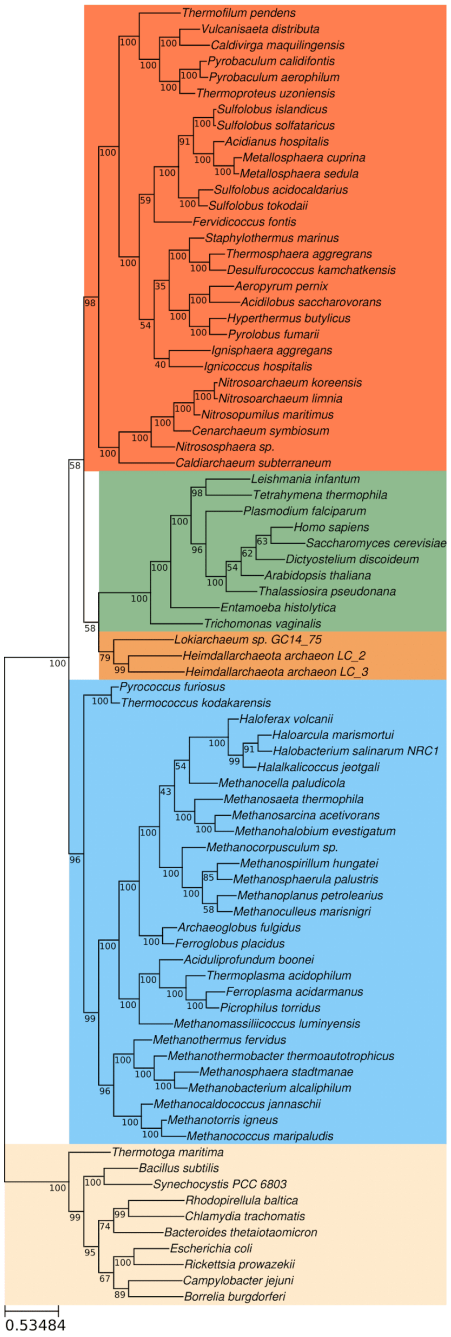


(b)

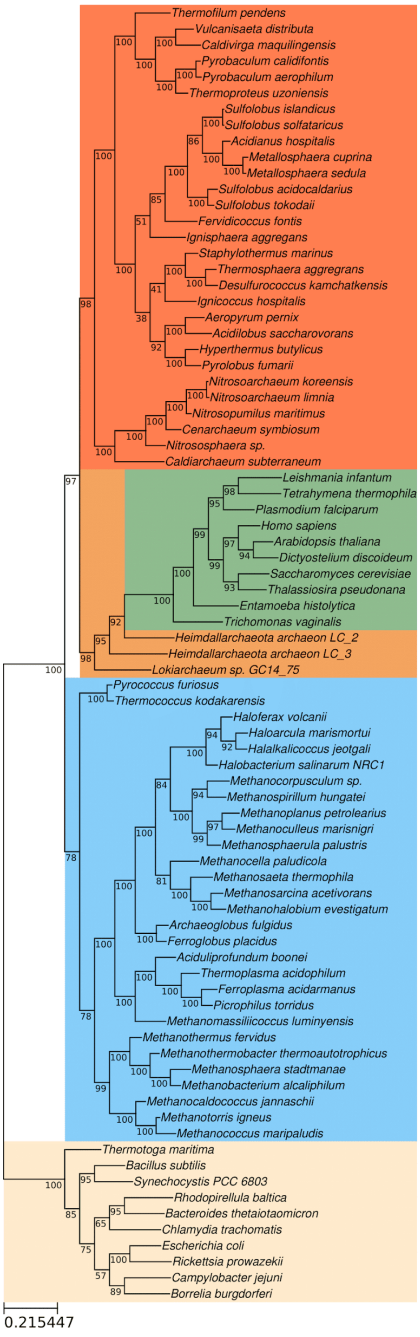


Supplementary Figure 5: Analysis of the 35-gene concatenation under the GHOST model of IQ-Tree, an edge-unlinked mixture model that accommodates heterotachy by allowing branch lengths, sequence compositions and exchangeabilities to vary across the sites of the alignment. (a) ML tree inference under a four-component LG+F model on the amino acid alignment. (b) ML tree inference under a four component GTR+F model on the SR4-recoded alignment. Each component has its own equilibrium frequencies and branch lengths on a shared underlying topology; in the GTR case, the exchangeabilities are also free to vary among components. Eukaryotes in green, TACK Archaea in dark orange, Asgard archaea in orange, Euryarchaeota in blue, Bacteria in beige. Branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar. To increase the legibility of support values very short branch lengths have been increased slightly in these supplemental visualisations; the Newick representations with the original branch lengths are included in the data supplement.

(a)



(b)

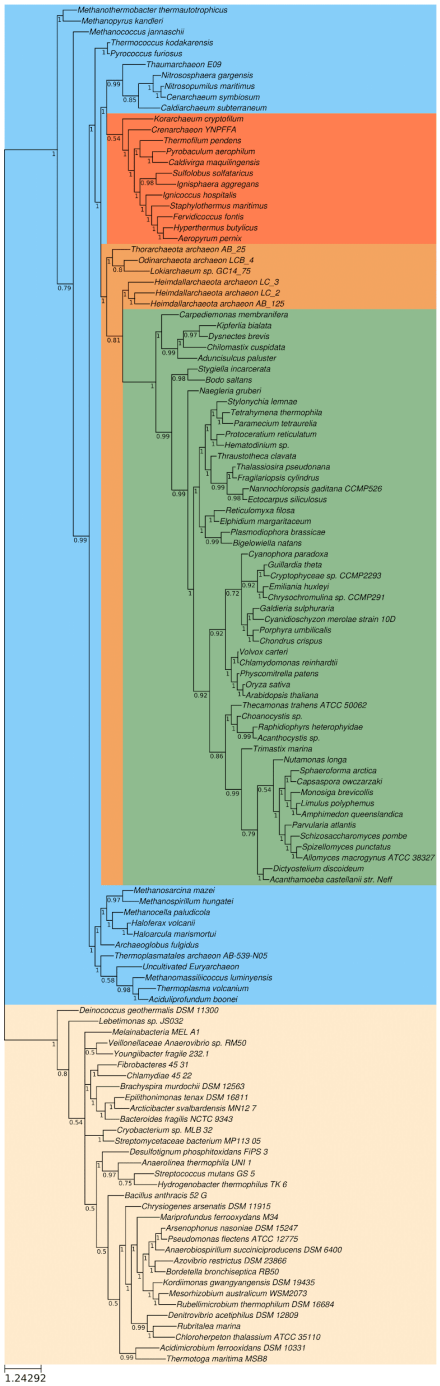


Supplementary Figure 6: Bayesian phylogenies of RNA polymerase subunits analyzed under the (a) LG+G4+F, (b) CAT+GTR+G4, and (c) CAT+GTR+G4 model in combination with SR4 recoding. Branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar. To increase the legibility of support values very short branch lengths have been increased slightly in these supplemental visualisations; the Newick representations with the original branch lengths are included in the data supplement.

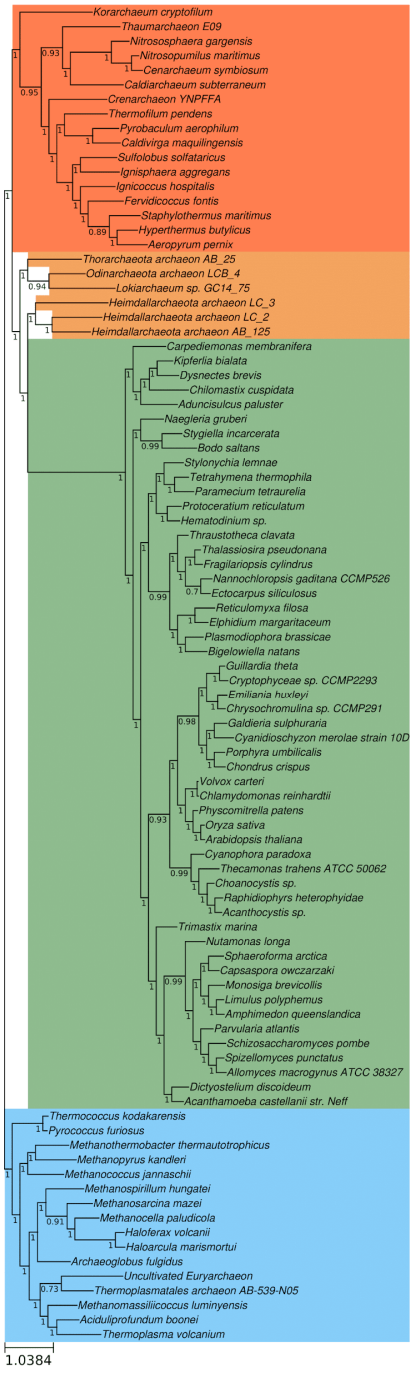
(a) (b) (c)

Supplementary Figure 7: Bayesian phylogenies of (a) 21 broadly-conserved genes in bacteria, archaea and eukaryotes; (b) 43 genes conserved in archaea and eukaryotes. In both analyses, eukaryotes form a clade with Heimdallarchaeota with moderate (PP = 0.81, (a)) to maximal (PP = 1, (b)) support. Trees inferred under the CAT+GTR+G4 model in PhyloBayes-MPI 1.8. Branch supports are posterior probabilities. Eukaryotes in green, TACK Archaea in dark orange, Asgard archaea in orange, Euryarchaeota in blue, Bacteria in beige. Branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar. To increase the legibility of support values very short branch lengths have been increased slightly in these supplemental visualisations; the Newick representations with the original branch lengths are included in the data supplement.

(a)

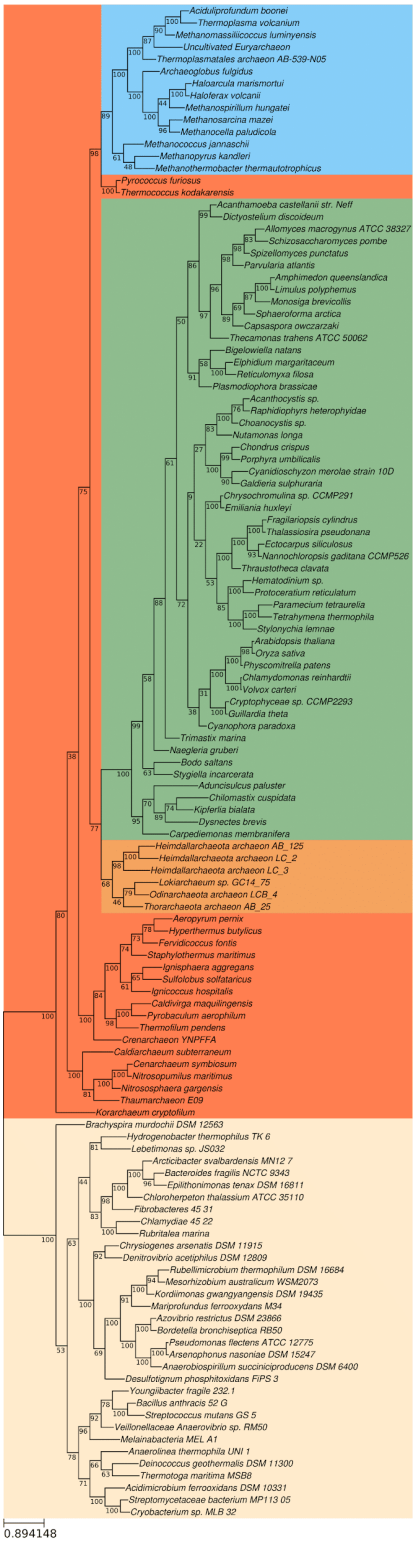


(b)

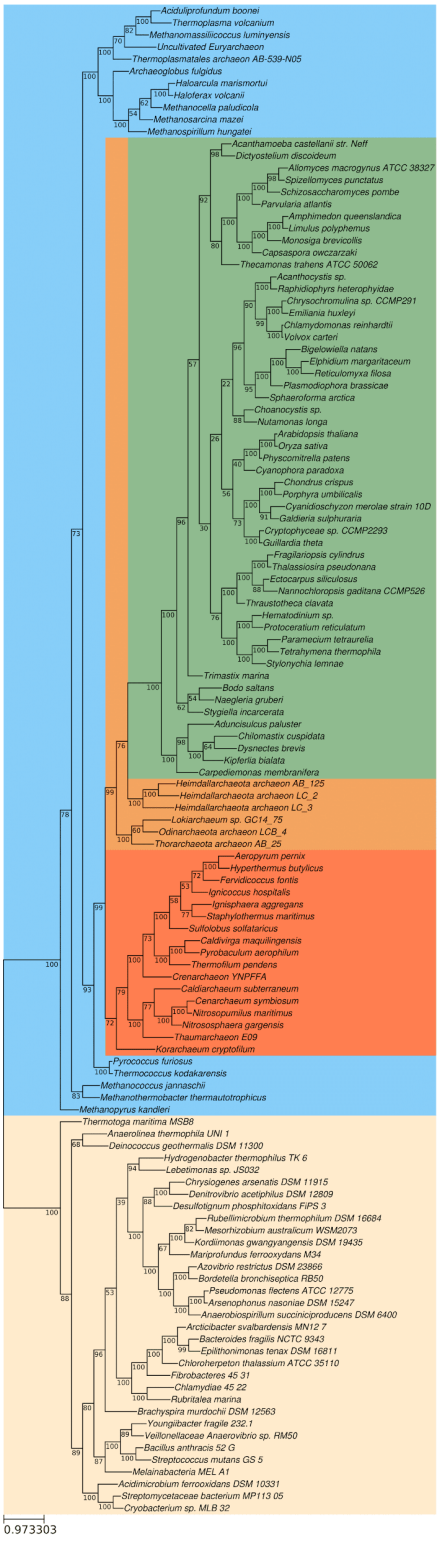


Supplementary Figure 8: Maximum likelihood phylogenies of the (a) 7-, (b) 27-, (c) 35- and (d) 50-gene concatenations. The 7-, 27- and 50-gene datasets represent genes found in all, at least two, or at least one of several recently analyzed core gene sets. The 35-gene set uses the same gene sampling as the 35-gene Spang dataset, but with updated taxon sampling. Eukaryotes in green, Asgard archaea in orange, TACK Archaea in dark orange, Euryarchaeota in blue, Bacteria in beige.The ML tree under the best-fitting substitution model was inferred for each concatenation, using 1000 UFBoot2 bootstraps as supports. In all cases, the LG+C60+G4+F model was chosen by the BIC criterion.

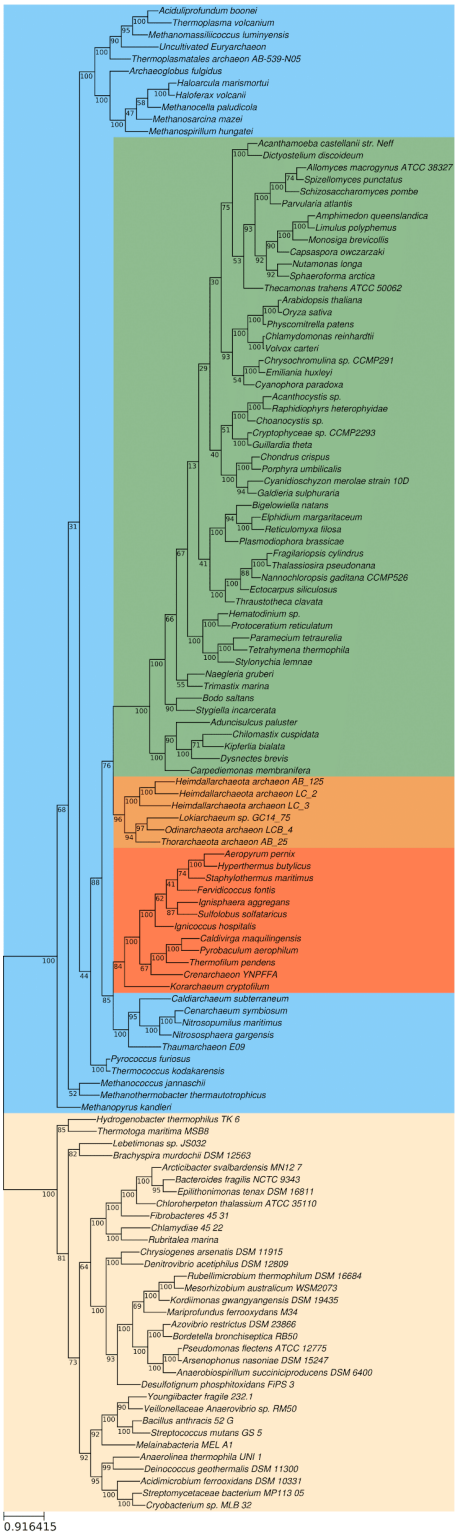
(a)



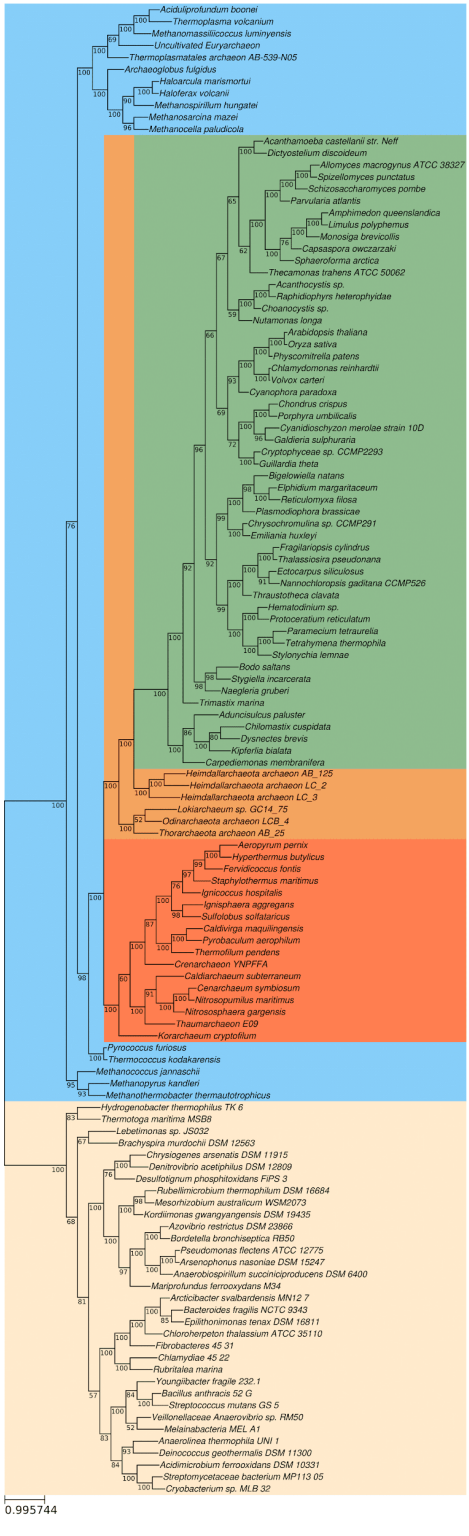
(b)



(c)

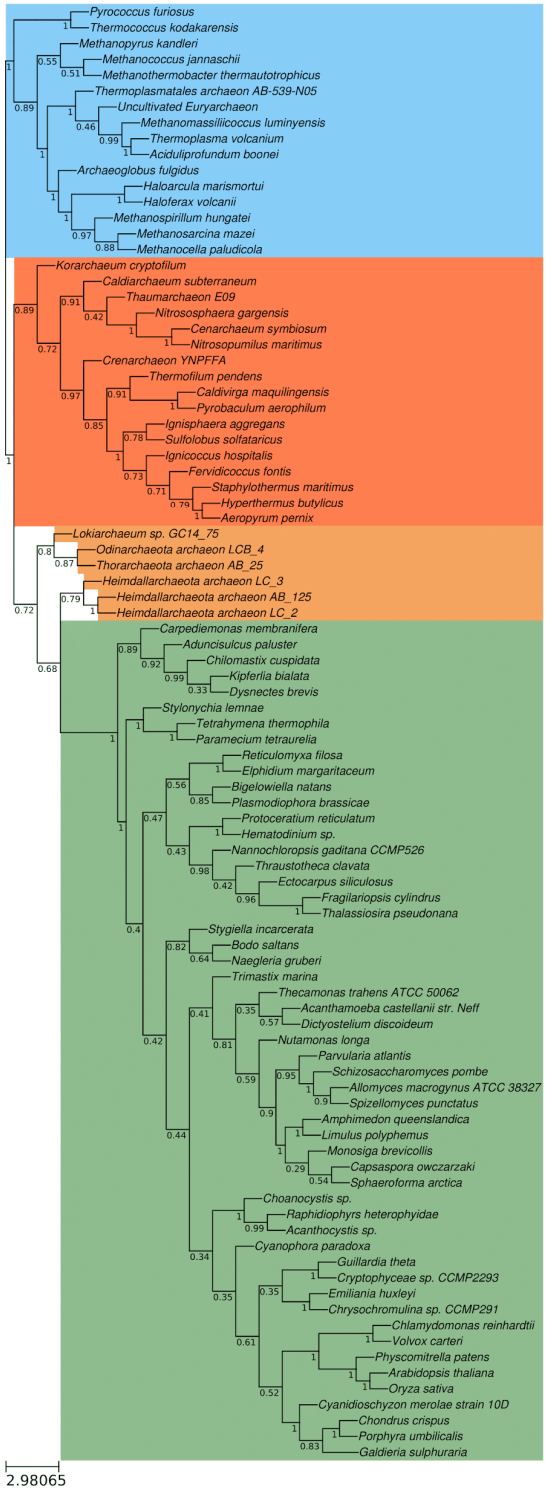

Aciduliprofundum boonei
Thermoplasma volcanium
Methanomassiliicoccus luminyensis
Uncultivated Euryarchaeon
Thermoplasmatales archaeon AB-539-N05
Archaeoglobus fulgidus
Haloarcula marismortui
Haloferax volcanii
Methanocella paludicola
Methanosarcina mazei
Methanospirillum hungatei
Acanthamoeba castellanii str. Neff
Dictyostelium discoideum
Allomyces macrogynus ATCC 38327
Spizellomyces punctatus
Schizosaccharomyces pombe
Parvularia atlantis
Amphimedon queenslandica
Limulus polyphemus
Monosiga brevicollis
Capsaspora owczarzaki
Nutamonas longa
Sphaeroforma arctica
Thecamonas trahens ATCC 50062
Arabidopsis thaliana
Oryza sativa
Physcomitrella patens
Chlamydomonas reinhardtii
Volvox carteri
Chrysochromulina sp. CCMP291
Emiliania huxleyi
Cyanophora paradoxa
Acanthocystis sp.
Raphidiophyrs heterophyidae
Choanocystis sp.
Cryptophyceae sp. CCMP2293
Guillardia theta
Chondrus crispus
Porphyra umbilicalis
Cyanidioschyzon merolae strain 10D
Galdieria sulphuraria
Bigelowiella natans
Elphidium margaritaceum
Reticulomyxa filosa
Plasmodiophora brassicae
Fragilariopsis cylindrus
Thalassiosira pseudonana
Nannochloropsis gaditana CCMP526
Ectocarpus siliculosus
Thraustotheca clavata
Hematodinium sp.
Protoceratium reticulatum
Paramecium tetraurelia
Tetrahymena thermophila
Stylonychia lemnae
Naegleria gruberi
Trimastix marina
Bodo saltans
Stygiella incarcerata
Aduncisulcus paluster
Chilomastix cuspidata
Kipferlia bialata
Dysnectes brevis
Carpediemonas membranifera
Heimdallarchaeota archaeon AB_125
Heimdallarchaeota archaeon LC_2
Heimdallarchaeota archaeon LC_3
Lokiarchaeum sp. GC14_75
Odinarchaeota archaeon LCB_4
Thorarchaeota archaeon AB_25
Aeropyrum pernix
Hyperthermus butylicus
Staphylothermus maritimus
Fervidicoccus fontis
Ignisphaera aggregans
Sulfolobus solfataricus
Ignicoccus hospitalis
Caldivirga maquilingensis
Pyrobaculum aerophilum
Thermofilum pendens
Crenarchaeon YNPFFA
Korarchaeum cryptofilum
Caldiarchaeum subterraneum
Cenarchaeum symbiosum
Nitrosopumilus maritimus
Nitrososphaera gargensis
Thaumarchaeon E09
Pyrococcus furiosus
Thermococcus kodakarensis
Methanococcus jannaschii
Methanothermobacter thermautotrophicus
Methanopyrus kandleri
Hydrogenobacter thermophilus TK 6
Thermotoga maritima MSB8
Lebetimonas sp. JS032
Brachyspira murdochii DSM 12563
Arcticibacter svalbardensis MN12 7
Bacteroides fragilis NCTC 9343
Epilithonimonas tenax DSM 16811
Chloroherpeton thalassium ATCC 35110
Fibrobacteres 45 31
Chlamydiae 45 22
Rubritalea marina
Chrysiogenes arsenatis DSM 11915
Denitrovibrio acetiphilus DSM 12809
Rubellimicrobium thermophilum DSM 16684
Mesorhizobium australicum WSM2073
Kordiimonas gwangyangensis DSM 19435
Mariprofundus ferrooxydans M34
Azovibrio restrictus DSM 23866
Bordetella bronchiseptica RB50
Pseudomonas flectens ATCC 12775
Arsenophonus nasoniae DSM 15247
Anaerobiospirillum succiniciproducens DSM 6400
Desulfotignum phosphitoxidans FiPS 3
Youngiibacter fragile 232.1
Veillonellaceae Anaerovibrio sp. RM50
Bacillus anthracis 52 G
Streptococcus mutans GS 5
Melainabacteria MEL A1
Anaerolinea thermophila UNI 1
Deinococcus geothermalis DSM 11300
Acidimicrobium ferrooxidans DSM 10331
Streptomycetaceae bacterium MP113 05
Cryobacterium sp. MLB 32
0.916415

(d)

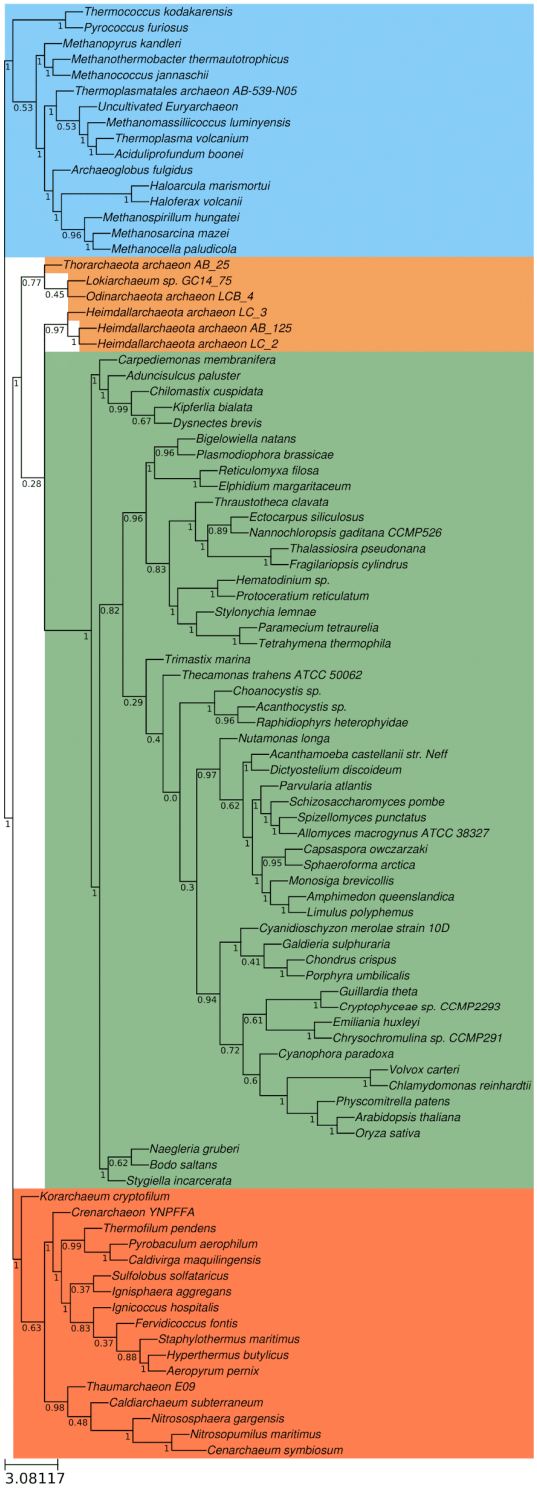

Aciduliprofundum boonei
Thermoplasma volcanium
Methanomassiliicoccus luminyensis
Uncultivated Euryarchaeon
Thermoplasmatales archaeon AB-539-N05
Archaeoglobus fulgidus
Haloarcula marismortui
Haloferax volcanii
Methanospirillum hungatei
Methanosarcina mazei
Methanocella paludicola
Acanthamoeba castellanii str. Neff
Dictyostelium discoideum
Allomyces macrogynus ATCC 38327
Spizellomyces punctatus
Schizosaccharomyces pombe
Parvularia atlantis
Amphimedon queenslandica
Limulus polyphemus
Monosiga brevicollis
Capsaspora owczarzaki
Sphaeroforma arctica
Thecamonas trahens ATCC 50062
Acanthocystis sp.
Raphidiophyrs heterophyidae
Choanocystis sp.
Nutamonas longa
Arabidopsis thaliana
Oryza sativa
Physcomitrella patens
Chlamydomonas reinhardtii
Volvox carteri
Cyanophora paradoxa
Chondrus crispus
Porphyra umbilicalis
Cyanidioschyzon merolae strain 10D
Galdieria sulphuraria
Cryptophyceae sp. CCMP2293
Guillardia theta
Bigelowiella natans
Elphidium margaritaceum
Reticulomyxa filosa
Plasmodiophora brassicae
Chrysochromulina sp. CCMP291
Emiliania huxleyi
Fragilariopsis cylindrus
Thalassiosira pseudonana
Ectocarpus siliculosus
Nannochloropsis gaditana CCMP526
Thraustotheca clavata
Hematodinium sp.
Protoceratium reticulatum
Paramecium tetraurelia
Tetrahymena thermophila
Stylonychia lemnae
Bodo saltans
Stygiella incarcerata
Naegleria gruberi
Trimastix marina
Aduncisulcus paluster
Chilomastix cuspidata
Dysnectes brevis
Kipferlia bialata
Carpediemonas membranifera
Heimdallarchaeota archaeon AB_125
Heimdallarchaeota archaeon LC_2
Heimdallarchaeota archaeon LC_3
Lokiarchaeum sp. GC14_75
Odinarchaeota archaeon LCB_4
Thorarchaeota archaeon AB_25
Aeropyrum pernix
Hyperthermus butylicus
Fervidicoccus fontis
Staphylothermus maritimus
Ignicoccus hospitalis
Ignisphaera aggregans
Sulfolobus solfataricus
Caldivirga maquilingensis
Pyrobaculum aerophilum
Thermofilum pendens
Crenarchaeon YNPFFA
Caldiarchaeum subterraneum
Cenarchaeum symbiosum
Nitrosopumilus maritimus
Nitrososphaera gargensis
Thaumarchaeon E09
Korarchaeum cryptofilum
Pyrococcus furiosus
Thermococcus kodakarensis
Methanococcus jannaschii
Methanopyrus kandleri
Methanothermobacter thermautotrophicus
Hydrogenobacter thermophilus TK 6
Thermotoga maritima MSB8
Lebetimonas sp. JS032
Brachyspira murdochii DSM 12563
Chrysiogenes arsenatis DSM 11915
Denitrovibrio acetiphilus DSM 12809
Desulfotignum phosphitoxidans FiPS 3
Rubellimicrobium thermophilum DSM 16684
Mesorhizobium australicum WSM2073
Kordiimonas gwangyangensis DSM 19435
Azovibrio restrictus DSM 23866
Bordetella bronchiseptica RB50
Pseudomonas flectens ATCC 12775
Arsenophonus nasoniae DSM 15247
Anaerobiospirillum succiniciproducens DSM 6400
Mariprofundus ferrooxydans M34
Arcticibacter svalbardensis MN12 7
Bacteroides fragilis NCTC 9343
Epilithonimonas tenax DSM 16811
Chloroherpeton thalassium ATCC 35110
Fibrobacteres 45 31
Chlamydiae 45 22
Rubritalea marina
Youngiibacter fragile 232.1
Bacillus anthracis 52 G
Streptococcus mutans GS 5
Veillonellaceae Anaerovibrio sp. RM50
Melainabacteria MEL A1
Anaerolinea thermophila UNI 1
Deinococcus geothermalis DSM 11300
Acidimicrobium ferrooxidans DSM 10331
Streptomycetaceae bacterium MP113 05
Cryobacterium sp. MLB 32
0.995744

Supplementary Figure 9: Multispecies coalescent tree inferred using ASTRAL-III[10] from (a) 43 and (b) 3199 single-copy orthologous genes found in archaea and eukaryotes. Eukaryotes in green, Asgard archaea in orange, TACK Archaea in dark orange, Euryarchaeota in blue. Input single gene trees were inferred using IQ-Tree 1.6.2 under the optimal substitution model, chosen for each gene using the BIC criterion. In addition to the default candidate model set, model selection also considered a set of empirical profile models (C10-C60). Input tree branches were collapsed when supported by 10% bootstrap or lower, as recommended by the authors of ASTRAL. Branch supports are quartet support values[11]. To increase the legibility of support values very short branch lengths have been increased slightly in these supplemental visualisations; the Newick representations with the original branch lengths are included in the data supplement.

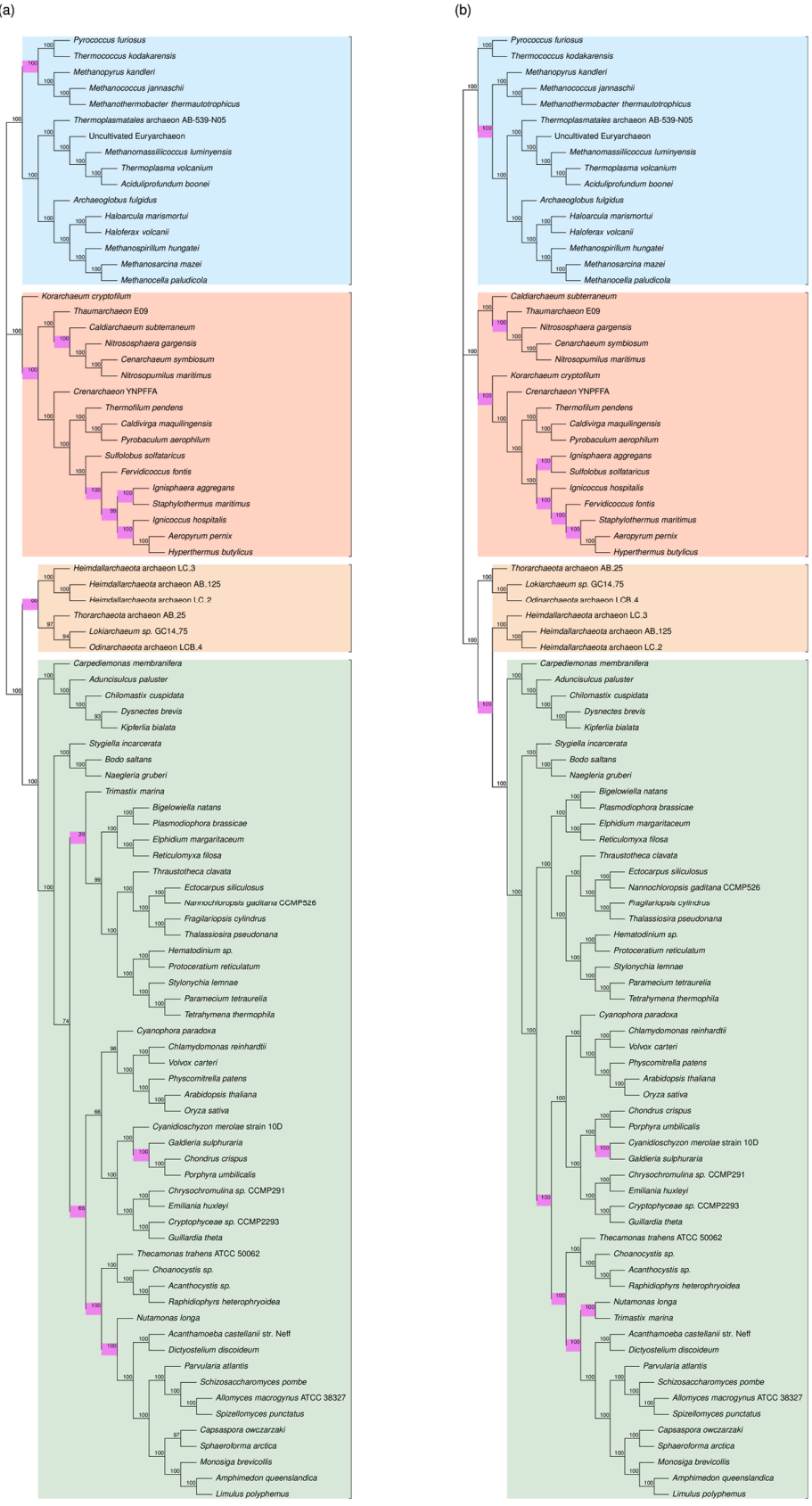
(a)



(b)



Supplementary Figure 10: Supertrees inferred from 3199 single-copy orthologous genes present in archaea and eukaryotes. (a) Bayesian MCMC with the model of Steel and Rodrigo[12], using symmetric distances and free beta. Support values are posterior probabilities as percent. (b) Bayesian MCMC with the SPA (Split Presence-Absence) model as implemented in p4[13]. Support values are posterior probabilities as percent. Euryarchaeota in blue, TACK Archaea in dark orange, Asgard archaea in orange, eukaryotes in green. Branches highlighted in violet were recovered with only one out of the two methods.



Supplementary Text

Does RNA polymerase support the three domains hypothesis?

One element of recent critiques of the two-domains tree has been the finding that phylogenies inferred from RNA polymerase support the three-domains tree[2]. In principle, RNA polymerase is a promising marker for deep phylogeny because the genes are universal and the largest subunits encode long proteins (>1000 amino acids), providing signal for phylogenetic reconstruction. Da Cunha et al. (2017) inferred phylogenies under the LG+G4+F and CAT+GTR+G4 models for a concatenation of three RNA polymerase subunits, both of which recovered a three-domains tree with 87/100 bootstrap for archaeal monophyly under LG, and PP = 0.94 under CAT+GTR+G4. The published analysis included most of the available RNA polymerase sequences from the Asgard archaea, with the exception of that from "Loki3" (now Heimdallarchaeon LC3); in the analyses that follow, we use the alignment of Da Cunha et al. but update it with the RNA polymerase sequences from this Asgard archaeon, which cluster with the other Asgard archaea in the resulting trees.

Phylogenies on the updated alignment inferred under the LG+G4+F and CAT+GTR+G4 models recover a strongly or moderately-supported three-domains tree: the support for archaeal monophyly is PP = 1 under LG+G4+F, and PP = 0.94 under CAT+GTR+G4 (Supplementary Figure 6). Posterior predictive simulations from these analyses indicate that the fit of both models is inadequate (Supplementary Table 8), although CAT+GTR+G4 fits better by comparison of Z-scores.

Since the most complex model available was inadequate, we explored data recoding (SR4, as above) as a means of ameliorating compositional heterogeneity in the RNA polymerase alignment. Analysis of the SR4-recoded alignment under CAT+GTR+G4 resulted in a loss of support for the three-domains tree (support for archaeal monophyly, PP = 0.01), while model fit was substantially improved compared to analyses on the unrecoded data (Supplementary Table 8). However, the position of eukaryotes within the Archaea was not well resolved; the position receiving the greatest support (PP = 0.81) places eukaryotes as sister to a clade comprising Asgard archaea, Euryarchaeota and Thaumarchaeota (Supplementary Figure 6).

Modelling protein fold presence-absence profiles to root the tree of life

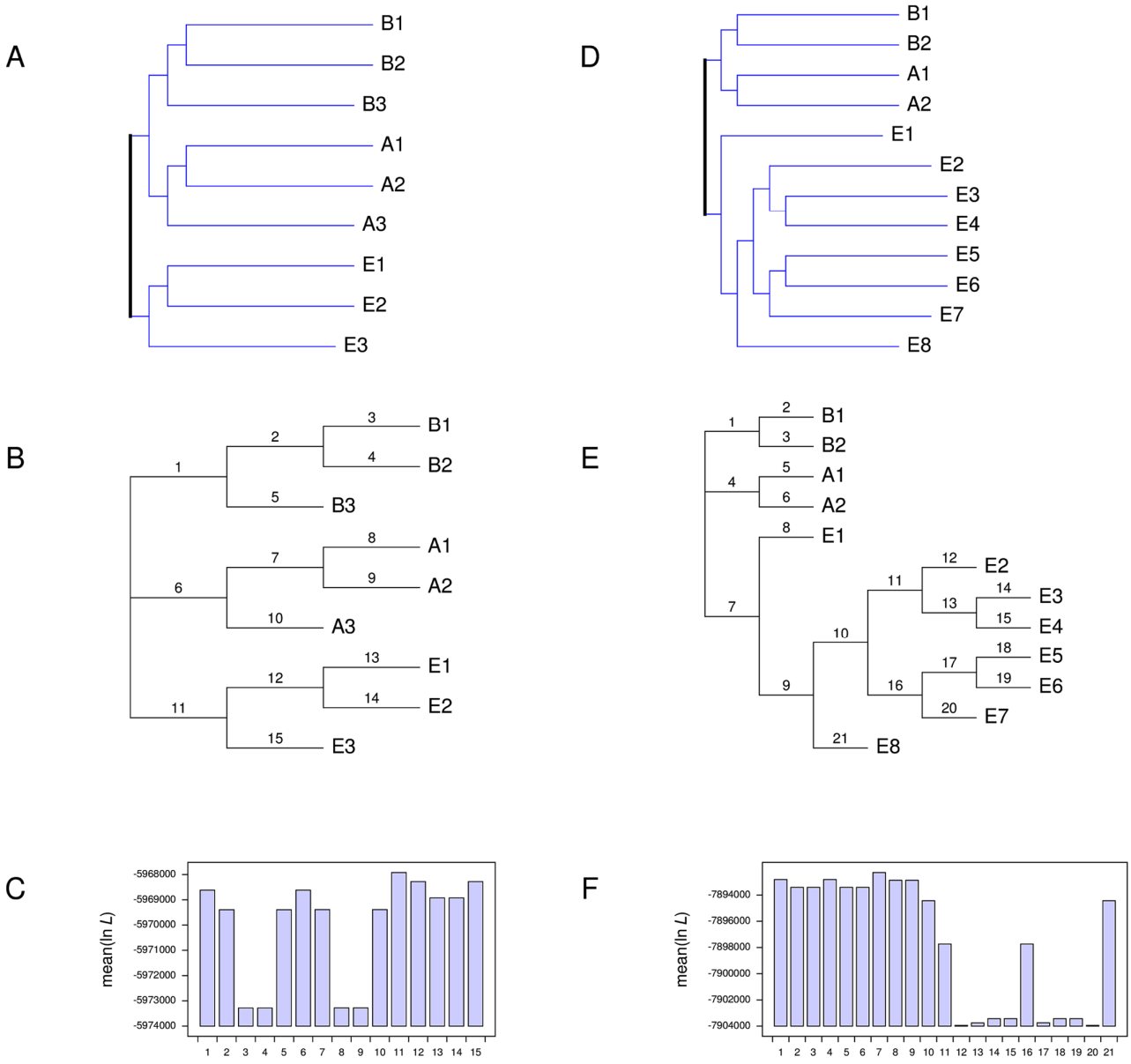
Tree topologies are usually inferred using stationary, reversible substitution models. These models assume that sequence composition does not change over time (stationarity) and that the probabilities of change on the tree do not depend on the direction of evolution (reversibility); they therefore lack a time dimension. The assumptions of stationarity and reversibility are not biologically motivated, but are simplifying assumptions that make the required calculations more tractable[14]. As a consequence, the resulting trees are formally unrooted (a root is not implied directly from the analysis), and must be oriented using some prior knowledge (e.g. outgroup designation), such as the assumption that the root lies on the bacterial branch. If the assumptions of stationarity or reversibility are relaxed, then the inferred trees will be intrinsically rooted in the sense that the probability of the data depends on the position of the root. A straightforward way to relax the assumption of stationarity is to have two different compositions, one at the root and one everywhere else on the tree. A model of this type (the KVR model) was proposed by Klopfstein et al. in the context of studying trends in hymenopteran morphological evolution[15], and this model was also used by Harish and Kurland (HK)[16] to infer a rooted tree of bacteria, archaea, and eukaryotes based on a binary matrix of protein fold presence/absence data.

As reported by HK, analysis of the protein fold data matrix (kindly supplied by Ajith Harish) under the 2-composition model (one composition at the root, one elsewhere) resulted in a root between prokaryotes and eukaryotes. We also obtained the same inference under a 3-composition model, in which the subtrees on either side of the root were fit with two different compositions (data not shown). As noted by HK, MCMC mixing is

challenging for this data and model; we therefore used maximum likelihood to compare support for all 277 possible roots on the consensus unrooted topology. To do so, we rooted on all 277 branches of the tree, optimised model parameters using maximum likelihood, and then compared the resulting likelihoods. The root position obtained in[16] --- on the branch separating eukaryotes and prokaryotes --- was the most likely using this approach.

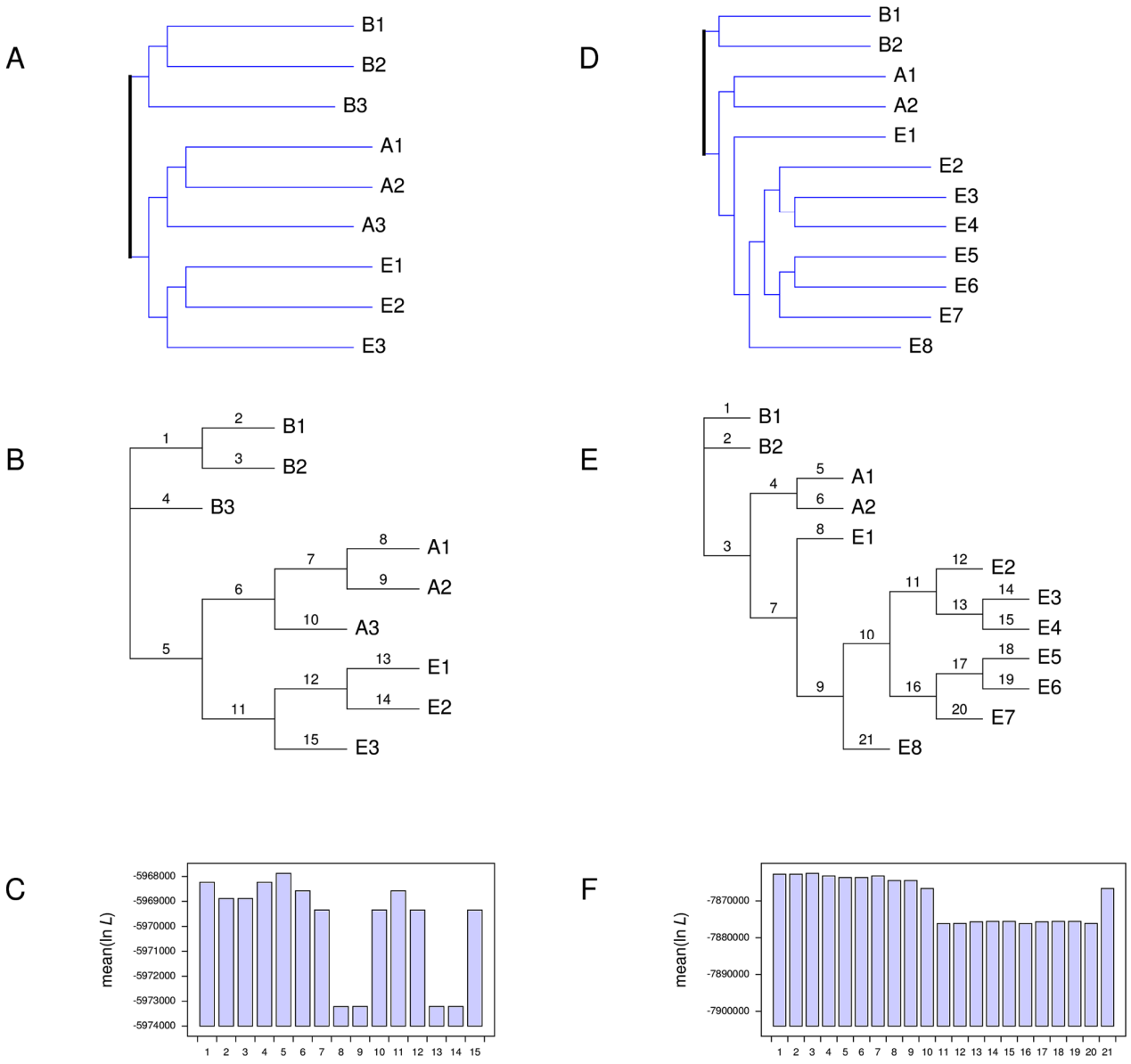
Simulations to evaluate the reliability of root inference from fold data using the KVR model

We next evaluated the ability of the KVR model to infer the root of the tree on simulated data, where the root is known. To evaluate the impact of tree shape, all the simulations were conducted on three different trees: a tree in which the root lay between eukaryotes and prokaryotes (Supplementary Figure 11), a 3D-like tree (Supplementary Figure 12), and a 2D-like tree (Supplementary Figure 13). When data were simulated under KVR (one composition at the root, a second composition everywhere else), the model recovered the true root under all conditions investigated (Supplementary Figures 11-13).

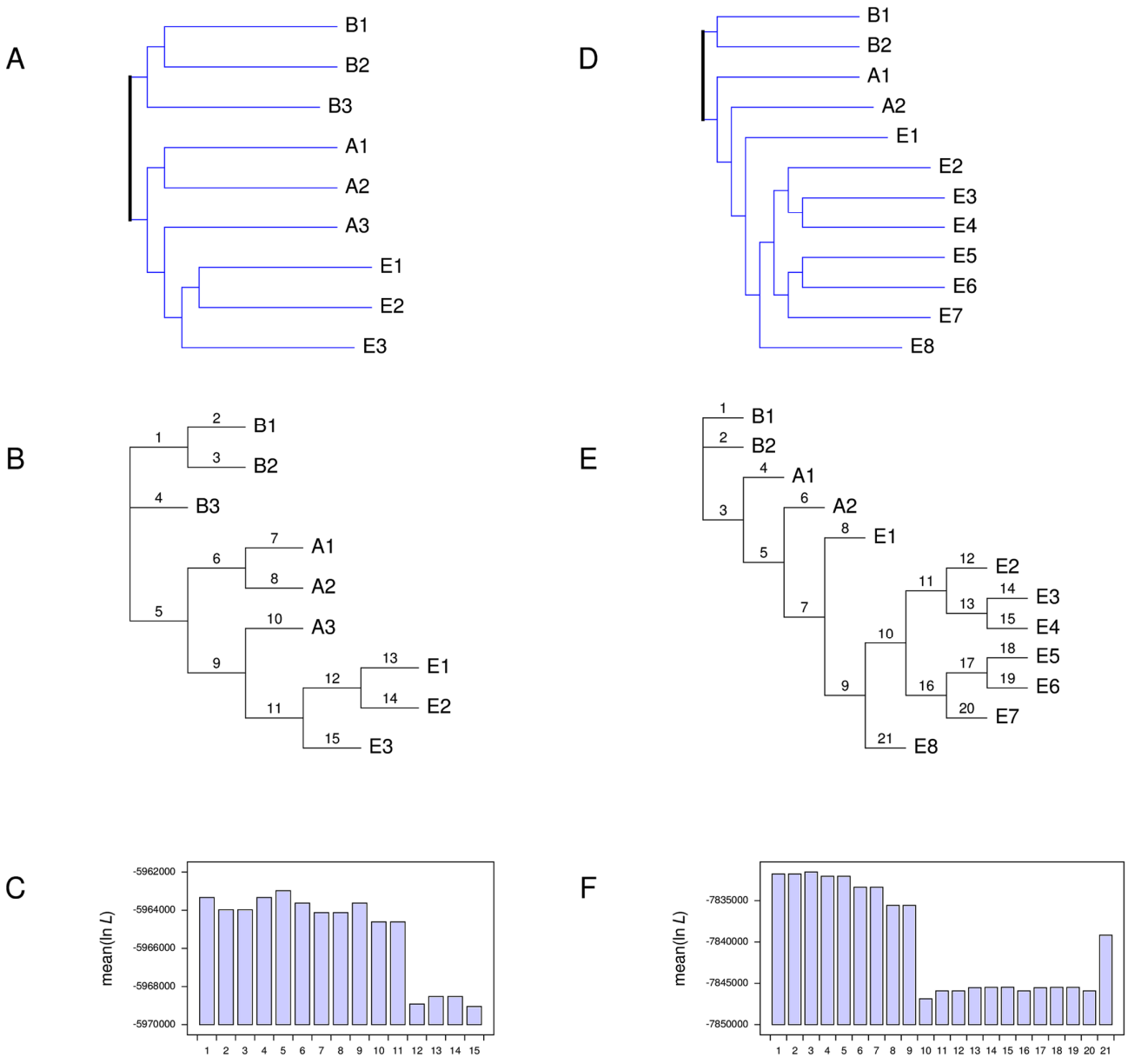


Supplementary Figure 11. Simulations to evaluate root inference under the KVR model when the simulation model is the KVR model -- I. Simulation using the KVR model (two compositions, with the root

given one composition and all other nodes given another) on a tree rooted at the base of the eukaryotes. Panels A and D show the simulation tree, where the two colours indicate the two compositions. The root (shown in black) was 0-rich, and all the other nodes (blue) were 1-rich. Panels B and E show all the possible rootings, each of which was evaluated by ML using the KVR model. Panels C and F show the mean log likelihoods of the simulated trees. The simulation root was root 11 in panel C and root 7 in panel F, and these were both recovered as the ML roots.



Supplementary Figure 12. Simulations to evaluate root inference under the KVR model when the simulation model is the KVR model -- II. Simulation using the KVR model (two compositions, with the root given one composition and all other nodes given another) on a simplified 3D-like tree. Panels A and D show the simulation tree, where the two colours indicate the two compositions. The root (shown in black) was 0-rich, and all the other nodes (blue) were 1-rich. Panels B and E show all the possible rootings, each of which was evaluated by ML. Panels C and F show the mean log likelihoods of the simulated trees using the KVR model. The simulation root was root 5 in panel C and root 3 in panel F, and these were both recovered as the ML roots.

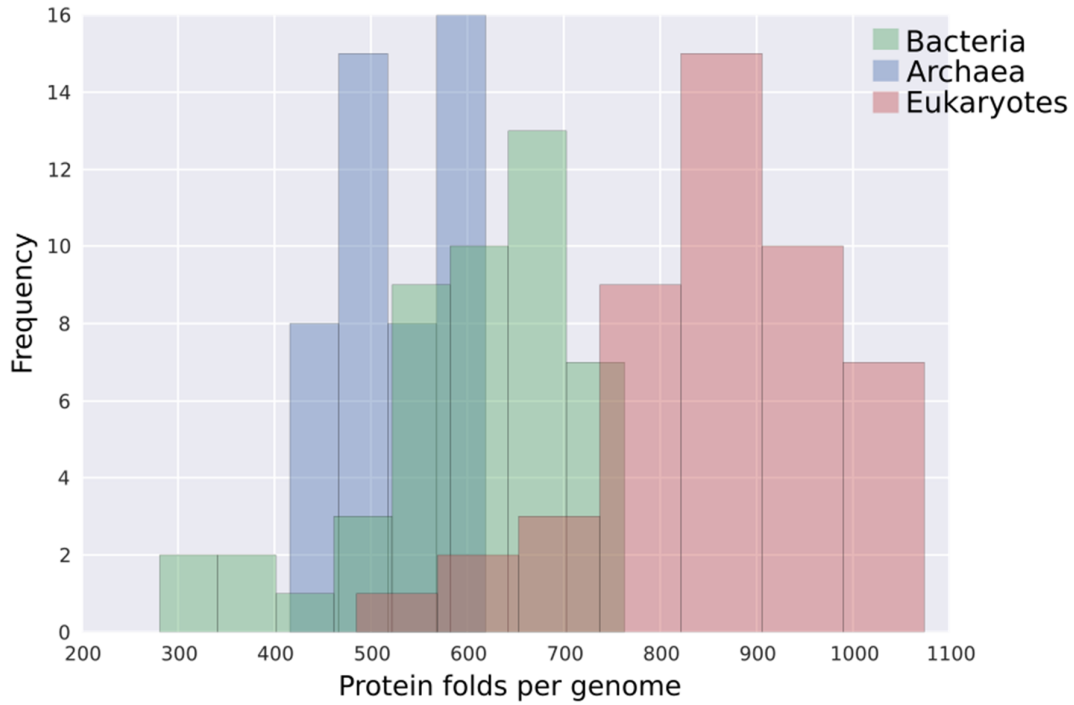


Supplementary Figure 13. Simulations to evaluate root inference under the KVR model when the simulation model is the KVR model -- III. Simulation using the KVR model (two compositions, with the root given one composition and all other nodes given another) on a simplified 2D-like tree. Panels A and D show the simulation tree, where the two colours indicate the two compositions. The root (shown in black) was 0-rich, and all the other nodes (blue) were 1-rich. Panels B and E show all the possible rootings, each of which was evaluated by ML. Panels C and F show the mean log likelihoods of the simulated trees using the KVR model. The simulation root was root 5 in panel C and root 3 in panel F, and these were both recovered as the ML roots.

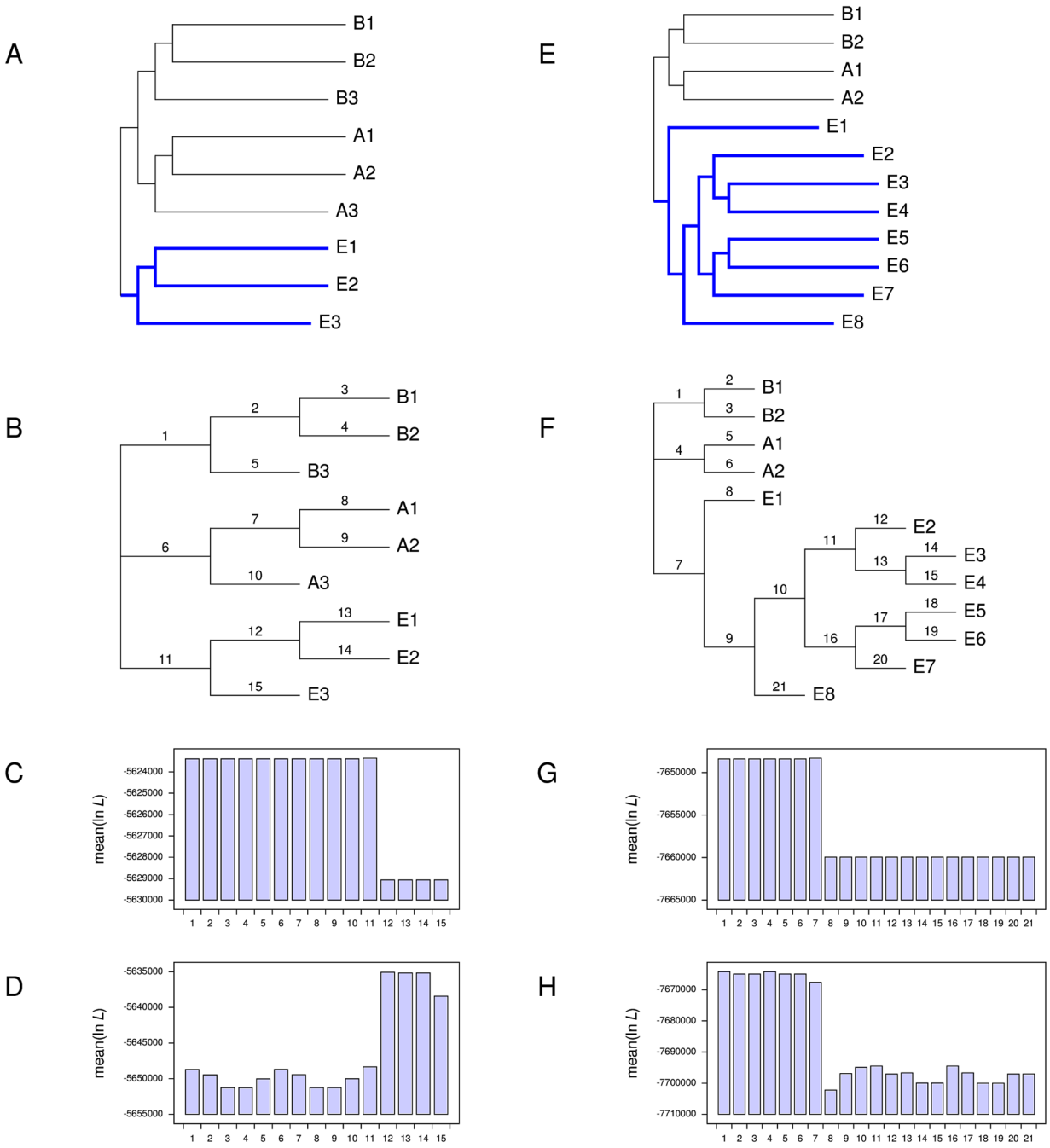
The real data is more heterogeneous than data simulated under the KVR model

However, the real data was not simulated under the KVR model. In particular, the data contain significant among-domain compositional variation. The compositions (that is, sum totals of protein fold presences) of the modern taxa used in the presence-absence matrix of HK vary substantially: from 281 folds in the parasitic bacterium *Ureaplasma urealyticum* to 1074 folds in the eukaryote, *Oryza sativa* (Asian rice). Fold compositions vary systematically among archaea (median 521 folds), bacteria (median 615 folds) and eukaryotes (median 871 folds), with these inter-domain differences being statistically significant in all cases ($P < 10^{-8}$ for the eukaryote-archaea and eukaryote-bacteria comparisons, $P = 0.000278$ comparing bacteria and archaea, Wilcoxon rank-

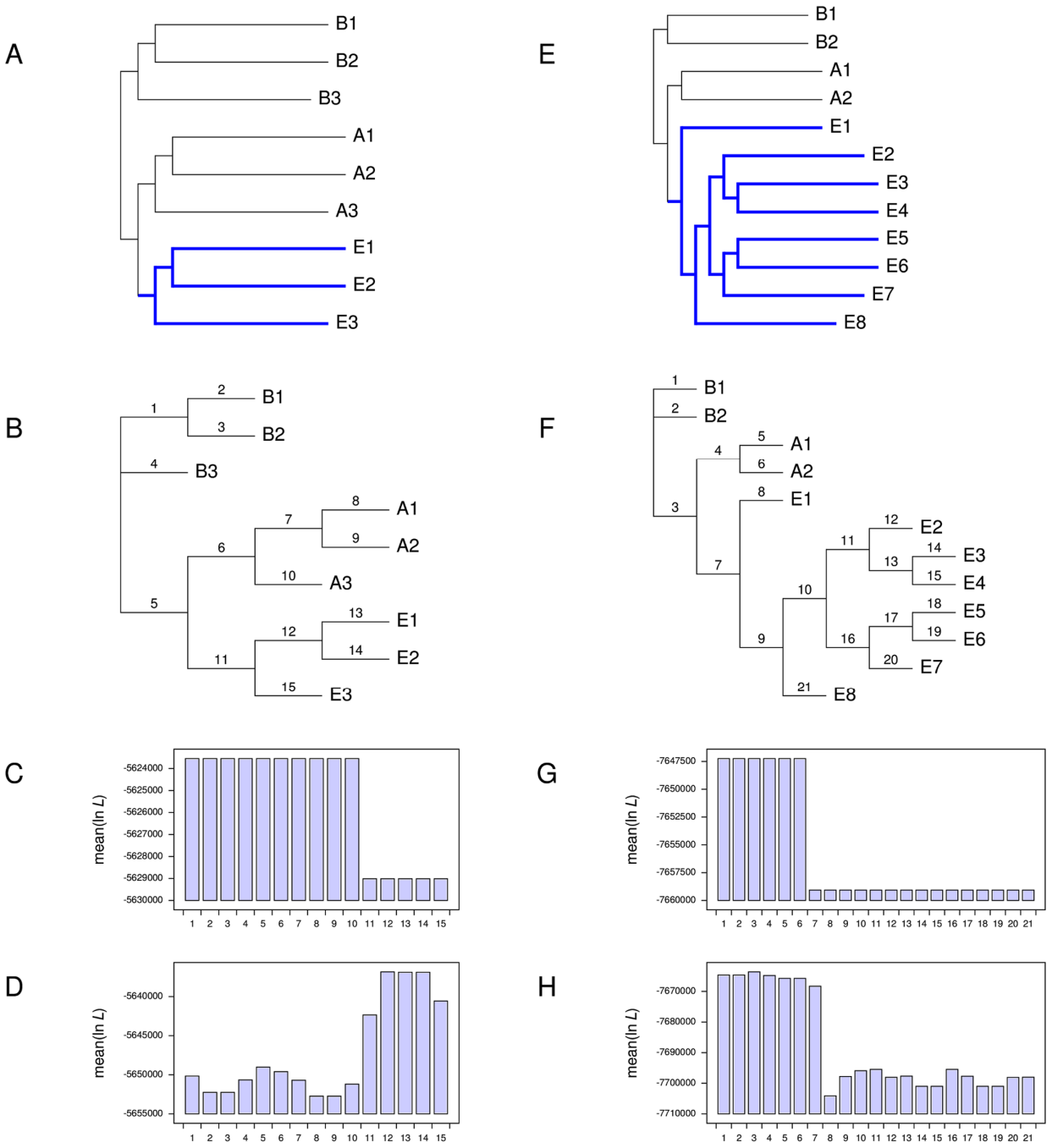
sum tests; see Supplementary Figure 14). Thus, the protein fold data analysed by HK contain abundant evidence of compositional variation across the tree of life following the divergence of the cellular domains from their last universal common ancestor. We therefore performed additional simulations to investigate the rooting behaviour of KVR in the presence of this kind of compositional heterogeneity over the tree (Supplementary Figures 15-17).



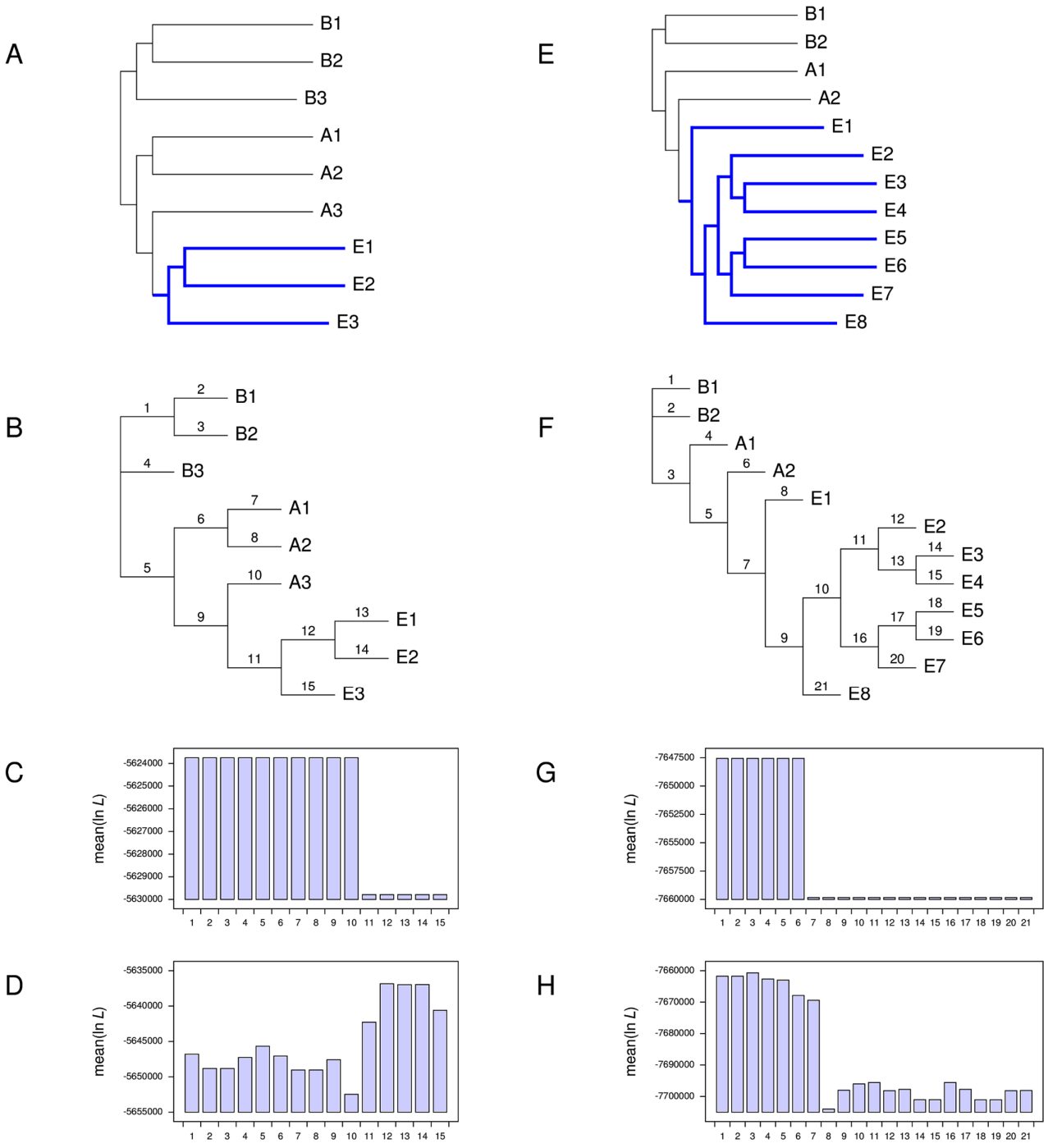
Supplementary Figure 14: The number of SCOP protein families represented on genomes varies significantly within and among the cellular domains. Eukaryotes (median 871 folds) encode significantly ($P < 10^{-8}$) more SCOP protein families than do bacteria (median 615) or archaea (median 521). The KVR model (one composition at the root, one composition elsewhere) assumes equal protein fold compositions within and among the domains.



Supplementary Figure 15. Simulations to evaluate root inference under the KVR model in the presence of compositional heterogeneity - I. Simulation using a two-composition model with eukaryotes given their own composition, on a tree rooted at the base of the eukaryotes. Panels A and E show the simulation trees, where the colours indicate the two compositions. Black lines are 0-rich, and blue lines are 1-rich. Panels B and F show all the possible rootings, each of which was evaluated by ML. Panels C and G show the ML roots for the simulation model. In panel C, the simulation root is root 11, and it has a slightly better mean ML compared to roots 1--10 (which are all equal). The simulation root is root 7 in panel G, and that root has the best mean ML. Panels D and H show ML roots for the KVR model. The simulation roots (roots 11 and 7, respectively) were not the ML roots when evaluated with the KVR model.



Supplementary Figure 16. Simulations to evaluate root inference under the KVR model in the presence of compositional heterogeneity -- II. Simulation using a two-composition model with eukaryotes given their own composition, on a simplified 3D-like tree. Panels A and E show the simulation trees, where the colours indicate the two compositions. Black lines are 0-rich, and blue lines are 1-rich. Panels B and F show all the possible rootings, each of which was evaluated by ML. Panels C and G show the ML roots for the simulation model. The simulation roots (roots 5 and 3, respectively) were one of the ML roots. Panels D and H show ML roots for the KVR model. The simulation root for panel D is root 5, and this is not one of the best roots. Root 3 is the simulation root for the results in panel H, and it was the ML root.



Supplementary Figure 17. Simulations to evaluate root inference under the KVR model in the presence of compositional heterogeneity -- III. Simulation using a two-composition model with eukaryotes given their own composition, on a simplified 2D-like tree. Panels A and E show the simulation trees, where the colours indicate the two compositions. Black lines are 0-rich, and blue lines are 1-rich. Panels B and F show all the possible rootings, each of which was evaluated by ML. Panels C and G show the ML roots for the simulation model. The simulation roots (roots 5 and 3, respectively) were one of the ML roots. Panels D and H show ML roots for the KVR model. The simulation root for panel D is root 5, and this was not one of the best roots. Root 3 is the simulation root for the results in panel H, and it was the ML root.

As before, all simulations were performed on a tree rooted between prokaryotes and eukaryotes (Supplementary Figure 15), a 3D-like tree (Supplementary Figure 16), and a 2D-like tree (Supplementary Figure 17); the conclusions from the simulations are the same in each case. We first simulated 200 binary datasets on the rooted

tree in panel A (Supplementary Figures 15-17), in which taxa proportions are equal to those in the real data matrix (equal proportions of archaea, bacteria and eukaryotes). We used a two-composition model for the simulation, where the eukaryote clade had a composition of (0.3, 0.7), *ie* 1-rich, and everything else, including the root, had a composition of (0.7, 0.3), *ie* 0-rich. To evaluate the best root position, roots were placed on all branches of the trees shown in panel B in each figure. We then fit the simulation model to each of these datasets, calculated the maximum likelihood (ML) values for each root position, and plotted the mean ML of each root in panel C; in all cases, the simulation root was the ML root, or had the joint highest likelihood. We then estimated mean ML values for the KVR model on these data, shown in panel D. KVR did not recover the true root, but instead rooted within the eukaryotes. The mean ln(likelihood) of the KVR evaluations were lower than the simulation model evaluations, indicating a poorer fit of the model to the data. The results in panel D (Supplementary Figures 15-17) show that the KVR model was unable to recover the true root from the simulated presence/absence data. The KVR model strongly favoured a root in the eukaryotes, in conflict with the true root on the simulation tree.

To further investigate the rooting behaviour of KVR, we performed an additional set of experiments on the set of simulation trees shown in panel E (Supplementary Figures 15-17), in which 2/3rds of the taxa had 1-rich (eukaryote-like) compositions; note that this is the opposite situation to the first set of simulations, in which eukaryote-like compositions were in a 1/3rds-minority. Again, roots were placed on all possible branches of the trees shown in panel F for analysis. As before, inference under the simulation model was correct but ambiguous; the simulation root was either the ML root or one of the ML roots. Under the KVR model (panel H), the ML root was either the simulation root (Supplementary Figures 16, 17) or was among the roots with the best likelihood (Supplementary Figure 15). Note that this result contrasts with the first set of simulations, in which KVR failed to find the correct root; what explains this difference? Comparison of the root support under KVR in the two sets of simulations (panels D and H in Supplementary Figures 15-17) reveals that KVR consistently favours root positions on branches with atypical (minority) compositions. In panels A-D, with equal proportions of archaea, bacteria, and eukaryotes, eukaryotic compositions are in the minority, and the root was placed within the eukaryotes; in panels E-H, it is the prokaryotic (0-rich) compositions that are in the minority, and the root was placed within the prokaryotes. When this model bias favouring atypical compositions synergises with the true root position --- that is, when the true root is among the atypical compositions (panels E-H in Supplementary Figures 15-17), then KVR can sometimes (Supplementary Figures 16, 17) identify the true root. But when the true root is among the more typical compositions (panels A-D in Supplementary Figures 15-17), the root inferred under KVR is determined by model bias.

It therefore appears that, when the data are compositionally heterogeneous, the root inferred under KVR will be drawn towards branches of the tree with atypical compositions. This behaviour indicates that root inferences under KVR are unreliable in the presence of compositional heterogeneity, and might explain the root inference under KVR on the HK matrix, in which (as in our first set of simulations) 1-rich taxa are in a 1/3rds minority.

Can published analyses of broadly-distributed protein folds be used to reject the two-domains tree?

In their published analyses, HK represented patterns of protein- fold presence and absence as a binary phylogenetic matrix, with each column representing a SCOP[17] protein fold and the patterns of 1s and 0s corresponding to presences or absence of that fold in a particular genome as determined by the assignments in the SUPERFAMILY database[18]. This representation is problematic for analyses aimed at resolving the relationships between the genes residing on prokaryotic and eukaryotic genomes, because it assumes that all protein folds in the matrix evolve on a single underlying phylogeny. This assumption ignores the possibility of reticulated evolution due to horizontal gene transfer (HGT) and the genetic contributions to eukaryotes of the bacterial endosymbionts that evolved into mitochondria and plastids. Among prokaryotes, analyses of genome evolution suggest that at least two thirds of ancient gene families have undergone HGT at least once in their evolutionary history[19,20]. The degree of HGT into eukaryotes is hotly debated, but the presence in eukaryotes of genes from the bacterial progenitors of mitochondria and chloroplasts is uncontroversial[21,22]. If the endosymbiotic theory is correct, then eukaryote genes with homology to prokaryotes will show a mix of ancestries, whether from

an archaeal host cell, the mitochondrial or plastid endosymbionts, or from other horizontal gene transfers into eukaryotes.

A further implication of the binary matrix representation for protein folds is that character changes do not have a unique interpretation: a change from 0 to 1 might indicate *de novo* origin of an existing fold by convergent evolution (which is likely to be rare), or the gain of an existing fold by HGT; if the latter, then the pattern of presences and absences for that fold cannot be reliably used to infer the underlying tree. When a column contains a mix of 0s and 1s, there is little information with which to determine whether the pattern is best explained by a single post-LUCA origin of the fold, early origin and parallel gene loss, or late origin and gene transfers.

To investigate the evolutionary histories of the protein folds analysed jointly in the HK matrix, we inferred amino acid sequence-level phylogenies for each. We assembled the sequence datasets by using the HMMs associated with each SCOP protein fold[23] to search against the set of bacterial, archaeal and eukaryotic genomes we analyze below, and retrieved and aligned protein regions with significant similarity using the HMMER3 package[24]. Maximum likelihood phylogenies were inferred under the best-fitting substitution model in IQ-Tree[5], using the built-in model selection tool, with 1000 rapid bootstraps[25]. We parsed the resulting trees to systematically assess the congruence of the phylogenetic signal among families. Although many trees were weakly supported due to the short length of most folds, this analysis nevertheless revealed substantial incongruence among the 1160 protein folds for which we were able to infer an ML tree (Supplementary Tables 13-15). Of these 1160 folds, 491 were present in all three domains (bacteria, archaea and eukaryotes). None of these 491 trees supported the monophyly of all three domains, and the eukaryotes were only monophyletic in 22 of the trees. These analyses suggest that the protein folds in the HK matrix cannot be assumed to have tracked the post-LUCA vertical evolution of bacteria, archaea and eukaryotes, and hence are potentially unreliable markers for inferring the relationships among these groups. While the trees are, in general, poorly resolved, they do provide some support for the endosymbiotic theory. Considering trees which contain a clan of eukaryotes but just one of the other domains, 6/17 trees with only eukaryotes and bacteria have Alphaproteobacteria as the sister group (the single highest count), and 28/35 trees containing only eukaryotes and archaea are consistent with a TACKL/eukaryote clan.

Supplementary Table 13: Nearest neighbour of the eukaryotic clan for the 22 out of 491 protein domain trees containing representatives from all three domains in which eukaryotes form a clan. Any clan on the tree has two neighbours; the smaller neighbour often has one or a small number of sequences, and its contents are reported here.

Tree number	Eukaryotes	Bacteria	Archaea	Tree file	Smaller neighbour clan
0	87	1	31	a.118.16.fas.dedup.contree	Three Crenarchaeota
1	2	23	15	a.184.1.phy.contree	PROE2_Bact_Epsilonproteobacteria
2	4	5	4	a.24.22.phy.contree	CHLAV1_Bact_Chlamydiae
3	3	8	15	a.246.2.phy.contree	HLC3_Arch_Asgardarchaeota
4	2	9	2	a.7.2.phy.uniqueseq.phy.contree	SPIR1_Bact_Spirochaetes
5	13	12	21	b.1.13.phy.contree	FIRM22_Bact_Firmicutes
6	2	8	13	b.1.9.phy.contree	CHLO1_Bact_Chloroflexi

7	2	4	4	b.107.1.phy.contree	PROA5_Bact_Alphaproteobacteria
8	27	1	1	b.2.4.phy.contree	CHLO1_Bact_Chloroflexi and HLC3_Arch_Asgardarchaeota
9	17	6	4	b.85.3.phy.contree	Two Euryarchaeota
10	50	7	13	c.103.1.phy.contree	Two Alphaproteobacteria
11	2	10	11	c.148.1.fas.dedup.contree	Three Bacteria
12	55	4	37	c.149.1.phy.uniqueseq.phy.contree	Aboon_Arch_Euryarchaeota
13	16	6	21	c.7.1.phy.contree	CHLO1_Bact_Chloroflexi and Lmira_Arch_Asgardarchaeota
14	4	3	35	c.8.2.phy.contree	PROA5_Bact_Alphaproteobacteria
15	27	3	38	d.208.1.phy.contree	Two Lmira_Arch_Asgardarchaeota
16	66	12	40	d.309.1.phy.uniqueseq.phy.contree	ELUS1_Bact_Elusimicrobia
17	2	4	7	d.349.1.phy.contree	FIRM10_Bact_Firmicutes
18	50	16	28	d.41.2.phy.uniqueseq.phy.contree	Mhung_Arch_Euryarchaeota
19	17	7	4	d.8.1.phy.contree	Csymb_Arch_Thaumarchaeota
20	83	38	20	d.87.2.phy.uniqueseq.phy.contree	Five Alphaproteobacteria and one Deinococcus
21	2	15	10	e.70.1.phy.contree	Three bacteria

Supplementary Table 14: Nearest neighbours of the eukaryotic clans for 17 out of 193 protein domain trees containing only eukaryotes and bacteria for which eukaryotes form a clan. Any clan on the tree has two neighbours; the smaller neighbour often has one or a small number of sequences, and its contents are reported here.

Tree number	Eukaryotes	Bacteria	Archaea	Tree file	Smaller neighbour clan	Larger neighbour clan
0	2	6	0	a.2.21.phy.contree	Firmicute and Tenericute	Four Firmicutes
1	10	6	0	a.290.1.phy.contree	One Betaproteobacteria	Beta-, Gamma-, and Zetaproteobacteria

2	13	3	0	a.293.1.phy.contree	One Alphaproteobacteria	Two Alphaproteobacteria
3	86	2	0	b.122.1.phy.uniqueseq.phy.contree	One Actinobacteria	One Actinobacteria
4	5	2	0	b.16.1.phy.contree	One Alphaproteobacteria	One Gammaproteobacteria
5	3	3	0	b.24.1.phy.contree	all Bacteroidetes	
6	48	2	0	b.42.8.fas.dedup.contree	two Bacteroidetes	
7	2	2	0	b.61.8.phy.contree	one Alphaproteobacteria	One Deinococcus
8	33	7	0	b.66.1.fas.dedup.contree	seven Actinobacteria	
9	114	2	0	d.124.1.phy.uniqueseq.phy.contree	one Alphaproteobacteria	one Epsilonproteobacteria
10	9	3	0	d.174.1.phy.contree	one Firmicute	Actinobacteria and Deinococcus
11	55	3	0	d.381.1.phy.contree	three Alphaproteobacteria	
12	25	3	0	d.58.10.phy.contree	one Bacteroidetes	Beta-, Alphaproteobacteria
13	50	2	0	d.73.1.phy.uniqueseq.phy.contree	Actinobacteria	Zetaproteobacteria
14	54	6	0	d.82.2.phy.contree	one Alphaproteobacteria	five various
15	2	3	0	e.64.1.phy.contree	one Gammaproteobacteria	Beta-, and Gammaproteobacteria
16	2	2	0	h.4.15.phy.contree	two Gammaproteobacteria	

Supplementary Table 15: Nearest neighbours of the eukaryotic clans for 35 out of 77 protein domain trees containing only eukaryotes and archaea for which eukaryotes form a clan. Any clan on the tree has two neighbours; the smaller neighbour often has one or a small number of sequences, and its contents are reported here.

Tree number	Eukaryotes	Bacteria	Archaea	Tree file	Smaller neighbour clan	Larger neighbour clan

0	40	0	33	a.137.1.phy.uniqueseq.phy.contree	Five Crenarchaeota	Crens, Asgards, Eury
1	579	0	2	a.207.1.fas.dedup.contree	Two Crenarchaeota	
2	350	0	2	a.238.1.phy.uniqueseq.phy.contree	Two Asgardarchaeota	
3	53	0	2	a.24.28.phy.uniqueseq.phy.contree	Two Asgards	
4	2	0	36	a.262.1.phy.contree	Two Eurys	mixture
5	69	0	5	a.278.1.phy.uniqueseq.phy.contree	One Crenarchaeota	Three Asgards, One Aig
6	308	0	8	a.289.1.fas.dedup.contree	Three Eury	Cren, Aig, Eury, Asgard
7	41	0	40	a.4.11.phy.uniqueseq.phy.contree	Two Eury	mixture
8	79	0	40	a.4.15.phy.uniqueseq.phy.contree	Asgard+Aig	mixture
9	6	0	3	a.47.4.phy.contree	Euryarchaeota	
10	88	0	39	a.5.8.phy.contree	Korarchaeota	mixture
11	65	0	38	a.60.14.phy.uniqueseq.phy.contree	One Asgardarchaeota	mixture
12	218	0	2	b.132.1.phy.uniqueseq.phy.contree	Two Asgards	
13	48	0	40	b.137.1.phy.uniqueseq.phy.contree	One Cren	mixture
14	52	0	37	b.38.4.phy.uniqueseq.phy.contree	Seven Cren, One Kor	mixture
15	69	0	30	c.116.1.phy.uniqueseq.phy.contree	One Asgard	mixture
16	72	0	2	c.52.3.phy.uniqueseq.phy.contree	Two Asgards	
17	143	0	42	c.9.2.phy.uniqueseq.phy.contree	Two Thau	mixture
18	51	0	4	d.110.4.phy.uniqueseq.phy.contr	Four Asgards	

				ee		
19	2	0	3	d.214.1.phy.contree	Three Eury	
20	99	0	39	d.235.1.phy.uniqueseq.phy.contree	One Asgard	mixture
21	3	0	4	d.274.1.phy.uniqueseq.phy.contree	Four Eury	
22	37	0	26	d.282.1.phy.uniqueseq.phy.contree	One Asgard	mixture
23	108	0	36	d.29.1.phy.uniqueseq.phy.contree	One Kor	mixture
24	38	0	30	d.329.1.phy.uniqueseq.phy.contree	One Asgard	mixture
25	87	0	28	d.355.1.phy.uniqueseq.phy.contree	One Eury	mixture
26	78	0	40	d.58.12.phy.uniqueseq.phy.contree	Two Asgards	mixture
27	64	0	32	d.58.59.phy.contree	One Cren, One Asgard, One Kor	mixture
28	158	0	58	d.68.6.phy.uniqueseq.phy.contree	One Thau	mixture
29	68	0	38	d.78.1.phy.uniqueseq.phy.contree	One Asgard	mixture
30	85	0	5	e.15.1.phy.uniqueseq.phy.contree	One Asgard	Aig + Thaum
31	49	0	34	f.23.28.phy.uniqueseq.phy.contree	One Eury	mixture
32	41	0	30	g.41.16.phy.uniqueseq.phy.contree	One Thaum	mixture
33	53	0	40	g.41.9.phy.uniqueseq.phy.contree	One Cren	mixture
34	7	0	2	h.1.27.phy.contree	Two Cren	

References

1. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 173–179 (2015).
2. Da Cunha, V., Gaia, M., Gadelle, D., Nasir, A. & Forterre, P. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.* **13**, e1006810 (2017).
3. Quang, L. S., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).
4. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst. Biol.* (2017). doi:10.1093/sysbio/syx068
5. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
6. Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R. & Embley, T. M. The archaebacterial origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20356–20361 (2008).
7. Foster, P. G., Cox, C. J. & Embley, T. M. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 2197–2207 (2009).
8. Williams, T. a., Foster, P. G., Nye, T. M. W., Cox, C. J. & Embley, T. M. A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc. Biol. Sci.* **279**, 4870–4879 (2012).
9. Williams, T. a. & Embley, T. M. Archaeal 'dark matter' and the origin of eukaryotes. *Genome Biol. Evol.* **6**, 474–481 (2014).
10. Zhang, C., Sayyari, E. & Mirarab, S. ASTRAL-III: Increased Scalability and Impacts of Contracting Low Support Branches. in *Comparative Genomics* 53–75 (Springer, Cham, 2017).
11. Sayyari, E. & Mirarab, S. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Mol. Biol. Evol.* **33**, 1654–1668 (2016).
12. Steel, M. & Rodrigo, A. Maximum likelihood supertrees. *Syst. Biol.* **57**, 243–250 (2008).
13. Foster, P. Modeling Compositional Heterogeneity. *Syst. Biol.* **53**, 485–495 (2004).
14. Williams, T. A. *et al.* New substitution models for rooting phylogenetic trees. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140336 (2015).

15. Klopfstein, S., Vilhelmsen, L. & Ronquist, F. A Nonstationary Markov Model Detects Directional Evolution in Hymenopteran Morphology. *Syst. Biol.* **64**, 1089–1103 (2015).
16. Harish, A. & Kurland, C. G. Empirical genome evolution models root the tree of life. *Biochimie* **138**, 137–155 (2017).
17. Wilson, D. *et al.* SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* **37**, D380–6 (2009).
18. Gough, J., Karplus, K., Hughey, R. & Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919 (2001).
19. Dagan, T. & Martin, W. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 870–875 (2007).
20. Williams, T. A. *et al.* Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4602-E4611. (2017) 10.1073/pnas.1618463114
21. Martin, W. F., Garg, S. & Zimorski, V. Endosymbiotic theories for eukaryote origin. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140330 (2015).
22. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The Origin and Diversification of Mitochondria. *Curr. Biol.* **27**, R1177–R1192 (2017).
23. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
24. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
25. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).