Origins of major archaeal clades correspond to gene acquisitions from bacteria

Shijulal Nelson-Sathi[1], Filipa L. Sousa[1], Mayo Röttger[1], Nabor Lozada-Chávez[1], Thorsten Thiergart[1], Arnold Janssen[2], David Bryant[3], Giddy Landan[4], Peter Schönheit[5], Bettina Siebers[6], James O. McInerney[7], William F. Martin[1*]

[1] Institute of Molecular Evolution, Heinrich-Heine University, 40225 Düsseldorf, Germany

[2] Mathematisches Institut, Heinrich-Heine University, 40225 Düsseldorf, Germany

[3] Department of Mathematics and Statistics, University of Otago, Dunedin, 9054, New Zealand

[4] Genomic Microbiology Group, Institute of Microbiology, Christian-Albrechts-Universität Kiel, Kiel, Germany.

[5] Institut für Allgemeine Mikrobiologie, Christian-Albrechts-Universität Kiel, Kiel, Germany.

[6] Faculty of Chemistry, Biofilm Centre, Molecular Enzyme Technology and Biochemistry, University of Duisburg-Essen, Essen, Germany

[7] Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland

*Corresponding author: bill@hhu.de

Abstract

The mechanisms that underlie the origin of major prokaryotic groups are poorly understood. In principle, the origin of both species and higher taxa among prokaryotes should entail similar mechanisms — ecological interactions with the environment paired with natural genetic variation involving lineage-specific gene innovations and lineage-specific gene acquisitions[1,2,3,4]. To investigate the origin of higher taxa in archaea, we have determined gene distributions and gene phylogenies for the 267,568 protein coding genes of 134 sequenced archaeal genomes in the context of their homologs from 1,847 reference bacterial genomes. Archaea-specific gene families define 13 traditionally recognized archaeal higher taxa in our sample. Here we report that the origins of these 13 groups unexpectedly correspond to 2,264 group-specific gene acquisitions from bacteria. Interdomain gene transfer is highly asymmetric, transfers from bacteria to archaea are more than 5-fold more frequent than vice versa. Gene transfers identified at major evolutionary transitions among prokaryotes specifically implicate gene acquisitions for metabolic functions from bacteria as key innovations in the origin of higher archaeal taxa.

Genome evolution in prokaryotes entails both tree-like components generated by vertical descent and network-like components generated by lateral gene transfer (LGT)[5,6]. Both processes operate in the formation of prokaryotic species[1,2,3,4,5,6]. While it is clear that LGT within prokaryotic groups such as cyanobacteria[7], proteobacteria[8] or halophiles[9] is important in genome evolution, the contribution of LGT to the formation of novel prokaryotic groups at higher taxonomic levels is unknown. Prokaryotic higher taxa are recognized and defined by rRNA phylogenetics[10], their existence is supported by phylogenomic studies of informational genes[11] that are universal to all genomes, or nearly so[12]. Such core genes encode about 30-40 proteins for ribosome biogenesis and information processing functions, but they comprise only about 1% of an average genome. While core phylogenomics studies provide useful prokaryotic classifications[13], they give little insight into the remaining 99% of the genome, because of LGT[14]. The core does not predict gene content across a given prokaryotic group, especially in groups with large pangenomes or broad ecological diversity[1,4], nor does the core itself reveal which gene innovations underlie the origin of major groups.

To examine the relationship between gene distributions and the origins of higher taxa among archaea, we clustered all 267,568 proteins encoded in 134 archaeal chromosomes using the Markov Cluster Algorithm (MCL)[15] at a ≥25% global amino acid identity threshold, thereby generating 25,762 archaeal protein families having ≥2 members. Clusters below that sequence identity threshold were not considered further. Among the 25,762 archaeal clusters, two thirds (16,983) are archaeal specific — they detect no homologs among 1,847 bacterial genomes. The presence of these archaea-specific genes in each of the 134 archaeal genomes is plotted in Fig. 1 against an unrooted reference tree (left panel) constructed from a concatenated alignment of the 70 single copy genes universal to archaea sampled. The gene distributions strongly correspond to the 13 recognized archaeal higher taxa present in our sample, with 14,416 families (85%) occurring in members of only one of the 13 groups indicated and 1,545 (11%) occurring in members of two groups only (Fig. 1). Another 4% of archaea-specific clusters are present in more than two groups, and 0.3% are present in all genomes sampled (Fig. 1).

The remaining one third of the archaeal families (8,779 families) have homologs that are present in anywhere from one to 1,495 bacterial genomes. The number of genes that each archaeal genome shares with 1,847 bacterial genomes and which bacterial genomes harbor those homologs is shown in the gene sharing matrix (Extended Data Fig. 1), which reveals major differences in the per-genome frequency of bacterial gene occurrences across archaeal lineages. We generated alignments and maximum likelihood trees for those 8,471 archaeal families having bacterial counterparts and containing ≥4 taxa. In 4,397 trees the archaeal sequences were monophyletic (Fig. 2), while in the remaining 4,074 trees the archaea were not monophyletic, interleaving with bacterial sequences. For all trees, we plotted the distribution of gene presence or absence data across archaeal taxa onto the reference tree.

Among the 4,397 cases of archaeal monophyly, 1,053 trees contained sequences from only one bacterial genome or bacterial phylum (Extended Data Figure 2), a distribution indicating gene export from archaea to bacteria. In the remaining 3,315 trees (Supplementary Table 3), the monophyletic archaea were nested within a broad bacterial gene distribution spanning many phyla. For 2,264 of those trees, the genes occur specifically in only one higher archaeal taxon (left portion of Fig. 2), but at the same time they are very widespread among diverse bacteria (lower panel of Fig. 2), clearly indicating that they are archaeal acquisitions from bacteria, or imports. Among the 2,264 imports, genes involved in metabolism (39%) are the most frequent (Supplementary Table 2).

Like the archaea-specific genes in Fig. 1, the imports in Fig. 2 correspond to the 13 archaeal groups. Does the origin of these groups coincide with the acquisition of the imports? If the imports were acquired at the origin of each group, their set of phylogenies should be similar to the set of phylogenies for the archaea-specific, or recipient, genes (Fig. 1) from the same group. As an alternative to single origin to account for monophyly, the imports might have been acquired in one lineage and then spread through the group, in which case the recipient and import tree sets should differ. Using Kolmogorov-Smirnov test adapted to non-identical leaf sets, we could not reject the null hypothesis H_0 that the import and recipient tree sets were drawn from the same distribution for six of the 13 higher taxa: Thermoproteales ($P = 0.32$), Desulfurococcales ($P = 0.3$), Methanobacteriales ($P = 0.96$), Methanococcales ($P = 0.19$), Methanosarcinales ($P = 0.16$), and Haloarchaea ($P = 0.22$), while the slightest possible perturbation of the import set, one random prune and graft LGT event per tree, did reject H_0 at $P < 0.002$ in those six cases, very strongly ($P < 10^{-42}$) for the Haloarchaea, where the largest tree sample is available (Extended Data Fig. 3, Extended Data Table 1). For these six archaeal higher taxa, the origin of their group-specific bacterial genes and the origin of the group are indistinguishable.

In 4,074 trees, the archaea were not monophyletic (Extended Data Fig. 4; Supplementary Table 4-5). Transfers in these phylogenies are not readily polarized and were scored neither as imports nor exports. Importantly, if we plot the gene distributions sorted for bacterial groups, rather than for archaeal groups, we do not find similar patterns such as those defining the 13 archaeal groups. That is, we do not detect patterns that would correspond to the acquisition of archaeal genes at the origin of bacterial groups (Extended Data Fig. 5), indicating that gene transfers from archaea to bacteria, though they clearly do occur, do not correspond to the origin of major bacterial groups sampled here.

In archaeal systematics, Haloarchaea, Archaeoglobales, and Thermoplasmatales branch within the methanogens[13,16], as in our reference tree (Fig. 2). All three groups hence derive from methanogenic ancestors. Previous studies have identified a large influx of bacterial genes into the halophile common ancestor[17], and gene fluxes between archaea at the origin of these major clades[16]. Fig. 2 shows that the acquisition of bacterial genes corresponds to the origin of these three groups from methanogenic ancestors, all of which have relinquished methanogenesis and harbour organotrophic forms[18,19]. Among the 2,264 bacteria-to-archaea transfers, 1,881 (83%) have been acquired by methanogens or ancestrally methanogenic lineages, which comprise 55% of the present archaeal sample.

Neither the archaea-specific genes nor the bacterial acquisitions showed evidence for any pattern of higher order archaeal relationships or hierarchical clustering[20] among the 13 higher taxa, with the exception of the crenarchaeote-euryarchaeote spilt (Extended Data Fig. 6). While 16,680 gene families (14,414 archaea-specific and 2,264 acquisitions) recover the groups themselves, only 4% as many genes (601: 491 archaea-specific and 110 acquisitions) recover any branch in the reference phylogeny linking those groups (Extended Data Fig. 7).

For 7,379 families present in 2-12 groups, we examined all 6,081,075 possible trees that preserve the crenarchaeote-euryarchaeote split by coding each group as an OTU (operational taxonomic unit) and scoring gene presence in one member of a group as present in the group. A random tree can account for 569 (8%) of the families, the best tree can account for 1,180 families (16%), while the reference tree accounts for 849 (11%) of the families (Extended Data Fig. 8). Thus, the gene distributions conflict with all trees and do not support a hierarchical relationship among groups.

Figure 3 shows the phylogenetic structure (gray branches) that is recovered by the individual phylogenies of the 70 genes that were used to make the reference tree. It reveals a tree of tips[21] in that, for deeper branches, no individual gene tree manifests the deeper branches of the concatenation tree. Even the crenarchaeote-euryarchaeote split is not recovered because of the inconsistent position of Thaumarchaea and Nanoarchaea. Projected upon the tree of tips are the bacterial acquisitions that correspond to the origin of the 13 archaeal groups studied here.

The direction of transfers between the two prokaryotic domains is highly asymmetric. The 2,264 imports plotted in Fig. 3 are transfers from bacteria to archaea, occurring only in one archaeal group (Extended Data Table 2, Supplementary Table 6). Yet only 391 converse transfers, exports from archaea to bacteria, were observed (Extended Data Table 2), the bacterial genomes most frequently receiving archaeal genes occurring in Thermotogae (Supplementary Table 7). Transfers from bacteria to archaea are thus >5-fold more frequent than *vice versa*, yet sample-scaled for equal number of bacterial and archaeal genomes, transfers from bacteria to archaea are 10.7-fold more frequent (see Supplementary Information). The bacteria-to-archaea transfers comprise predominantly metabolic functions, with amino acid import and metabolism (208 genes), energy production and conversion (175 genes), inorganic ion transport and metabolism (123 genes) and carbohydrate transport and metabolism (139 genes) being the four most frequent functional classifications (Extended Data Table 2).

The extreme asymmetry in interdomain gene transfers likely relates to the specialized lifestyle of methanogens, which served as recipients for 83% of the polarized gene transfers observed (Supplementary Table 8). Hydrogen-dependent methanogens are specialized chemolithoautotrophs, the route to more generalist organotrophic lifestyles that are not H_2-CO_2 dependent entails either gene invention or gene acquisition. For Haloarchaea, Archaeoglobales and Thermoplasmatales, gene acquisition from bacteria provided the key innovations that transformed methanogenic ancestors into founders of novel higher taxa with access to new niches, whereby several methanogen lineages have acquired numerous bacterial genes[22] but have retained the methanogenic lifestyle.

Gene transfers from bacteria to archaea not only underpin the origin of major archaeal groups, they also underpin the origin of eukaryotes, because the host that acquired the mitochondrion was, phylogenetically, an archaeon[23,24]. Our current findings support the theory of rapid expansion and slow reduction currently emerging from studies of genome evolution[25]. Subsequent to genome expansion via acquisition, lineage-specific gene loss predominates, as evident in Figs. 1 and 2. In principle, the bacterial genes that correspond to the origin of major archaeal groups could have been acquired by independent LGT events[9,14], via unique combinations in founder lineage pangenomes[3,4], or via mass transfers involving symbiotic associations, similar to the origin of eukaryotes[23,24]. For lineages in which the origin of bacterial genes and the origin of the higher archaeal taxon are indistinguishable, the latter two mechanisms seem more likely.

Figure 1: Distribution of genes in archaea-specific families. Maximum-likelihood (ML) trees were generated for 16,983 archaea-specific clusters. Ticks indicate presence (black) or absence (white) of genes in genomes within groups indicated on the left. The number of trees containing taxa specific to each group is indicated at top. To generate clusters, 134 archaeal and 1,847 bacterial genomes were downloaded from the NCBI website [www.ncbi.nlm.hih.gov, version June 2012]. An all-against-all BLAST[26] of archaeal proteins yielded 11,372,438 reciprocal best BLAST hits[27] (rBBH) having an e-value $<10^{-10}$ and ≥25% local amino acid identity. These protein pairs were globally aligned using the Needleman-Wunsch algorithm[28] resulting in a total of 10,382,314 protein pairs (267,568 proteins, 86.6%). These 267,568 proteins were clustered into 25,762 families using the standard Markov Chain clustering procedure[15]. There were 41,560 archaeal proteins (13.4% of the total) that did not have archaeal homologs, these were classified as singletons and excluded from further analysis. The 23 bacterial groups were defined using phylum names except for Firmicutes and Proteobacteria. All 25,752 archaeal protein families were aligned using MAFFT[29] (version v6.864b). Archaeal specific gene families were defined as those that lack bacterial homologs at the e-value $<10^{-10}$ and ≥25% global amino acid identity threshold. For those archaeal clusters having hits in multiple bacterial strains of a species, only the most similar sequence among the strains was considered for the alignment. Maximum likelihood trees were reconstructed using RAxML[30] program for all cases where the alignment had four or more protein sequences. Archaeal species, named in order, are given in Supplementary Table 1. Clusters, including gene identifiers and corresponding COG functional annotations, are given in Supplementary Table 2. The unrooted reference tree at left was constructed as described in Fig. 2.

Figure 2: Bacterial gene acquisitions in archaeal genomes. Upper panel ticks indicate gene presence in the 3,315 ML trees in which archaea are monophyletic. Archaeal genomes listed as in Fig. 1. The lower panel shows the occurrence of homologs among bacterial groups. Gene identifiers including functional annotations are given in Supplementary Table 2. The number of trees containing taxa specific to each archaeal group (or groups) is indicated at top. The *Methanopyrus kandleri* branch (dot) subtends all methanogens in the tree. The 56 genes at right occur in all 13 groups and were likely present in the prokaryote common ancestor. Bacterial homologs of archaeal protein families were identified as described in Figure 1 (rBBH and ≥25% global identity), yielding 8,779 archaeal families having one or more bacterial homologs. An archaeal reference tree was constructed from a

weighted concatenation alignment[29] of 70 archaeal single copy genes using RAxML[30]. The 70 genes used to construct the unrooted reference tree are *rpsJ, rpsK, rps15p, rpsQ, rps19e, rpsB, rps28e, rpsD, rps4e, rpsE, rps7, rpsH, rpl, rpl15, rpsC, rplP, rpl18p, rplR, rplK, rplU, rl22, rpl24, rplW, rpl30P, rplC, rpl4lp, rplE, rpl7ae, rplB, rpsM, rpsH, rplF, rpsS, rpsI, rimM, gsp-3, rli, rpoE, rpoA, rpoB, dnaG, recA, drg, yyaF, gcp, hisS, map, metG, trm, pheS, pheT, rio1, ansA, flpA, gate, glyS, rplA, infB, arf1, pth, SecY, proS, rnhB, rfcL, rnz, cca, eif2A, eif5a, eif2G, valS.*

Figure 3: Archaeal gene acquisition network. Vertical edges represent the archaeal reference phylogeny in Fig. 1 based on 70 concatenated genes, gray shading indicates how often the branch was recovered by the 70 genes analyzed individually. The vertical edge weight of each branch in the reference tree (scale bar at left) was calculated as the number of times associated node was present within the single gene trees (see Source Data). Lateral edges indicate 2,264 bacterial acquisitions in archaea. The number of acquisitions per group is indicated in parentheses, the number of times the bacterial taxon appeared within the inferred donor clade is color coded (scale bar at right). The strongest lateral edge links Haloarchaea with Actinobacteria. Archaea were arbitrarily rooted on the Korarchaeota branch (dotted line). Bacterial taxon labels are (from left to right) Chlorobi, Bacteroidetes, Acidobacteria, Chlamydiae, Planctomycetes, Spirochaetes, ε-Proteobacteria, δ-Proteobacteria, β-Proteobacteria, γ-Proteobacteria, α-Proteobacteria, Actinobacteria, Bacilli, Tenericutes, Negativicutes, Clostridia, Cyanobacteria, Chloroflexi, Deinococcus-Thermococcus, Fusobacteria, Aquificae, Thermotogae. The order of archaeal genomes (from left to right) is as in Fig. 1 (from bottom to top).

References

1. Doolittle, W. F. & Papke, R. T. Genomics and the bacterial species problem. *Genome Biol.* **7**, 116 (2006).
2. Retchless, A. C. & Lawrence, J.G. Temporal fragmentation of speciation in Bacteria. *Science* **317**, 1093-1096 (2007).
3. Achtmann, M. & Wagner, M. Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* **6**, 431-440 (2008)
4. Fraser, C., Alm, E.J., Polz, M. F., Spratt, B. G. & Hanage, W. P. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**, 741–746 (2009).
5. Puigbo, P., Wolf, Y.I. & Koonin, E. V. The tree and net components of prokaryote genome evolution. *Genome Biol. Evol.* **2**: 745-756 (2010)
6. Dagan, T. Phylogenomic networks. *Trends Microbiol.* **19**, 483-491 (2011).
7. Hess, W. R. Genome analysis of marine photosynthetic microbes and their global role. *Curr. Opin. Biotechnol.* **15**, 191-198 (2004).
8. Kloesges, T. *et al.* Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Mol. Biol. Evol.* **28**, 1057–1074 (2011).
9. Williams, D., Gogarten, J. P. & Papke, R. T. Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biol. Evol.* **4**, 1223-1244 (2012).
10. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51**, 221-271 (1987).
11. Rivera, M.C., Jain, R., Moore, J.E., Lake, J.A. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. USA* **95**, 6239-6244 (1998).
12. Puigbo, P., Wolf, Y. I. & Koonin, E. V. Search for a tree of life in the thicket of the phylogenetic forest. *J. Biol.* **8,** 59 (2009).
13. Brochier-Armanet, C., Forterre, P. & Gribaldo, S. Phylogeny and evolution of the Archaea: One hundred genomes later. *Curr. Opin. Microbiol.* **14**, 274-281 (2011).
14. Lake, J. A. & Rivera, M. C. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol. Biol. Evol.* **21**, 681-690 (2004).
15. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
16. Wolf, Y. I., Makarova, K. S., Yutin, N., Koonin E. V. Updated clusters of orthologous

genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. *Biol. Direct* **7**, 46 (2012).

17. Nelson-Sathi, S. *et al.* Acquisitions of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. USA* **109**, 20537-20542 (2012).

18. Bräsen, C., Esser, D., Rauch, B. & Siebers, B. Carbohydrate metabolism in Archaea: Current insights into unusual enzymes and pathways and their regulation. *Microbiol. Mol. Biol. Rev.* **78**, 89-175 (2014).

19. Siebers, B. & Schönheit, P. Unusual pathways and enzymes of central carbohydrate metabolism in Archaea. *Curr. Opin. Microbiol.* **8**: 695-705 (2005).

20. Doolittle, W. F. & Bapteste, E. Pattern pluralism and the tree of life hypothesis. *Proc. Natl. Acad. Sci. USA* **104**: 2043–2049 (2007).

21. Creevey, C. J. *et al.* Does a tree-like phylogeny only exist at the tips in the tree of prokaryotes? *Proc. R. Soc. B.* **271**, 2551–2558 (2004).

22. Deppenmeier, U. *et al.* The genome of *Methanosarcina mazei*: Evidence for lateral gene transfer between bacteria and archaea. *J. Mol. Microbiol. Biotechnol.* **4**, 453–461 (2002).

23. Williams, T. A., Foster, G.F., Cox, C. Y. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504, 231-236 (2013).

24. McInerney, J. O., O'Connell, M. J. & Pisani, D. The hybrid nature of eukaryota and a consilient view of life on Earth. *Nat. Rev. Microbiol.* **12**, 449–455 (2014).

25. Wolf, Y. I. & Koonin, E. V. Genome reduction as the dominant mode of evolution. *BioEssays* **35,** 829–837 (2013).

26. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

27. Tatusov, R. L., Koonin, E.V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).

28. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

29. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).

30. Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**, 456–463 (2005).

Supplemental Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements

We gratefully acknowledge funding from European Research Council (ERC 232975 to W.F.M.), the graduate school E-Norm of the University of Düsseldorf (W.F.M.), the DFG (Scho 316/11-1 to P.S.; SI 642/10-1 to B.S.), and BMBF (0316188A, B.S). G.L. is supported by an ERC grant (281357 to Tal Dagan), D.B. thanks the Alexander von Humbold Foundation for a Fellowship. Computational support of the Zentrum für Informations- und Medientechnologie (ZIM) at the University of Düsseldorf is gratefully acknowledged.

Author Contributions

S.N.-S., F.L.S., M.R., N.L.-C., and T.T. performed bioinformatic analyses; A.J., D.B., and G.L. performed statistical analyses; P.S., B.S., J.O.M., and W.F.M. interpreted results; S.N.-S., F.L.S., G.L., J.O.M., and W.F.M. wrote the paper; S.N.-S., G.L., and W.F.M. designed the study. All authors discussed the results and commented on the manuscript.

Author Information

Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to W.F.M (bill@hhu.de).

Extended Data Figure Legends

Extended Data Figure 1: Inter-domain gene sharing network. Each cell in the matrix indicates the number of genes (e-value $\leq 10^{-10}$ and ≥25% global identity) shared between 134 archaeal and 1,847 bacterial genomes in each pairwise inter-domain comparison (scale bar at lower right). Archaeal genomes are listed as in Fig. 1. Bacterial genomes are presented in 23 groups corresponding to phylum or class in the Genbank nomenclature: *a* = Clostridia; *b* = Erysipelotrichi, Negativicutes; *c* = Bacilli; *d* = Firmicutes; *e* = Chlamydia; *f* = Verrucomicrobia, Planctomycete; *g* = Spirochaete; *h* = Gemmatimonadetes, Synergisteles, Elusimicrobia, Dyctyoglomi, Nitrospirae; *i* = Actinobacteria; *j* = Fibrobacter, Chlorobi; *k* = Bacteroidetes; *l* = Fusobacteria; Thermatogae, Aquificae, Chloroflexi; *m* = Deinococcus-Thermus; *n* = Cyanobacteria; *o* = Acidobacteria; δ,ε,α,β,γ = Delta, Epsilon, Alpha, Beta and Gamma proteobacteria; *p* = Thermosulfurobateria, Caldiserica, Chysiogenete, Ignavibacteria. Bacterial genome size in number of proteins is indicated at top.

Extended Data Figure 2: Presence absence patterns of archaeal genes with sparse distribution among bacteria sampled. Archaeal export families are sorted according to the reference tree on the left. The figure shows the 391 cases of archaea to bacteria export (≥2 archaea and ≥2 bacteria from one phylum only), 662 cases of bacterial singleton trees (≥3 archaea, one bacterium). The 25,762 clusters were classified into the following categories (Supplementary Table 2): 16,983 archaeal specific, 3,315 imports, 391 exports, 662 cases of bacterial singletons with ≥3 archaea in the tree, 308 cases with three sequences (a bacterial singleton and 2 archaea) in the cluster, 4,074 trees in which archaea were non-monophyletic, and 29 ambiguous cases among trees showing archaeal monophyly. The bacterial taxonomic distribution shown in the lower panel. Gene identifiers and trees are given in Supplemental Table 3.

Extended Data Figure 3: Comparison of sets of trees for single-copy genes in 11 archaeal groups. Cumulative distribution functions for scores of tree compatibility with the recipient dataset. Values are *P*-values of the two-sided Kolmogorov–Smirnov two-sample goodness-of-fit in the comparison of the *Recipient* (blue) datasets against the *Imports* (green) dataset and three synthetic datasets, *One-LGT* (red), *Two-LGT* (pink) and *Random* (cyan). **a**, Thermoproteales **b**, Desulfurococcales **c**, Sulfolobales, **d**, Thermococcales **e**, Methanobacteriales **f**, Methanococcales **g**, Thermoplasmatales **h**, Archaeoglobales **i**, Methanococcales **j**, Methanosarcinales **k**, Halobacteriales.

Extended Data Figure 4: Presence absence patterns of all archaeal non-monophyletic genes. Archaeal families that did not generate monophyly for archaeal sequences in ML trees are plotted according the reference tree on left, the distribution across bacterial genomes groups is shown in the lower panel. These trees include 693 cases in which archaea showed non-monophyly by the misplacement of a single archaeal branch. Gene identifiers and trees are given in Supplemental Table 4-5.

Extended Data Figure 5: Sorting by bacterial presence absence patterns for archaeal imports, exports and archaeal non-monophyletic families. Archaeal families and their homologue distribution in 1,847 bacterial genomes are sorted by archaeal (top) and bacterial (bottom) gene distributions for direct comparison. Distributions of archaeal imports sorted by archaeal groups (**a**) and by bacterial groups **(b**); distributions of archaeal exports sorted by archaeal groups (**c**) and by bacterial groups **(d**); distributions of archaeal non-monophyletic gene families sorted by archaeal groups (**e**) and by bacterial groups **(f**).

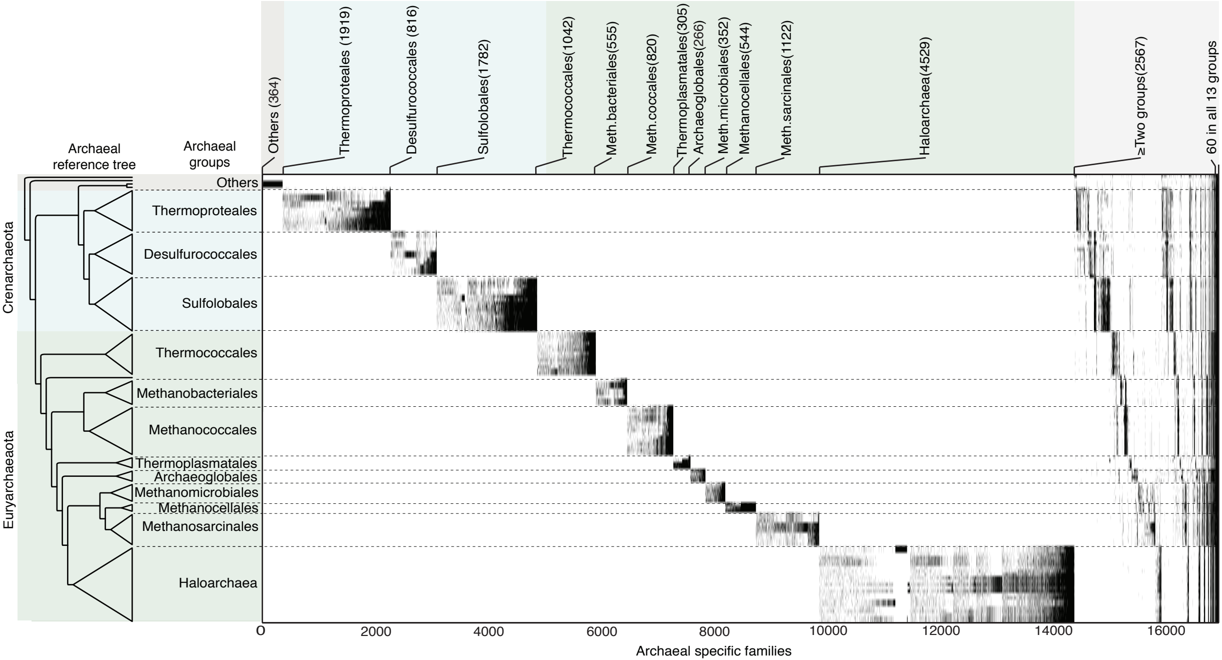
Extended Data Figure 6: Testing for evidence of higher order archaeal relationships using a permutation tail probability (PTP) test. Comparison of pairwise Euclidian distance distributions between archaeal real and conditional random gene family patterns. **a, Archaeal specific families:** Distribution of 2,471 archaeal specific families present in at least 2 and less than 11 groups (top), Comparison between real data and conditional random patterns generated by shuffling the entries within Crenarchaeota and Euryarchaeota separately, Comparison between real data and conditional random patterns generated by including Nanoarchaea and Thaumarchaea into Crenarchaeota (middle) or into Euryarchaeota (bottom). **b, Archaeal import families:** Distribution of 989 archaeal import families present in at least 2 and less than 11 groups (top). Comparison between real data and conditional random patterns generated by shuffling the entries within Crenarchaeota and Euryarchaeota separately by including Nanoarchaea and Thaumarchaea into Crenarchaeota (middle), iii) Comparison between real data and random patterns generated by including Nanoarchaea and Thaumarchaea into Euryarchaeota (bottom).

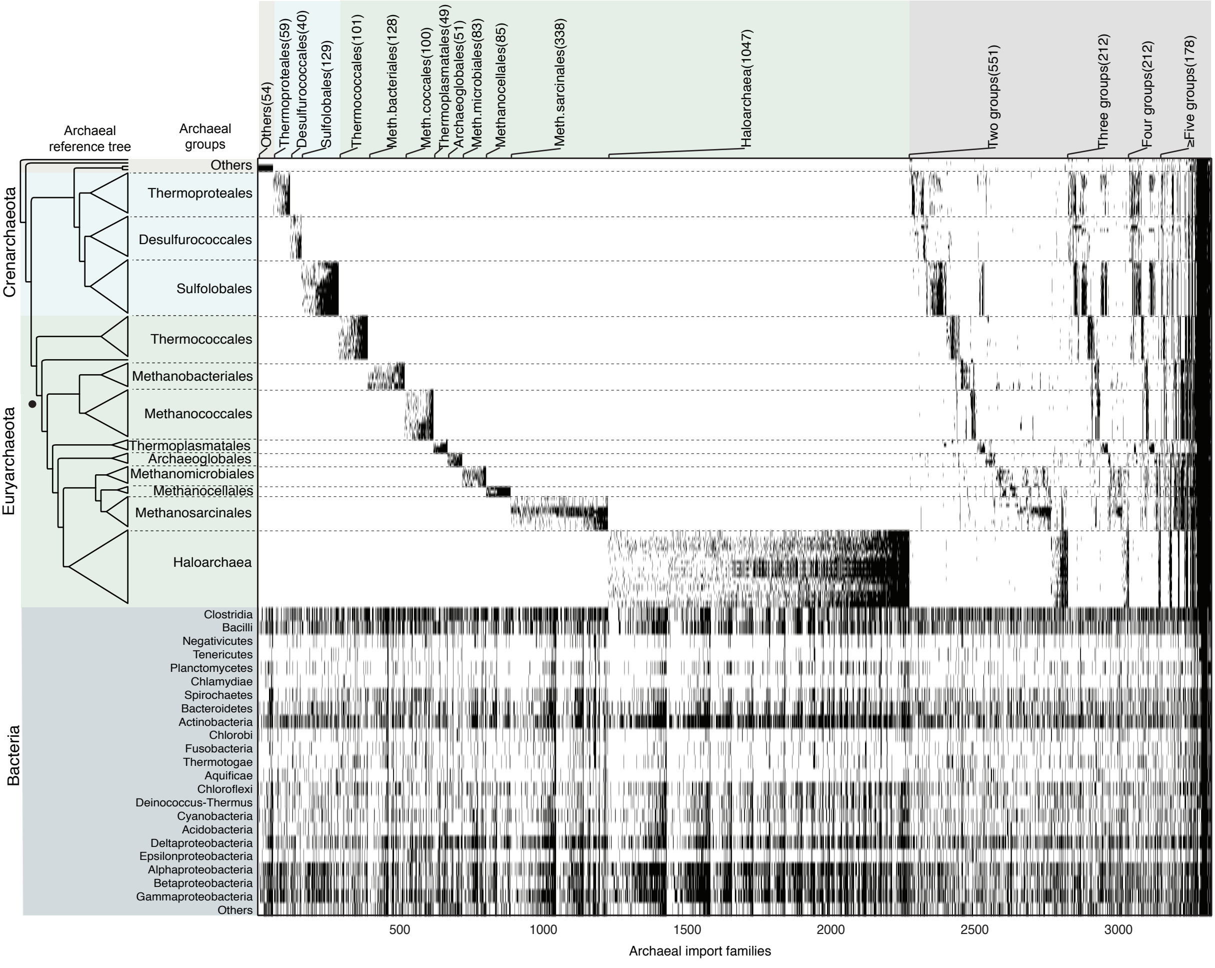
Extended Data Figure 7: Archaeal specific and import gene counts on a reference tree. Number of archaeal specific and import families corresponding to each node in the reference tree are shown in the order of ‘specific/imports’. Numbers at internal nodes indicate the number of archaeal-specific families and families with bacterial homologues that correspond to the reference tree topology. Values at the left indicate the number of archaeal-specific families and families with bacterial homologues that are present in all archaeal groups.

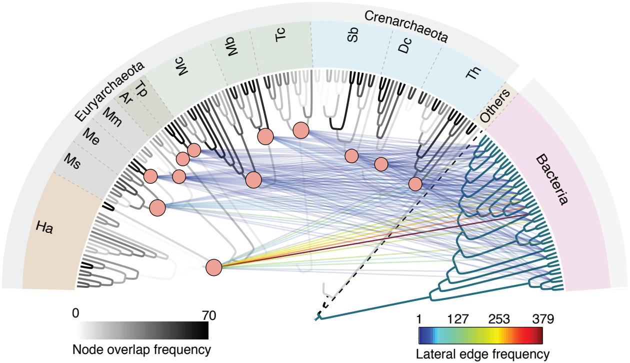
Extended Data Figure 8: Non tree-like structure of archaeal protein families. Proportion of archaeal families whose distributions are congruent with the reference tree and with all possible trees. Filled circles indicate the proportion of archaeal families that are congruent to the reference tree allowing no losses (with a single origin) and different increments of losses allowed. Red, blue, green, magenta and black circles represent the proportion of families that can be explained using a single origin (849, 11.5%), single origin + 1 loss (22.4%), single origin + 2 losses (15%), single origin + 3 losses (13%) and single origin + ≥ 4 losses (38%) respectively. Lines indicate the proportion of families that can be explained by each of the 60,81,075 possible trees that preserve euryarchaeote and crenarchaeote monophyly. Note that on average, any given tree can explain 569 (8%) of the archaeal families using a single origin event in the tree, and the best tree can explain only 1,180 families (16%). In the present data, 208,019 trees explain the gene distributions better than the archaeal reference tree without loss events, underscoring the discordance between core gene phylogeny and gene distributions in the remainder of the genome.

Extended Data Table 1: Comparison of sets of trees for single-copy genes in 11 archaeal groups. Values are *P*-values of the Kolmogorov–Smirnov two-sample goodness-of-fit test operating on scores of tree compatibility with the recipient dataset.

Extended Data Table 2: Functional annotations for archaeal genes according to gene family distribution and phylogeny. Specific: genes that occur in at least two archaea but no bacteria in our clusters. M: archaeal genes that have bacterial homologs and the archaea (≥ 2 genomes) are monophyletic. NM: archaeal genes that have bacterial homologs but the archaea (≥ 2 genomes) are not monophyletic. Exp: exports, the gene occurs in ≥2 archaea but with extremely restricted distribution among bacteria (Supplementary Table 6). Imp: imports, archaeal genes with homologs that are widespread among bacterial lineages, while the archaea (≥ 2 genomes) are monophyletic and the archaeal gene distribution is specific to the groups shown in Figs. 1 and 2.


Archaeal reference tree
Archaeal groups
Crenarchaeota
Euryarchaeota
Others
Thermoproteales
Desulfurococcales
Sulfolobales
Thermococcales
Methanobacteriales
Methanococcales
Thermoplasmatales
Archaeoglobales
Methanomicrobiales
Methanocellales
Methanosarcinales
Haloarchaea
Others (364)
Thermoproteales (1919)
Desulfurococcales (816)
Sulfolobales(1782)
Thermococcales(1042)
Meth.bacteriales(555)
Meth.coccales(820)
Thermoplasmatales(305)
Archaeoglobales(266)
Meth.microbiales(352)
Methanocellales(544)
Meth.sarcinales(1122)
Haloarchaea(4529)
≥Two groups(2567
60 in all 13 groups
0
2000
4000
6000
8000
10000
12000
14000
16000
Archaeal specific families


Archaeal reference tree
Archaeal groups
Others(54)
Thermoproteales(59)
Desulfurococcales(40)
Sulfolobales(129)
Thermococcales(101)
Meth.bacteriales(128)
Meth.coccales(100)
Thermoplasmatales(49)
Archaeoglobales(51)
Meth.microbiales(83)
Methanocellales(85)
Meth.sarcinales(338)
Haloarchaea(1047)
Two groups(551)
Three groups(212)
Four groups(212)
≥Five groups(178)
Crenarchaeota
Euryarchaeota
Bacteria
Others
Thermoproteales
Desulfurococcales
Sulfolobales
Thermococcales
Methanobacteriales
Methanococcales
Thermoplasmatales
Archaeoglobales
Methanomicrobiales
Methanocellales
Methanosarcinales
Haloarchaea
Clostridia
Bacilli
Negativicutes
Tenericutes
Planctomycetes
Chlamydiae
Spirochaetes
Bacteroidetes
Actinobacteria
Chlorobi
Fusobacteria
Thermotogae
Aquificae
Chloroflexi
Deinococcus-Thermus
Cyanobacteria
Acidobacteria
Deltaproteobacteria
Epsilonproteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Others
500
1000
1500
2000
2500
3000
Archaeal import families



Ha - Haloarchaea (1047)
Ms - Methanosarcinales (338)
Me - Methanocellales (83)
Mm - Methanomicrobiales (85)

Ar - Archaeoglobus (51)
Tp - Thermoplasma (49)
Mc - Methanococcales (100)
Mb - Methanobacteriales (128)

Tc - Thermococcales (101)
Sb - Sulfolobales (129)
Dc - Desulfurococcales (40)
Th - Thermoproteales (59)

Others - Korarchaeota, Nanoarchaeota and Thaumarchaeota