doi:10.1038/nature13805

Supplementary Methods

Inferring lateral gene transfer events

Archaeal families having bacterial homologs, showing archaeal monophyly and corresponding to the 13 higher taxa in Fig. 1 were designated as imports (acquisitions specific to an archaea higher taxon). Archaeal genes generating two or more archaeal clades were designated as non-monophyletic, among which those generating exactly two monophyletc archaeal clades were designated as replacements. Trees were classified according to the branching topology of archaeal and bacterial clades using a Perl script. An archaeal clade is monophyletic if there exists a bipartition (branch) in the tree that separates all archaeal from all bacterial leaves. Among trees containing ≥2 bacteria, there were 3,315 cases of archaeal monophyly, 2,264 of which encompassed members of only one archaeal group. Trees in which archaea were monophyletic and the bacterial taxa contain members of only one bacterial phylum (391 cases) were designated as exports (Supplementary Table 6). There are 662 trees containing ≥3 archaea and one bacterium (shown in Extended Data Figure 2), these 662 cases were not scored as exports because they represent singleton cases and, like archaeal singletons, they were excluded from further analysis. The functional classification of protein families in Extended Data Table 2 corresponds to the cluster of orthologous groups database (COG) database[31]. Donor clade membership in Fig. 3 was inferred as previously described[32]. The topology of the bacterial tree shown in Fig. 3 was generated previously from a concatenated alignment of 48 proteins[33].

Permutation Tail Probability (PTP) test

All archaeal recipient families present in more than one group and less than 10 archaeal groups were considered (2,471 families, 14.5% of the data). Gene distributions were condensed into a 13 x 2,471 matrix A, where A_{ij} =1 if at least one member from $family_j$ present in $group_i$. To analyze the relationship between groups, each column in the matrix is randomized within phylum and pairwise Euclidian distance distributions between real and random data were compared using the two-sided Kolmogorov-Smirnov two-sample goodness-of-fit test[34].

Scaling of Genome Samples

To avoid bias in estimates for the ratio of imports to exports due to the larger number of bacterial genomes relative to archaeal genomes in the sample, bacterial sample-scaling for the set of clusters containing bacterial and archaeal homologues was performed by randomly selecting 134 bacterial species while maintaining bacterial higher taxon representation as in the initial dataset. All gene families were then reduced to the scaled sample, and the resulting trees were examined again to detect LGT and classify events, imports and exports were counted. This procedure was repeated 100 times, and the ratio of archaeal imports to exports was determined as the mean.

Comparison of tree sets

The task at hand is to compare two collections of single copy gene trees for each higher taxonomic group, a set reconstructed from recipient genes and a set reconstructed from imported genes. The trees in each set differ from one another, either due to noisy data or due to estimation errors and biases, but our null hypothesis is that genes in both sets evolved along the same phylogeny from a single origin and therefore should display the same phylogenetic signal. In the alternative scenarios, the trees are not related by the same underlying phylogeny, either because of multiple origins or due to lateral gene transfer (LGT) between lineages.

The datasets. The recipient set A and the import set B differ in sample size, and within each dataset, some gene families may be present in only a subset of taxa. We denote the sample sizes by $N^{\mathrm{A}} = \{n_m^A\}$ and $N^B = \{n_m^B\}$ where $m = 1..number\ of\ taxa$. To avoid biases stemming from unequal sample sizes we created two down-sampled datasets A' and B', with a common sample size of $N = \{n_m = min(n_m^A, n_m^B)\}$. Thus, for each m the smaller sample was retained intact, while the larger one was replaced by a random sample (without replacement) of the common, smaller, size. To estimate the alternative scenario of non-vertical inheritance, we generated three additional synthetic datasets: (C') one-LGT trees, constructed by a minimal perturbation of the imported dataset B' where for each tree a random branch was pruned and then re-grafted at a random branch of the remaining trunk, thereby simulating a single lateral transfer event from the grafting branch to the pruned clade; (D') Two-LGT trees where the C' trees were further perturbed by an

additional prune-and-graft operation; and (E') n_m random trees sampled uniformly from the entire tree space.

Measuring phylogenetic congruence. As a reference reflecting the vertical inheritance phylogenetic structure we used the recipient set A (full sample size). This choice avoids committing to some specific phylogenetic tree, while allowing representation of both strong and weak phylogenetic signals. Each tree in the datasets A', B', C', D' and E' was assigned a score designed to reflect its congruence with the reference set, as follows.

For two splits, s_1 from tree t_1 over leaf set L_1, and s_2 from tree t_2 over leaf set L_2, we first create two reduced splits, s_1' and s_2', over the common set of leafs $L_{1,2} = L_1 \cap L_2$. The pairwise split compatibility[35,36] is defined as:

$$comp(s_1, s_2) = \begin{cases} -1 & if\ |L_{1,2}| < 4\ or\ if\ either\ reduced\ split\ is\ uninformative \\ 0 & if\ the\ reduced\ splits\ s_1'\ and\ s_2'\ are\ incompatible \\ 1 & if\ the\ reduced\ splits\ s_1'\ and\ s_2'\ are\ compatible \end{cases}$$

A split s is considered compatible with a tree t if it is compatible with all splits in the tree:

$$comp(s, t) = \begin{cases} -1 & \forall_{s_i \in t}\ comp(s, s_i) = -1 \\ 0 & \exists_{s_i \in t}\ comp(s, s_i) = 0 \\ 1 & otherwise \end{cases}$$

The compatibility of a split s with a set of trees T is defined as the fraction of trees that are compatible with the split:

$$comp(s, T) = \frac{|comp(s, t) = 1|_{t \in T}}{|comp(s, t) = 0|_{t \in T} + |comp(s, t) = 1|_{t \in T}}$$

And finally, the compatibility of a tree t with a reference set of trees T is defined as the minimal compatibility observed among its splits:

$$comp(t, T) = \min_{s \in t} comp(s, T)$$

We note that this definition of tree compatibility, especially the use of the minimal value at the last step, is designed to be sensitive to those features of a tree which are least compatible with the reference set.

Tests. All trees in datasets A', B', C', D' and E' were scored against the reference dataset A, and the sets of compatibility scores were compared using the two-sided Kolmogorov–Smirnov two-sample goodness-of-fit test[11]. For each of the archaeal groups we conducted four tests, comparing A' to each of B', C', D' and E' in turn.

Supplementary References

31. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
32. Thiergart T., *et al.* An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. *Genome Biol. Evol.* **4**, 466-485 (2012).
33. Sousa, F. L. *et al.* Early bioenergetic evolution. *Phil. Trans. R. Soc. B* **368**, 20130088 (2013).
34. Zar, J. H. *Biostatistical analysis*. (Pearson College Div, 2010).
35. Charles, S., and Steel, M., *Tree reconstruction via a closure operation on partial splits*. (Springer Berlin Heidelberg, 2001).
36. Buneman, P. in *The recovery of trees from measures of dissimilarity*. (eds Hodson, F. R., Kendall, D. G. and Tautu P.) 387–395 (Mathematics in the Archaeological and Historical Sciences, Edinburgh University Press, 1971).