Supporting Information

Nelson-Sathi et al. 10.1073/pnas.1209119109

SI Text

Statistical Methods. The task at hand is to compare two collections of trees, 367 trees reconstructed from recipient genes and 109 trees 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. To gain a perspective on how those alternate scenarios will look, we generated two additional synthetic datasets: 109 random trees sampled uniformly from the entire tree space and 109 one-LGT trees, constructed by a minimal perturbation of the imported dataset where a random subtree was pruned and then regrafted at a random branch of the remaining trunk. This simulates a single lateral transfer event from the grafting branch to the pruned clade.

The phylogenetic signal contained within each tree can be summarized in several ways (1). We have examined three basic units of phylogenetic information: phylogenetic partitions (splits), taxa quartets assertions, and triple taxa assertions. Splits and quartets were applied to both the rooted and unrooted versions of the trees, for a total of five phylogenetic signal units.

To test the hypotheses: H_0: Trees in the two sets are drawn from the same underlying tree distribution, vs. H_1: The two sets of trees differ in their underlying phylogenetic signal, we have developed three methodologies: goodness of fit between tree distributions, Euclidean distance between frequencies of phylogenetic assertions, and comparison of distances to a common consensus tree.

Goodness of fit between tree distributions. The two sets of trees were recorded into a $2\times m$ contingency table, where the m categories were defined in an adaptive procedure based on one of the five phylogenetic units. First, the two samples were pooled together into a single set of size n, and the n trees converted into tuples of phylogenetic assertions, or states. Each state was ranked according to its frequency in the pooled state sets. Next, each tree was labeled by the rank of its lowest ranking state, and the pooled tree set was sorted by this label. Bins were defined as a collection of states by sequential addition of states from the sorted list, and creation of a new bin when the current bin included at least $\sqrt{n}$ trees, resulting in $m \leq \sqrt{n}$ bins (the choice of $\sqrt{n}$ is a common practice to ensure a balance between the number of bins and the average sample size for each bin). In the last step, trees from the two sets were added to a $2\times m$ contingency table (with the two rows corresponding to the two sets) based on their label, i.e., their least ranked state. The resulting contingency table was used to derive a standard goodness-of-fit statistic (2). The significance of the goodness-of-fit statistic was tested in a permutation test and the P value estimated from a Monte Carlo simulation with 10^5 permutations. One advantage of the goodness-of-fit statistic is that asymptotically it is χ^2 distributed with $m-1$ degrees of freedom, and the P value can be approximated using the χ^2_{m-1} cumulative distribution function (Table S5*A*).

Euclidean distance between frequencies of phylogenetic assertions. Each of the two sets of trees was converted to a set of phylogenetic assertions, using one of the five phylogenetic units. The two distributions of phylogenetic states were represented as frequency vectors, and the similarity between the two sets was measured by the Euclidean distance between the two frequency vectors. The significance of the Euclidean distance statistic was tested in a permutation test and the P value was estimated from a Monte Carlo simulation with 10^5 permutations (Table S5*B*).

Comparison of distances to a common consensus tree. First, a greedy consensus tree (3) was computed from the pooled set of trees. Next, the distance from the pooled consensus to each tree in the two tree sets was calculated based on one of the five phylogenetic units (1). The distributions of the tree distances for the two sets of trees were compared using the Kolmogorov–Smirnov test (2). (Table S5*C*).

Phylogenetic compatibility with a reference set. The comparison of sets of trees by the foregoing methodologies is applicable only when all trees include the same set of taxa. To extend the analysis to trees that include only a subset of taxa, we examined such trees in terms of their phylogenetic compatibility with a reference set comprised of all recipient trees that do include the full set of taxa. Recipient and imported trees that include only a subset of taxa were grouped based on the number of taxa n, and each group was analyzed separately. Each n taxon tree was decomposed into its *(n-3)* splits, and each split was scored by the fraction of splits in the reference set that are phylogenetically compatible with it. The split compatibility scores for all splits of all trees in the group forms the split compatibility distributions of the group. Additionally, the *(n-3)* split compatibility scores of a specific tree were averaged to produce a tree compatibility score. The distributions of compatibility scores for the recipient and imported groups of trees were compared using the Kolmogorov–Smirnov test (2). (Table S5*D*).

Multiple copy genes. The foregoing tests can be applied only when gene families are present as (at most) single copies (SC) in the several genomes. To apply the tests to trees where multiple copies (MC) of a gene are present in some genomes, we converted the MC trees into SC-like trees by removal of some of the additional copies, using several removal strategies:

i) Condensing of tips: When all copies of a gene in a specific genome form a monophyletic clade in the tree, they can be condensed into a single leaf without affecting the phylogenetic relationships between the several taxa. Only a few MC trees could be converted into SC trees using this strategy.

ii) Retaining exactly one copy per genome: In this approach, we created two sets of SC-like trees, one containing the copies that best fit a reference tree and the second containing the copies with the worst fit to the reference tree. A MC tree was first reduced to a collection of SC-like subtrees by taking all possible combinations of a single copy form each of the several genomes. Next we scored each of the subtrees by its compatibility with the reference tree and retained the two extreme scoring trees as members of the best/worst sets. When several trees were tied with minimal/maximal score, we randomly selected one of the tied trees. We restricted this approach to cases where there are less than 1,024 possible subtrees, only a few cases of very high copy number MC trees were omitted due to this restriction.

iii) Retaining only those genomes where the gene is present in a single copy. This approach can be applied to all MC trees, but some of the resulting SC-like subtrees have less than four taxa and are therefore uninformative.

The goodness-of-fit tests are shown in Table S5*E*, and the tree compatibility tests in Table S5*F*.

Power of the goodness-of-fit test. The goodness-of-fit test based on unrooted splits is powerful enough to reject the recipient vs. one-

LGT comparison. In the one-LGT dataset, every gene is affected by one LGT, raising the question how will the test fair if only some of the genes are affected by LGT. To address this question, we repeated the analysis using random mixtures of the one-LGT and imported datasets (Fig. S3). The goodness-of-fit test based on unrooted splits is powerful enough to reject a mixture of 34% LGT/66% imports at the 5% level.

Common conflicting splits. In Fig. 2*B* (modified version reproduced as Fig. S4), we observed that the six most common splits are compatible and that the tree they define is identical to the haloarchaeal phylogeny generated by 56 universally distributed archaebacterial genes. Moreover, these six splits comprise 51% and 46% of the splits in the recipient and imported sets, respectively. However, other splits are also present in a sizeable proportion of the trees. For example, splits ranking as the 7th to 20th most common are present in about 10% of the trees. The question arises whether these splits indicate an alternative biological signal or whether they are the result of random phylogenetic reconstruction error. If the next 12 or so splits are attributable to random phylogeny errors (as opposed to a biological signal), then the most frequent splits should correspond to alternative topologies that are very close to the reference tree (only one branch being "wrong," for example). If, on the other hand, it is a biological signal, there should be no correlation between split frequency and topological distance to (compatibility with) the reference tree. In Fig. S4, which is a modified version of Fig. 2*B*, we plotted the compatibility of splits with the reference tree (which is also the tree for the first six splits), alongside the split frequencies in the recipient and imported trees.

Clearly, the most frequent splits that are incompatible with the reference tree are also those that are most compatible with it. The correlation is very high (Spearman rank correlation $r = 0.75$; $P = 7{\cdot}10^{-13}$ for the recipients, $r = 0.76$; $P = 7{\cdot}10^{-19}$ for the imports). This strongly indicates that there is no alternative biological signal in this data, but that the second-best splits are behaving exactly as one would expect for the case that phylogeny methods are doing the best they can, but are slightly imperfect.

1. Felsenstein J (2004) *Inferring Phylogenies* (Sinauer, Sunderland, MA).
2. Zar JH (2010) *Biostatistical Analysis* (Prentice Hall, Upper Saddle River, NJ), 5th Ed.
3. Bryant D (2003) A classification of consensus methods for phylogenies. *BioConsensus*, eds Janowitz M, Lapointe FJ, McMorris FR, Mirkin B, Roberts FS (American Mathematical Society, Providence, RI), pp 163–183.

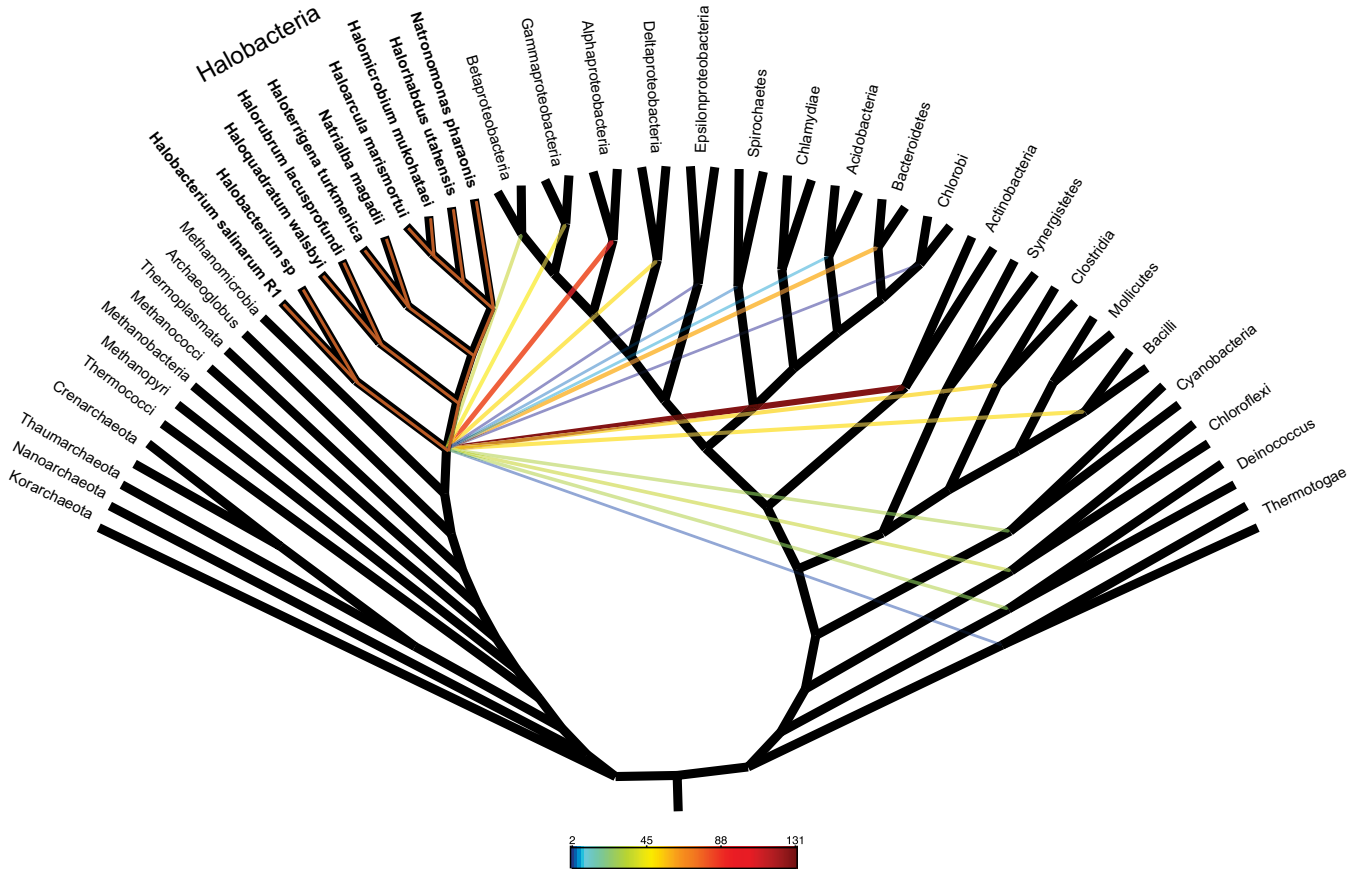

Fig. S1. Acquisition network showing sole donor lineages in what is best understood as a single acquisition from a chimeric donor genome.

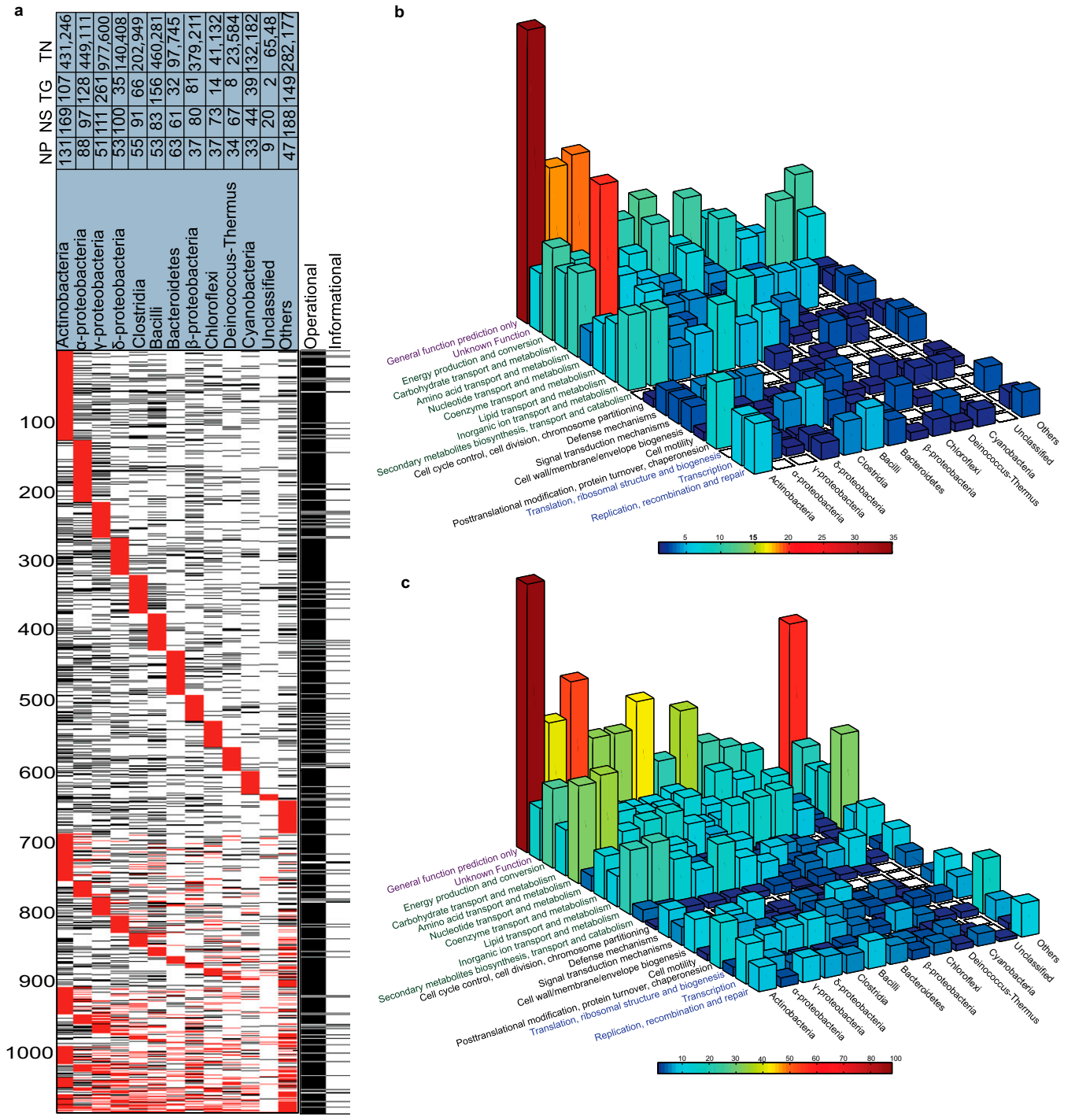

Fig. S2. Phylogenetic affinities and functional classes of eubacterial genes imported into Haloarchaea. (*A*) Presence of eubacterial groups in the sister clade to the haloarchaeal imports (red) and presence in the tree but not in the sister clade (black). The assignment to informational and operational classes for each import is indicated on the right hand side of *A*. Numbers in *A*, *Top* are as follows: NP, number of trees in which the taxon was the only taxon present in the sister clade to the Haloarchaea (the top 691 entries); NS, number of times that the taxon was present in the sister clade to the Haloarchaea (either the sole taxon present or in addition to other taxa); TG, number of genomes sampled for the taxon; TN, total number of genes sampled for the taxon. (*B*) Number of trees in which the taxon was the only taxon present in the sister clade to the Haloarchaea plotted against functional categories. (*C*) Number of trees in which the taxon was present in a mixed sister clade plotted against functional categories.

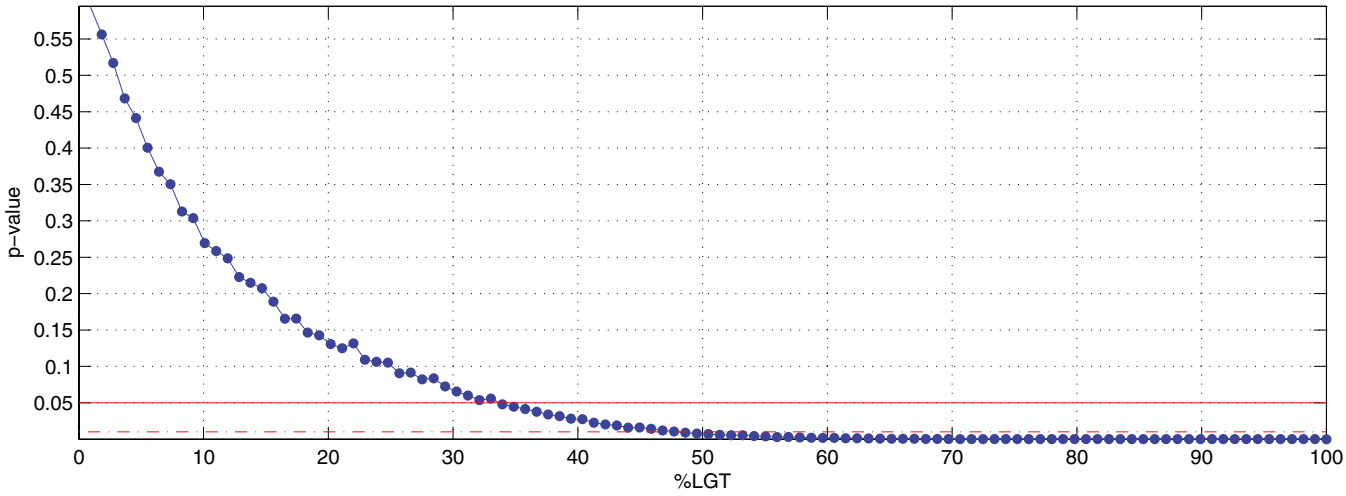

Fig. S3. Power of the goodness-of-fit test.

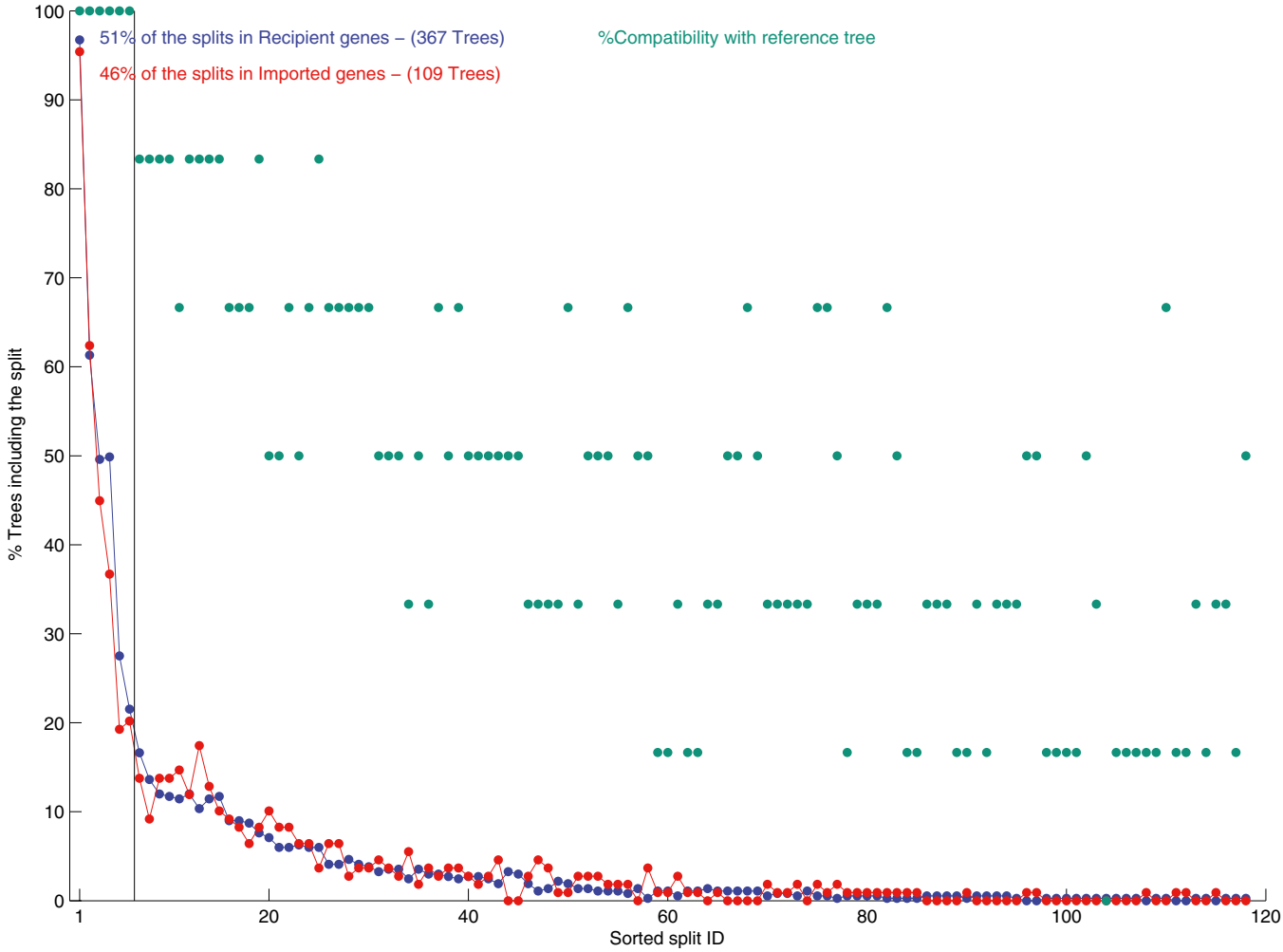

Fig. S4. Split frequencies and their compatibility with the reference tree.

Table S1. Functional categories and distribution of eubacterial imports in Haloarchaea

Table S1

Table S2. Total of 1,089 imports and their functional annotations

Table S2

Table S3. Functional categories and distribution of eubacterial imports in Methanosarcinales (Ms), Methanomicrobialesm (Mm), and Methanocellales (Mc)

Table S3

Table S4. Gene names, functional annotations, and gene distribution among Haloarchaea for components of the respiratory chain

Table S4

Table S5. Statistical tests

Table S5