Quantifying Homologous Replacement of Loci between Haloarchaeal Species

David Williams[1], J. Peter Gogarten, and R. Thane Papke*

Department of Molecular and Cell Biology, University of Connecticut

*Corresponding author: E-mail: thane@uconn.edu.

[1]Present address: Microbial Diversity Institute, Yale University West Campus, West Haven, CT.

Accepted: November 5, 2012

Abstract

In vitro studies of the haloarchaeal genus *Haloferax* have demonstrated their ability to frequently exchange DNA between species, whereas rates of homologous recombination estimated from natural populations in the genus *Halorubrum* are high enough to maintain random association of alleles between five loci. To quantify the effects of gene transfer and recombination of commonly held (relaxed core) genes during the evolution of the class Halobacteria (haloarchaea), we reconstructed the history of 21 genomes representing all major groups. Using a novel algorithm and a concatenated ribosomal protein phylogeny as a reference, we created a directed horizontal genetic transfer (HGT) network of contemporary and ancestral genomes. Gene order analysis revealed that 90% of testable HGTs were by direct homologous replacement, rather than nonhomologous integration followed by a loss. Network analysis revealed an inverse log-linear relationship between HGT frequency and ribosomal protein evolutionary distance that is maintained across the deepest divergences in Halobacteria. We use this mathematical relationship to estimate the total transfers and amino acid substitutions delivered by HGTs in each genome, providing a measure of chimerism. For the relaxed core genes of each genome, we conservatively estimate that 11–20% of their evolution occurred in other haloarchaea. Our findings are unexpected, because the transfer and homologous recombination of relaxed core genes between members of the class Halobacteria disrupts the coevolution of genes; however, the generation of new combinations of divergent but functionally related genes may lead to adaptive phenotypes not available through cumulative mutations and recombination within a single population.

Key words: homologous recombination, horizontal gene transfer, lateral gene transfer, fitness landscape, populations, microbial evolution.

Introduction

Important evolutionary consequences can be caused by a small number of amino acid changes. Hedge and Spratt (1985) demonstrated that a single amino acid change in a penicillin-binding protein (PBP) of *Escherichia coli* decreased the affinity for penicillin, increasing resistance to the antibiotic. Furthermore, Spratt (1988) observed sequence changes in the PBP-2 genes of penicillin-resistant isolates of *Neisseria gonorrhoeae* in the region coding for the penicillin-sensitive domain that likely arose by homologous replacement (HR) from a closely related species. It is now clear that homologous recombination in microbial populations is often a major source of genome diversity. Analyses of large data sets from closely related microorganisms have often revealed panmictic populations in which a high rate of homologous recombination has caused a random association between loci: for example, *Helicobacter pylori* (Salaün et al. 1998); *Halorubrum* spp. (Papke et al. 2004); and *Burkholderia pseudomallei* (Pearson et al. 2009).

The acquisition of novel genes from distant lineages (e.g., Hilario and Gogarten 1993; Nelson et al. 1999), in contrast to replacing preexisting genetic material as in the penicillin resistance example, has led to the innovation of new biochemical pathways (Boucher et al. 2003; Fournier and Gogarten 2008; Khomyakova et al. 2011). Horizontal genetic transfer (HGT), combined with high rates of gene loss, can lead to substantial differences in gene content between members of the same species (Makarova et al. 1999; Welch et al. 2002; Thompson et al. 2005; Normand et al. 2007). These and other reports in the literature imply the generation of genetic and phenotypic diversity during microbial evolution has been from mutations in apparently clonal or recombining populations, and HGT between populations and species.

HGT generates mosaic-like microbial genomes (Lawrence and Ochman 1998; Welch et al. 2002), a result that questions the validity of describing prokaryote diversity and evolutionary history with a tree-like model or using terms such as lineages or species (Hilario and Gogarten 1993; Doolittle 1999; Martin 1999; Doolittle and Zhaxybayeva 2009; Koonin et al. 2011; Williams et al. 2011). Genome-scale comparative analyses provide unprecedented insight into the evolutionary histories of organisms allowing us to characterize and quantify the processes involved. One set of organisms for which several whole-genome sequences are available is the haloarchaea (class: Halobacteria; division Euryarchaeota [Grant et al. 2001]).

Haloarchaea are typically found in salterns, hypersaline marshes and lakes, and inland seas such as the Dead Sea and the Great Salt Lake where they often dominate the microbial community (Antón et al. 1999; Oren 2008). Most members are extreme halophiles requiring >10% (w/v) NaCl for growth and K^+ as a compatible solute, with the associated adaptation of an acidic proteome (Danson and Hough 1997). However, some estuarine isolates grow at 2.5% (w/v) NaCl (Purdy et al. 2004). As a group, the haloarchaea are metabolically diverse heterotrophs (Falb et al. 2008) that respire using oxygen and sometimes nitrate (Oren 2008), although *Haloquadratum walsbyi* demonstrates a narrow range of compounds required for growth: for example, glycerol (Bolhuis et al. 2004), pyruvate (Burns et al. 2007), or dihydroxyacetone (Bardavid and Oren 2008). Unusual for archaea, many representatives of the haloarchaea harbor multiple large (>100 kbp) replicons classed as chromosomes if they host "essential" genes or as megaplasmids if they do not (DasSarma et al. 2009).

Haloarchaea have an unusual mating system involving intercellular cytoplasmic bridges between cells (Rosenshine et al. 1989) and can be artificially transformed in the laboratory (Cline and Doolittle 1992). A recent report by Naor et al. (2012) demonstrated the frequent formation of viable hybrids following recombination between two *Haloferax* species, and there is also evidence that haloarchaea are highly recombinogenic in nature. Multilocus sequence analysis (MLSA) of strains classified as belonging to the genus *Halorubrum* and sampled from different salinities and geographic locations revealed panmictic populations and genetic exchange between populations defined by a phylogeny of concatenated genes (Papke et al. 2004, 2007). A comparison of the completely sequenced genome of an *Haloquadratum walsbyi* isolate (DSM 16790) with an environmental metagenome, both sampled from the same solar saltern in Spain, showed multiple chromosomal regions to be underrepresented in the environment, whereas contiguous regions of environmental fragments only partially matched regions of the sequenced genome. It was concluded that the wider *Haloquadratum walsbyi* community contains a large gene repertoire and is highly recombinogenic (Legault et al. 2006; Cuadros-Orellana et al. 2007). A comparison of the genome with that of another *Haloquadratum walsbyi* strain (DSM 16854) isolated from Australia revealed a putative, mechanistically coupled insertion and deletion system, causing different sequences to be integrated at exactly the same position on the chromosome (Dyall-Smith et al. 2011).

Although monophyly of Haloarchaea was observed for the 16S ribosomal RNA (rRNA) gene (Arahal et al. 2002), this genetic marker is not without criticism for establishing within group relationships because strains from several genera (*Haloarcula* [Mylvaganam and Dennis 1992], *Halosimplex* [Vreeland et al. 2002], and *Natrinema* [Boucher et al. 2004]) can contain multiple copies differing by approximately 7% of the nucleotide positions. Such allelic divergence values are used to distinguish species. In an attempt to increase reliability of inferred evolutionary history, Walsh et al. (2004) performed phylogenetic analysis of a fragment of the single copy, universally distributed beta-subunit of RNA polymerase, rpoB'. They recovered two well-resolved clades: Clade I, containing species of the genera *Halobiforma*, *Haloterrigena*, *Natrialba*, *Natrinema*, *Natronobacterium*, and *Natronorubrum* and Clade II containing the genera *Halobaculum*, *Haloferax*, *Halogeometricum*, and *Halorubrum*. A subsequent analysis by Minegishi et al. (2010) using full-length rpoB' sequences further assigned *Halopiger*, *Halostagnicola,* and *Halovivax* to Clade I and suggested the proximity of the pyrD gene to the 16S rRNA genes as diagnostic of Clade I. Papke et al. (2011) recovered Clade I with strong support and Clade II with moderate support using multiple concatenated loci, but analysis of individual genes revealed only moderate or no support for those clades. MLSA analysis by Andam et al. (2012) found Clades I and II with high support, as well as a newly identified Clade III containing *Haloarcula*, *Halomicrobium,* and *Halorhabdus*.

Complete genome sequences from representatives of Halobacteria have provided insights into their evolution and metabolism. In a comparison of 10 genomes, Anderson et al. (2011) observed a greater potential capacity for polysaccharide degradation, siderophore synthesis, and cell wall modification in soil/sediment isolated organisms. A supermatrix constructed from a concatenation of the aligned amino acid sequences of all inferred gene families with four or more members and analyzed using maximum likelihood (ML) and maximum parsimony (MP) clustering algorithms produced congruent trees broadly in agreement with Clades I–III described earlier.

In an attempt to further characterize the processes responsible for diversity in the Halobacteria, we report a phylogenetic analysis of the widely distributed genes (relaxed core) of 14 previously sequenced genomes representing 13 halobacterial genera, with an additional seven draft genomes augmenting the sample in two of those genera (*Haloferax* and *Haloarcula*). We employed Quartet Decomposition (Zhaxybayeva 2009,

Zhaxybayeva et al. 2006, 2009) for inferring which gene families have been affected by HGT by detecting incongruities among phylogenies. This method analyzes four tip topologies (quartets) embedded in larger phylogenies within a bootstrapped sample to answer the question "do these gene families share the same evolutionary history?" It is immune to the loss of resolution suffered by bootstrap node support (bipartition frequencies) as more sequences are added to a phylogenetic reconstruction or when only a minority of sequences contains a weak phylogenetic signal (Mao et al. 2012); and it avoids problems due to poor taxon sampling often associated with quartet approaches (Zhaxybayeva and Gogarten 2003).

We further enhanced resolution by combining embedded quartets inferred from DNA and amino acid data and used a novel algorithm to infer explicit HGT events and exchange partners using these quartet topologies. To provide insight into the mode of chromosomal integration, we tested the regional homology of the integration sites in sampled descendants of recipient lineages relative to donors and nonrecipients. Finally, we modeled the quantitative relationship between transfer frequency and evolutionary distance of exchange partners allowing us to quantify the degree of chimerism in each genome: the proportion of amino acid changes in the relaxed core genome that had happened in other, divergent haloarchaeal populations, before returning via HGT to a recipient genome.

Material and Methods

Sequence data analysis was implemented in BioPython (Cock et al. 2009) using iPython (Pérez and Granger 2007) and BioPerl (Stajich et al. 2002); phylogenetic computation was implemented using DendroPy 3.10.0 (Sukumaran and Holder 2010) in Python 2.7.2 (www.python.org). Scripts are available from the authors upon request. Other tools were used as described later.

Sequence Data Sources and Genome Annotation

Haloarcula californiae ATCC 33799, *Haloarcula sinaiiensis* ATCC 33800, *Haloarcula vallismortis* ATCC 29715, and *Haloferax denitrificans* ATCC 35960, *Haloferax mediterranei* ATCC 33500, *Haloferax mucosum* ATCC BAA 1512, and *Haloferax sulfurifontis* ATCC BAA 897 were draft genome sequences and (Lynch et al. 2012) were downloaded from http://edwards.sdsu.edu/halophiles/fasta/ (last accessed March 19, 2012). *Haloferax volcanii* DS2 (NC_013965.1, NC_013967.1, NC_013964.1, NC_013966.1, NC_013968.1; Hartman et al. 2010) ATCC 29605, *Haloarcula marismortui* ATCC 43049 (NC_006393.1, NC_006392.1, NC_006396.1, NC_006389.1, NC_006397.1, NC_006391.1, NC_006394.1, NC_006395.1, NC_006390.1; Baliga et al. 2004), *Halobacterium salinarum* R1 (NC_010364.1, NC_010366.1, NC_010367.1, NC_010369.1, NC_010368.1; Pfeiffer et al. 2008), *Halogeometricum borinquense* DSM 11551 (NC_014731.1, NC_014736.1, NC_007428.1, NC_014732.1, NC_014737.1, NC_014735.1; Malfatti et al. 2009), *Halomicrobium mukohataei* DSM 12286 (NC_013201.1, NC_013202.1; Tindall et al. 2009), *Haloquadratum walsbyi* C23, DSM 16854 (FR746099.1, FR746100.1, FR746101.1, FR746102.1; Dyall-Smith et al. 2011), *Haloquadratum walsbyi* HBSQ001, DSM 16790 (AM180089.1, AM180088.1; Bolhuis et al. 2006), *Halorhabdus utahensis* AX-2 (NC_013925.1, CP001687.1; Anderson et al. 2009), DSM 12940, *Halorubrum lacusprofundi* ATCC 49239 (CP001365.1, CP001366.1, CP001367.1; Anderson et al. in preparation), *Haloterrigena turkmenica* DSM 5511 (Saunders et al. 2010), *Natrialba magadii* ATCC 43099 (NC_013922.1, NC_013923.1, NC_013924.1; Siddaramappa et al. 2012), *Natronobacterium pharaonis* Gabara, DSM 2160 (NC_007426.1, NC_007427.1; Falb et al. 2005), and *Halalkalicoccus jeotgali* B3 (NC_014297.1, NC_014301.1, NC_014303.1, NC_014300.1, NC_014302.1, NC_014298.1, NC_014299.1; Roh et al. 2010) and *Halopiger xanaduensis* SH-6 (NC_015666.1, NC_015659.1, NC_015658.1, NC_015667.1; Anderson et al. 2012) were completed genome sequences and were downloaded from the National Center for Biotechnology Information's RefSeq database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). For consistency, new open reading frame (ORF) calls were made using the Rapid Annotation Using Subsystem Technology server (Aziz et al. 2008), and gene functions were predicted and linked to Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology numbers using the KEGG Automatic Annotation Server (KAAS, Moriya et al. 2007). Codon frequencies were counted using the "cusp" program, and effective number of codons was calculated using the "chips" program both from the EMBOSS package version 6.3.1 (Rice et al. 2000).

Gene Family Allocation

Haloarchaeal genomes are known to harbor inteins (Perler 2002). Before assigning ORFs to families according to sequence homology, intein sequences were identified and removed from protein coding sequences because they are not present in all homologous ORFs (Gogarten et al. 2002). Each known intein sequence from InBase (Perler 2002) was used as a seed to build position-specific scoring matrices with Position-Specific Initiated Basic Local Alignment Search Tool (BLAST) 2.2.23+ (Camacho et al. 2009) against InBase and the haloarchaeal protein sequences with an acceptance threshold *e* value of 0.0001. Each matrix was used to query the haloarchaeal protein sequences, and alignments with an *e* value $< 1e{-}10$ were searched at each end with regular expressions designed to match the N-terminus ([ACS][AGFIHMLQSVY]) and C-terminus ([GFHKNS][QSN][CGSTVY]) intein splicing sites from known InBase Bacteria- and Archaea-derived sequences. Multiple alignments of protein sequences

with shared KAAS inferred KEGG orthology numbers that included putative intein containing sequences were performed using Muscle 3.8.31 (Edgar 2004) with the default settings to confirm presence of inteins. Inferred intein sequences were removed and are listed in supplementary table S2, Supplementary Material online.

To establish superfamily clusters of ORFs, each protein sequence was used as a BLASTP (Camacho et al. 2009) query against all proteins, and groups were formed based on *e* values 1e−4. After single-linkage clustering, the MCL algorithm (Enright et al. 2002) was applied with $I = 1.2$ to each group using the lesser of hit-query bidirectional BLAST bitscores, normalized to self-hit bitscores, as edge weights but with hit-query length mismatches 30% set to zero to lessen the influence of less than full-length alignments on ORF clusters formation. The MCL algorithm was repeated on clusters > 210 with increasing *I* values: 1.8, 2.4, 3.0, 3.6, and 4.2 as some very large superfamilies remained after applying smaller *I* values. The resulting superfamilies of sequences were aligned with Muscle 3.8.31, and any remaining distantly homologous sequences were removed from each with scan_orphanerrs from the RASCAL package (Thompson et al. 2003). The superfamilies were realigned, phylogenies inferred with FastTree version 2.1.2 SSE3 (Price et al. 2010), and gene families inferred using the BranchClust algorithm (Poptsova and Gogarten 2007) with many = 11. BranchClust was started at each terminal edge (see Poptsova and Gogarten 2007 for algorithm details) and the run resulting in the most families and greatest inter family edge length (as a tie breaker) was selected.

Phylogenetic Reconstruction of Widely Distributed Gene Families

All ORF family amino acid sequences were aligned using AQUA (Müller et al. 2010) with default settings (Muscle 3.8.31, MAFFT v6.861b (Katoh et al. 2002), RASCAL 1.34 (Thompson et al. 2003), and norMD 1.2 (Thompson et al. 2001), except for -maxiters 32 in Muscle). Nucleotide sequences were aligned to these using Tranalign from the EMBOSS package version 6.3.1 (Rice et al. 2000). Most haloarchaeal genomes have a higher proportion guanine and cytosine bases that cause an increase in erroneous identification of start and stop codons for most gene calling algorithms (Aivaliotis et al. 2007). N-terminal extensions were removed to mitigate phylogenetic reconstruction artifacts caused by inclusion of nonprotein coding sequences. Homology information from ORF family multiple alignments was used to identify putatively erroneous N-terminal extensions defined as regions of ORFs starting in the multiple alignment earlier than the majority of other members that include 1 or more methionine or valine and had predicted isoelectric point (pI) >6 (the predicted pI of most Haloarchaeal ORFs is <5); predictions were made using computePI() from SeqinR library 3.0-5 (Charif and Lobry 2007) for the R statistical computing environment 2.13.2 (Ihaka and Gentleman 1996). C-terminal extensions were rare enough to not warrant similar screening.

Phylogenies were inferred for ORF families with one representative from at least 15 of the 21 genomes from amino acid and nucleotide alignments. Families with more than one ORF from any one genome were excluded from the analysis to minimize ambiguity of histories caused by potential paralogy. For each alignment, substitution model selection for ML reconstruction were made for amino acid alignments with ProtTest (Abascal et al. 2005) using the Akaike Information Criterion (AIC) criterion and for nucleotide alignments with ModelTest (Posada and Crandall 1998) implemented in HyPhy (Kosakovsky Pond et al. 2005). Guide trees were constructed using PhyML 3.0 (Guindon and Gascuel 2003) using the best of NNI and SPR search operations, estimating a proportion of invariant sites and a gamma distribution of among site rate variation with four rate categories by ML using LG substitution matrix (Le and Gascuel 2008) for amino acids and the Hasegawa–Kishino–Yano substitution model (Hasegawa et al. 1985) for nucleotide data. Phylogenies with 100 nonparametric bootstrap replicates were inferred as for the guide trees except where the selected models differed.

Quartet Decomposition

Topologies of all quartets of homologous ORF sequences (each representing a genome) embedded in each set of 100 nonparametric bootstrap replicate phylogenies were extracted from distance matrices of the phylogenies according to the four-point condition of Buneman (1974). This numerical approach proved to be more computationally efficient than inferring embedded quartet topologies by directly manipulating phylogenies represented as data objects. For each embedded quartet in each phylogeny in each set of bootstrap replicates (per gene family), the frequency of each of the three topologies was counted providing a bootstrap score (BSS) of resolution out of (and adding up to) 100. In simulations performed by Zhaxybayeva et al. (2006) to investigate error rates of false-positive and -negative HGT inference by embedded quartet decomposition, they found that omitting embedded quartets with ≥80% BSS in less than 30% of the genomes in which that quartet exists (i.e., poorly resolved in most cases) provided a negligibly low rate of false positives. They also found that excluding those quartets increased the number of false-negative inferences (missed HGTs). The relatively smaller rate of false-positive than false-negative inferences provided a conservative estimate of transfers. The excluded quartets were probably vulnerable to stochastic noise, that is, occasionally well supported but potentially false-positive topologies due to chance in a finite data set. This definition of a "well resolved" quartet as having a bootstrap score of ≥80% is used in the present analysis.

The greatest of the three scores per quartet was taken from the amino acid phylogenies unless it was <80% BSS and that of the nucleotide quartet was ≥80% BSS in which case the latter was taken as the score for that quartet. This approach mitigated loss of information if only considering amino acid sequences when the corresponding nucleotide data provided better resolution as expected for closely related genes. The score for each topology of a quartet across all families in which it is found was summed, and the topology with the highest score was designated the plurality topology for that quartet of genomes (Zhaxybayeva et al. 2006). Embedded quartets may have been affected by long-branch attraction (Felsenstein 1978) when two adjacent long edges in the full phylogeny share a node with the quartet internal edge. Embedded quartets with these characteristics were omitted from the analysis to mitigate false-positive inferences of HGT due to long-branch attraction artifacts (LBAA). Potentially affected quartets were defined as having the shorter of two external adjacent edges on one side of the quartet's central edge more than five times the length of the central edge. Simulations have demonstrated ML estimation accounting for among-site variation to be unaffected by LBAA within these relative long versus short edge length differences (Zhaxybayeva and Gogarten, unpublished). However, the phylogeny inference that provided the embedded quartets was only subject to long-branch attraction with respect to edge lengths in the full phylogeny not each embedded quartet. Therefore, the lengths used for the external adjacent edges were the inner most with respect to nodes in the full phylogeny. If the outer edge of an embedded quartet formed a terminal edge in the full phylogeny, the whole length of the quartet outer edge was considered.

Phylogenies from Genome Sequences

Concatenated Ribosomal Protein Sequences

We inferred a well-resolved, rooted phylogeny for comparison with each ORF family using a concatenation of ribosomal protein coding genes from the 21 haloarchaeal genomes rooted with three outgroup taxa. Steps were taken to avoid model violations due to nonstationarity caused by compositional heterogeneity and systematic errors caused by long-branch attraction (Felsenstein 1978) most likely to affect the edge leading to the outgroup. To decrease the length of the edge to the in group, we selected outgroup taxa from two divergent groups: Nanohaloarchaea and Methanomicrobia. Alignments of each homologous ribosomal protein from the in and out groups were screened for compositional homogeneity using the test of Foster (2004) implemented in PhyloBayes 3.3b using posterior predictive resampling (Lartillot and Philippe 2004). We omitted sequences with a Z score > 2 in an alignment, that is, those with larger deviations in composition, from a concatenation of 59 ribosomal proteins. Sequences from two mesophilic euryarchaea: *Methanosarcina acetivorans* C2A and *Methanococcus aeolicus*. Nankai-3 were also screened in this way. The latter was selected because it had fewer proteins contributing to compositional heterogeneity. An ML phylogenetic reconstruction was performed with RAxML 7.3.0 starting from 20 randomized parsimony trees with a gamma distribution of among site substitution rates using per partition substitution models selected using ProtTest with the AIC criterion (Abascal et al. 2005). Bipartition support was assessed by frequency in 100 nonparametric bootstrap replicates.

Genome Gene Family Composition

For each genome, the presence of a gene family was treated as a character. An MP phylogeny was inferred using the September 2011 version of TNT (Goloboff et al. 2008) with the traditional search, tree bisection reconnection method, 20 search levels, 20 replicated Wagner trees, up to 100 steps for Bremer support (Bremer 1988), and 100 nonparametric bootstrap replicates calculated by frequency differences. To allow an ML phylogenetic reconstruction using PhyML version 20110919 (Guindon and Gascuel 2003), presence was encoded as a cysteine base and absence as an adenosine base with the F84 model of nucleotide substitutions (allows unequal base frequencies and independent rates of transitions and transversions) inferring a proportion of invariable sites and a free distribution of rate categories across a mixture model by ML.

Embedded Quartet Supertree

Plurality-embedded quartet topologies of the strict core gene families were encoded in a matrix according to the method of Baum (1992) and Ragan (1992) used in an MP phylogeny search (MRP) using the September 2011 version of TNT (Goloboff et al. 2008) with the same settings as for gene family composition analysis.

Genome Rearrangements

The strand, order, and chromosome of the core gene families in the subset of genome sequences that were previously fully assembled (*Haloferax volcanii*, *Haloarcula marismortui*, *Halobacterium*, *Halogeometricum*, *Halomicrobium*, *Haloquadratum* DSM 16854 and 16790, *Halorhabdus*, *Halorubrum*, *Haloterrigena*, *Natrialba*, *Natronobacterium*, *Halalkalicoccus,* and *Halopiger*) were used for neighbor-joining phylogenetic reconstruction (Saitou and Nei 1987) from multichromosomal gene rearrangement distances inferred under the "double-cut-and-join" model implemented in TIBA: Tree Inference with Bootstrap Analysis (Lin et al. 2011; http://lcbb.epfl.ch/softwares/tiba.html, last accessed February 12, 2012).

Inference of HGTs

Screening for Transfers from beyond the Sampled Haloarchaea

It was important not to confuse HGT from unsampled donors with ancient HGTs among ancestors of sampled genomes, else interpretation of HGT donor–recipient partners would suffer inaccuracies. If a homolog is horizontally transferred into the sampled haloarchaea from either an unsampled haloarchaeal lineage sister to the sampled group or a nonhaloarchaeal lineage, the recipient would become a cousin clan (sensu Wilkinson et al. 2007, the unrooted analogue of monophyletic group or clade appropriate for phylogenies in which the root is unknown) in the gene tree to the lineage that is deepest in the rooted reference phylogeny. This would be indistinguishable from an HGT from the deepest sampled lineage by analysis of topological incongruities alone. HGT from a donor outside of the sampled group would, in most cases, deliver a homolog with lower sequence similarity than any sampled donor and would resemble an out group often used for rooting phylogenies, that is, an unexpectedly long edge. The following procedure considering branch lengths was used to identify gene families in which incongruities may be due to HGT from unsampled donors from outside of the sampled group, as opposed to HGT among haloarchaea. Gene family phylogenies with unexpectedly long edges were partitioned into sets of homologs either side of those unexpectedly long edges. Unexpectedly long edges were those that were >75% longer than the mean edge length for that phylogeny. This arbitrary length threshold was used to provide a list of potentially problematic gene families which were then screened by BLAST analysis. If a set of homologs had lower BLAST expect scores to non-Haloarchaea than to the other sets from that gene family, an HGT from outside of the haloarchaea was concluded and that set of homologs was excluded from the following analyses to avoid false inference of HGT by phylogenetic incongruity.

Identifying Ancestral HGT Recipient–Donor Pairs within the Sampled Haloarchaea

Statistically supported incongruities between a gene family phylogeny and that of vertical descent can be interpreted as an HGT between a pair of ancestral lineages assuming the descendant of the donor lineage is sampled (see previous section). The difference in topologies caused by a single HGT will result in a different number of conflicting embedded quartets depending on how many nontrivial splits in the reference topology were traversed. For example, two HGTs crossing a small numbers of splits can cause fewer conflicting embedded quartets than one HGT crossing a large number of splits. The following algorithm infers recipient–donor pairs by analysis of conflicting embedded quartets corresponding to topological incongruities. Embedded quartets taken from bootstrap replicates, which provide better resolution than bipartition supports in full gene phylogenies, were compared with those of the concatenated ribosomal protein phylogeny taken to be a proxy for that of vertical descent. The embedded quartets differing between the ribosomal protein phylogeny and gene family with adequate resolution (>80% BSS) were divided into groups that described the same incongruities (a phylogeny may be affected by more than one HGT). Each group was reduced to a single quartet in which each tip represented regions of the full topologies that were congruent (sometimes referred to as "branch and bind"). This was achieved by combining all two-member quartet topology defined sets if they had shared membership ("single-linkage clustering"). This yielded several sets containing homologs or groups of homologs corresponding to congruent regions of the two topologies. Two of these groups represent exchange partners and are cousin clans (sensu Wilkinson et al. 2007) in the gene family phylogeny but are not sister clades in the genome lineage phylogeny.

HGT exchange partners that appear adjacent in the gene family phylogeny can be recovered by discarding those sets that are sisters in the genome lineage phylogeny. Where several homologs are recovered, an ancestral HGT affecting more than one sampled descendant has been inferred. Repeating this process using a genome reference phylogeny on which previously inferred transfers are applied by subtree pruning and regrafting operations, nested and overlapping transfers in a single gene phylogeny can be recovered. Rearrangements involving sister clades with two members or four member comb phylogenies were inferred by a set of simple conditions for each scenario. When HGT pairs cannot be recovered but conflicting embedded quartets remain, only nonspecific evidence of HGT in that gene family can be concluded due to insufficient resolution in the data. The recipient in the HGT pair can be inferred by assessing which is in a different phylogenetic context in the gene family phylogeny.

Characterization of HGTs

Transfer of Multiple Homologs

For HGT donor–recipient lineage pairs inferred from conflicting embedded quartets for a specific homolog, the hypothesis that its neighboring ORFs were also transferred in the same event was tested. First, the homology of the next ORF in the 5′ direction along the chromosomes of the donor, recipient, and nonrecipient was tested (i.e., did it belong to the same gene family?) allowing up to four inserted or deleted ORFs in each strand. If homologous and in a single copy per genome, widely distributed gene family for which embedded quartets were obtained indicating the same donor–recipient lineage HGT, it was included in the same multi-ORF HGT event. This process was continued along both strand directions until a homolog was not transferred or not identified between the pair.

Additionally, for donor–recipient lineage pairs separated by distance *D* along the edges of the ribosome phylogeny, where the recipient was within $D \times 0.85$ to other genomes unaffected by HGT for that gene family (nonrecipients), a multiple ORF transfer was inferred if the ML estimate of substitutions per site distance (inferred using the WAG substitution model [Whelan and Goldman 2001] with five rate categories in a gamma distribution as implemented in RAxML 7.3.0 [Stamatakis 2006] from a multiple sequence alignment of all homologs in the sampled genomes) was smaller to the donor than to the nonrecipient, that is, if the ratio of pairwise distances for that homolog was in conflict with that of the concatenated ribosomal protein phylogeny (see fig. 2 for an example). Many donor–recipient pairs had several sampled descendants in which case the analysis with the shortest multi-ORF transfer was retained to provide a conservative estimate of HGT unit size. Chromosome gene maps to aid in this analysis were plotted using the R package genoPlotR (Guy et al. 2010).

Mode of Chromosomal Integration

The transferred homolog or homologs were inferred as HR if they were located in a chromosomal region with orthology to the region containing the ancestral versions in the reference genome (described in the previous section). The use of a reference genome allowed confirmation that an ORF underwent HR within an orthologous region with common ancestry between the donor and recipient by excluding the possibility of transfer of that whole region or genomic island (a xenologous region) causing syntenic conservation. If the transferred ORF or ORFs were found in a region other than that identified in the putative donor and close relative, nonhomologous insertion (NHI) followed by loss of the pre-existing version from the orthologous region was inferred. Chromosomal rearrangements during evolution means the probability of identifying homologous regions decreases with evolutionary distance, and for many HGTs, the recipient did not have close relatives with orthology for the gene. Whether these requirements were met for each HGT therefore depended on the phylogenetic placement of the donor and recipient among the available genomes.

If homologous regions were not identified, the mode of integration could not be inferred. If homologous regions were identified and the region of HGT ORF(s) intersects a window of eight ORFs around the center of the region in the recipient, HR was inferred, else NHI (followed by loss of the original homolog for the single copy families analyzed here) was inferred as the mode of chromosomal integration.

Initially, the reference genome chromosomes were scanned with a moving window of eight ORFs. If a single region in the nonrecipient contained two of the same homologs found within four ORFs in or around the HGT unit in the recipient chromosome, those regions were considered homologous. The fewest gene families per genome was 2,212 in the *Halobacterium salinarum,* whereas the average 3,077; the probability of finding any two of four homologs in a window of eight in a genome of 2,212 homologs is $(4 \times 8 \times [1/2{,}212])^2 = 0.0002$ providing a false-positive rate of 0.02% for transfers to *Halobacterium* but for the majority of inferences 0.01% on average.

Modeling Exchange Partner Sequence Similarity versus *Frequency of HGTs*

The frequency of HGT was calculated as the quantity of HGT events during the time a pair of HGT partner lineages coexisted. Time of coexistence was estimated as the length of overlapping edges in a maximum clade credibility phylogeny (e.g., the region labeled "t" in fig. 1*B*) from a Bayesian posterior distribution of phylogenies using the ribosomal protein sequences described earlier under an uncorrelated log-normal relaxed molecular clock (Drummond et al. 2006). The data were partitioned into large and small ribosomal subunit associated sets of sequences, the tree prior set to a Yule model, and the substitution model to WAG (Whelan and Goldman 2001) with five categories in a gamma distribution of among site rate variation. Four Markov chain Monte Carlo sampling chains of 20,000,000 and one of 14,000,000 generations with a discarded burnin of 800,000 generations using BEAST v1.6.1 (Drummond and Rambaut 2007) and BEAGLE v1.0 (Ayres et al. 2011) with an MSI (City of Industry, CA) N560GTX-TI TWIN FROZR II 2G GPU were calculated. The smallest effective sample size was 170 as calculated by Tracer v1.4.1 (Rambaut and Drummond 2007) as five separate trace files or after serializing with LogCombiner (part of the BEAST package) indicating both an adequate burnin and convergence.

The sequence similarity was taken to be the substitutions per site across the RAxML inferred ribosomal protein phylogeny described earlier. Although rates of evolution will vary between gene families, relative rates among lineages within gene families may be similar to those of the ribosomal proteins. Specifically, between the points on the donor–recipient edges mid-way along the overlapping region in the relaxed molecular clock tree (e.g., the region labeled "t" in fig. 1*B*) scaled to the equivalent point in the substitutions per site tree (e.g., the terminal ends of the regions labeled "d1" and "d4" in fig. 1*A*) spanning the edges lengths since the donor–recipient last common ancestor (e.g., the regions labeled "d1" to "d4" in fig. 1*A*). The distances between partners may be underestimated when the phylogenetic resolution within a clan of putative transfer partner homologs (either recipient or donor descendents) was insufficient to infer the precise edge of horizontal transfer: the next deepest edge of resolution 80 would have been returned by the algorithm used to infer HGT by phylogenetic incongruity. The resolution in the gene phylogenies within the inferred HGT partner groups was

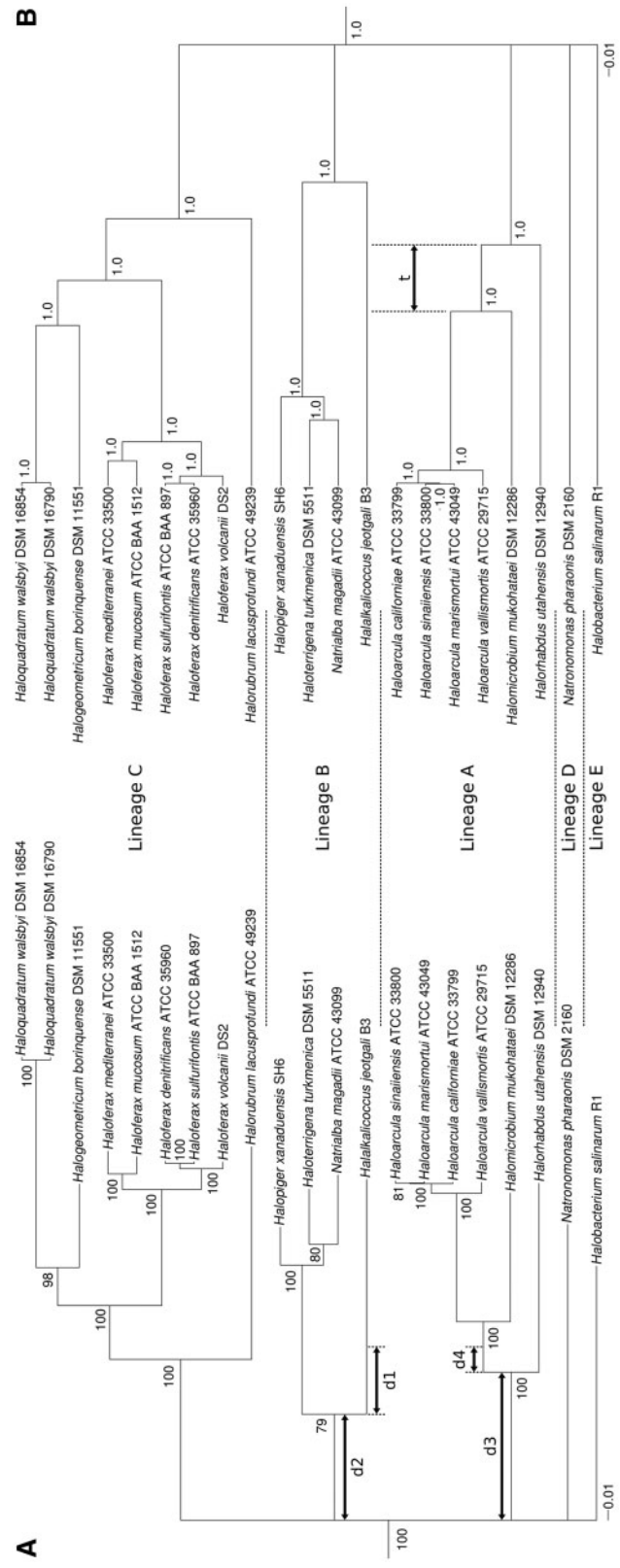

Fig. 1.—(*A*) ML phylogenetic reconstruction from 59 concatenated ribosomal protein sequences from 21 haloarchaea with edge lengths scaled to substitutions per site. Two sets of nanohaloarchaeal and one mesophilic methanogen from Methanomicrobia were used as an outgroup. Protein homologs inferred as causing compositional heterogeneity were excluded, and the deepest bipartitions were collapsed due to inconsistency among nonparametric bootstrapped replicates and evidence of LBAA. (*B*) Bayesian sampled phylogeny inferred from the same data set with edge lengths scaled to a relaxed molecular dock. As an example, the edges marked d1–4 in (*A*) and the regions labeled "t" in (*B*) indicate the genetic distance between and the duration of coexistence respectively of the ancestral lineages of *Halalkalicoccus* and of *Haloarcula* and *Halomicrobium* used in HGT frequency versus genetic distance modeling. All pairwise, coexisting, nonsister edges were included.

tested by checking for embedded quartets that supported each of the next edges within the regions of the gene family phylogeny associated with either exchange pair until supported. The mean distance into the ribosome phylogeny along unresolved edges was added to the distance between exchange partners to account for this uncertainty. A linear model was fitted with the lm() function after a log transformation of the HGT frequency data using the log() function of the base package of R 2.14.2 (Ihaka and Gentleman 1996).

Inferring the Relative Contributions of "in-lineage" and "out-lineage" Sequence Substitutions in Relaxed Core Genes

The total "in lineage" substitutions for ORFs in single copy relaxed core families were calculated as the distance from each tip to the root of the ML ribosomal protein phylogeny multiplied by the quantity of such ORFs in the genome sampled for that lineage (units: ORF.substitutions.site^{-1}).

The total "out-lineage" substitutions were calculated by predicting the HGT frequency for each edge between a tip and the root with each of all other coexisting lineages, according to the relaxed molecular clock phylogeny, using the corresponding distances in the substitutions per site phylogeny as the distance for the fitted linear model. For each edge pair, the HGT frequency (units: HGT.time^{-1}) was multiplied by the mean number of ORFs per HGT (units: ORF.time^{-1}) and then by the average of half of the edge lengths in each lineage since the last common ancestor to (assuming equal transfers in each direction: otherwise the edge length in the donor lineage would be used) give horizontally acquired substitutions (units: ORF.substitutions.time^{-1}.site^{-1}), finally multiplying by the length of overlapping edges (units: ORF.substitutions.site^{-1}).

Results

Genes, Families, and Inteins

To ensure consistency across genome analyses, we inferred and reannotated the coding regions for all genomes in this study using the same algorithms (KAAS for functional annotation, Moriya et al. 2007; RAST for gene calls, Aziz et al. 2008). *Halobacterium salinarum* R1 had the fewest predicted ORFs with 2,619, whereas *Haloarcula sinaiiensis* ATCC 33800 had the most with 4,311. *Natronomonas pharaonis* DSM 2160 had the highest ORF density with 90.3%. ORF density was more than 82% in the largest chromosome of each genome except for *Haloquadratum walsbyi* DSM 16854 and 16790, which were 77.4% and 75.7% respectively. The total ORFs, mean length, and density per chromosome are reported in supplementary table S1, Supplementary Material online, for the 14 fully assembled genomes.

Accurately identifying homologous genes distributed across a set of genomes is crucial for reconstructing their histories. Inteins are parasitic genetic elements that rely on homing endonucleases to insert themselves in chromosomes at highly conserved sites in homologous genes and splice themselves out of the protein product (i.e., not the mRNA) without disrupting gene function (Gogarten et al. 2002). They were screened for and removed from this data set because if they are not present in all members of a gene family, they may cause artifacts in sequence-based comparative analyses including homology inference and phylogenetic reconstruction. Those identified and removed are listed in supplementary table S2, Supplementary Material online. Seventy-one sequences from 20 integration sites were removed across 10 gene families (as annotated by KAAS, Moriya et al. 2007).

After assigning homologs to broad clusters (superfamilies), we used the phylogeny-aware BranchClust algorithm (Poptsova and Gogarten 2007) to identify gene families, including those with several homologs per genome. The strict core, defined as present in all 21 genomes, revealed 893 gene families of which 643 only had single homologs per genome. An additional nine had recent, lineage-specific paralogs not affecting the relationships between genomes. However, we included draft genomes, which by their nature are incomplete, and a strict core analysis may overlook genes that are present but not sequenced or annotated. To ameliorate this potential problem and increase the sample size, we examined a relaxed core, defined here as gene families present in 15 or more of the 21 sampled genomes. The relaxed core comprised 1,814 gene families, of which 1,000 had a single copy in each occupied genome. An additional 19 that had lineage-specific paralogs were also included providing 1,019 total gene families for all analyses presented later. Restricting this analysis to single-copy genes avoids the ambiguity of having one genome represented in several places in a topology but still retains a large data set from which to infer evolutionary processes.

Phylogenies from Genome Sequences

We desired a phylogeny representing the vertical descent of populations from which each genome sequence was sampled to infer incongruent gene family phylogenies as HGTs. Any gene family may have experienced HGT, so such a reconstruction required a larger set of characters than that of a single set of gene sequences.

Concatenated Ribosomal Protein Phylogeny

Associated with high levels of cytosolic potassium ions, haloarchaeal proteins contain an over abundance of acidic amino acids (Danson and Hough 1997). Most current models of sequence evolution assume stationarity in the substitution matrix requiring compositional homogeneity across the data set. Unfortunately, this assumption is likely violated and the accuracy of the rooting compromised when using an

outgroup containing less acidic proteins. Posterior predictive resampling under a Bayesian framework (Lartillot and Philippe 2004) was used to test whether a model that assumes stationarity fits the composition of each ribosomal protein (Foster 2004) to identify sequences that may cause model violations and compromise the accuracy of phylogenetic reconstruction. The 56 ribosomal proteins shared with a new candidate class of halophilic Euryarchaea (Nanohaloarchaea: Narasingarao et al. 2012) were first tested because adaptation to their hypersaline environment may have led to similar molecular adaptations as in the haloarchaeal genes (i.e., class Halobacteria) and therefore similar amino acid composition. Eight *Candidatus* Nanosalina sp. J07AB43 and seven *Candidatus* Nanosalinarum sp. J07AB56 sequences failed the test (supplementary table S3, Supplementary Material online). To mitigate further systematic error caused by a long edge leading to the outgroup, two additional Euryarchaea were tested: *M. acetivorans* C2A and *Met. aeolicus* Nankai 3. The former failed on 28 sequences, whereas the latter failed on 43, so the outgroup was constructed using two representatives from Nanohaloarchaea and one from *Methanosarcina*, omitting sequences that failed the compositional homogeneity test.

ML reconstruction of the concatenated ribosomal protein phylogeny using RAxML (Stamatakis 2006) (supplementary fig. S1, Supplementary Material online) and a Bayesian relaxed molecular clock phylogenetic reconstruction using BEAST v1.6.1 (Drummond and Rambaut 2007) placed the genus *Halobacterium* as the deepest haloarchaeal lineage with a bipartition support of 74% in the former and with the longest edge in both. Omitting the outgroup postreconstruction put *Halobacterium* as a cousin to a clan (sensu Wilkinson et al. 2007) consisting of *Natronomonas*, *Halorhabdus*, *Halomicrobium,* and *Haloarcula* and to a clan consisting of *Halalkalicoccus*, *Halopiger*, *Haloterrigena*, *Natrialba*, *Halorubrum*, *Haloferax*, *Halogeometricum,* and *Haloquadratum* but when omitting the outgroup prereconstruction this arrangement changed, placing *Halobacterium* within the second of these two clans (supplementary fig. S2, Supplementary Material online). The original placement of the root was consistent with an LBAA (e.g., see Felsenstein 1978) and therefore in doubt. A reconstruction omitting the *Halobacterium* sequences placed the root on the edge between *Natronomonas*, and *Halorhabdus*, *Halomicrobium,* and *Haloarcula* with bipartition support of 40% on the edges leading to and sister to *Natronomonas* (supplementary fig. S3, Supplementary Material online), thus the placement of the root was conclusive only in as far as resolving five basal lineages (fig. 1). This inference was corroborated by a Bayesian reconstruction (supplementary fig. S4, Supplementary Material online) as implemented in PhyloBayes 3.3b, an algorithm previously shown to be less susceptible to LBAA (Lartillot et al. 2007).

Genome Composition, Organization, and Core-Embedded Quartet Phylogenies

The MP phylogeny inferred from genome gene family composition (encoding presence or absence as a character) differed from the concatenated ribosomal protein phylogeny in the placement of *Haloquadratum* one split further from *Halogeometricum* and the placement *Halopiger* and *Haloterrigena* as cousins, whereas *Natrialba* and *Haloterrigena* were sisters in the rooted concatenated ribosomal protein phylogeny (supplementary fig. S5, Supplementary Material online). Among the observed lineages, the placement of *Halobacterium*, *Halorhabdus*, *Natronomonas,* and *Halorubrum* was unresolved, as was the relationship among *Haloarcula sinaiiensis*, *Haloarcula Marismortui,* and *Haloarcula californiae* within Lineage A. Identical differences were observed using an ML reconstruction of the same gene family presence or absence character encoding, and *Halorubrum* and *Halorhabdus* were placed outside of their respective lineages with more confidence (supplementary fig. S6, Supplementary Material online). The MRP supertree of plurality topologies of embedded quartets had maximum bootstrap and Bremer bipartition support and differed from the concatenated ribosomal protein phylogeny in the placement only of *Halopiger* and *Haloterrigena* as cousins (supplementary fig. S7, Supplementary Material online). A chromosomal rearrangement phylogeny inferred from a "double-cut-and-join" rearrangement distance matrix (Lin et al. 2011) differed from the concatenated ribosomal protein phylogeny by placing *Halogeometricum* and *Haloferax* as cousins, *Halalkalicoccus* and *Halobacterium* as cousins, and *Natrialba* and *Halopiger* as cousins (supplementary fig. S8, Supplementary Material online). Notably, the terminal edge to *Halorubrum* was the longest in the phylogeny, and an *x–y* scatter plot of these pairwise rearrangement distances against the concatenated ribosomal protein phylogeny distances shows the distance from *Halorubrum* to the other haloarchaea to be greater than distances within the haloarchaea (supplementary fig. S9, Supplementary Material online). A better sampling of *Halorubrum* genomes could confirm whether this lineage is unusual in its chromosome architecture among the haloarchaea.

Selecting a Proxy for the Phylogeny of Vertical Descent

Though conflict was found among every method or set of characters used to reconstruct trees, the five groups defined by the ribosomal protein-based phylogenies were largely in agreement across the range of different reconstruction methods and genome characters. Some of the within-group relationships were unresolved by some methods or differed in others, but overall, there was enough in common to suggest a signal of vertical descent had been recovered. The ribosomal protein sequence phylogenies have the advantage of outgroup rooting and the availability of sophisticated and

well-characterized algorithms for scaling edge lengths to substitutions per site or time that the other characters do not have. As this additional information was desirable for further analyses, the concatenated ribosomal protein phylogeny was selected as a proxy for lineages. The genomes affected by HGTs between internal edges of an unrooted reference phylogeny cannot be conclusively identified. However, because there was uncertainty associated with the position of the outgroup rooting, the conservative decision of placing the five deepest lineages in an unresolved basal polytomy was taken. This avoided inferences due to unreliable incongruities over these deeper bipartitions. The Bayesian relaxed molecular clock phylogenetic reconstruction from BEAST v1.6.1 used for subsequent analyses of HGTs is shown in figure 1*B*.

Quantification of HGT

We developed a novel algorithm to infer exchange partners and direction of HGT within each gene family by examining how each statistically supported embedded quartet topology conflicts with the same quartet topology in the concatenated ribosomal protein reference phylogeny (≥80% bootstrap score, see Materials and Methods for details). This method for detecting differences in tree phylogenies provides better sensitivity than bipartition (nodal) support approaches because it relies on differences in embedded quartet topologies (see Mao et al. [2012] for a comparison of embedded quartet topology and bipartition in bootstrap replicates). The evolution of homologous genes during diversification of the populations they reside in (orthology) may have included duplications within (paralogy) and transfers between (xenology) those populations resulting in chimeric genomes with complex histories (Gogarten and Townsend 2005). Given the relative rareness of paralogy to xenology in prokaryotes (Treangen and Rocha 2011; see also section Quartet Decomposition), it is most parsimonious to interpret highly supported discordant trees as evidence for HGT.

A large majority of relaxed core gene families (97%) had at least one embedded quartet in conflict with the reference phylogeny at ≥80% BSS. Therefore, 97% of these gene families were affected by HGT at some point during their evolution. Of these 1,019 relaxed core gene family phylogenies, 812 contained enough embedded quartets conflicting with the reference phylogeny for our novel algorithm to infer explicit exchange partners. We detected transfer partners for 1,682 genes among those 812 gene families. Based on a simulation performed by Zhaxybayeva et al. (2006), use of the 80% BSS threshold for statistical support in conflicting embedded quartets provides a conservative trade-off between false-positive and false-negative inferences of HGT events. Increasing the required threshold for conflicts between embedded quartets to ≥85% BSS yields fewer inferred HGTs (92% gene families affected), whereas decreasing to ≥75% BSS yields more (98% gene families affected). According to the simulation of Zhaxybayeva et al. (2006), the higher threshold of ≥85% BSS is likely to miss more HGT events (false negatives), whereas the lower threshold of ≥75% BSS is likely to miss fewer HGT events but report stochastic noise as HGT more often (false positives).

HGT from a nonhaloarchaeon that is found only in a fraction of haloarchaeal genomes will cause a longer than usual edge in gene family phylogeny and may also cause a change in topology because the recipient will be placed next to the root. To account for such events and avoid false inference of HGT among the sampled haloarchaea, we screened for unexpectedly long edges (see Materials and Methods for details). We tested the statistical support for homology between representatives from either side of a long edge to currently available non-Halobacteria genomes by BLAST analysis (Camacho et al. 2009). Members from 34 gene families with better-supported homology to nonhaloarchaea were excluded from subsequent analyses because they were likely to have been acquired by HGT from a nonhaloarchaeon. Two gene families contained members affiliated to Nanohaloarchaea, three to Bacteria, and 11 to other Euryarchaeota. The remainder was ambiguous with respect to donor (see supplementary table S4, Supplementary Material online, for gene families affected by non-Halobacteria HGT donations and supplementary table S5, Supplementary Material online, for gene families with more ambiguous scenarios and the sequences excluded in each case).

We would expect contiguous genes, for instance, genes in operons, to sometimes be transferred together in a single episode. By using the number of conflicts per gene family phylogeny to quantify HGTs and assuming a single ORF is transferred in each HGT event, we are overestimating the actual number of HGTs. To obtain a more accurate estimate for the total number of transfer events, we assigned multiple ORFs from different gene families to the same HGT event if they were adjacent in a chromosome sequence and shared the same phylogenetic incongruities as determined from embedded quartets. In nearly all cases, ORFs appeared to have been transferred individually. We estimated that the 1,652 relaxed core ORFs determined as having been transferred could be explained by 1,610 HGT events of which 1,571 delivered one ORF, 36 delivered two ORFs, and three delivered three ORFs. See the Discussion section for consideration of within ORF recombination events.

Although our estimate of total HGTs was improved, we had only included ORFs from the relaxed core gene families in the embedded quartet topology analysis. Some of these HGT events may also have included other ORFs in less widely distributed gene families. For example, an ancestral ORF inferred to have been transferred by topology analysis might be flanked by nonrelaxed core ORFs in genomes of the extant recipient and donor. To improve the estimate of ORFs per HGT, we assigned these additional nonrelaxed core ORFs to already inferred HGTs depending on their

evolutionary distance to the putative donor. Each had to be more closely related to their homolog in the putative donor than to a genome that, according to the ribosomal protein phylogeny, ought to have the more closely related homolog in the absence of HGT (see Materials and Methods for more details and figure 2 for a visual explanation of a real example). This use of evolutionary distances is similar to the commonly used approach of ratios of BLAST bit scores to infer HGTs (see earlier and supplementary tables S4 and S5, Supplementary Material online) but is expected to be more accurate because molecular evolution is explicitly modeled in estimating the distance. In this context, it is also more conservative as inferences of horizontal transfer are only made if associated with a transferred ORF from the relaxed core that was inferred using embedded quartet topology analysis. Of the 1,610 inferred HGT events, 65 delivered two ORFs and 38 delivered three or more. The most ORFs transferred in a single HGT was 15. An average 1.12 ORFs were transferred per HGT event. This result was surprising in light of recent in vitro analyses between *Haloferax* species that demonstrated a capacity for enormous fragment size recombination events (e.g., >500 kbp: Naor et al. 2012).

The functional role that a gene product carries out may affect the likelihood of it being retained in a recipient genome. Table 1 lists the proportions of the relaxed core gene families assigned to a single KEGG Orthology Level 1 functional category ("Metabolism," "Genetic Information Processing," "Environmental Information Processing," and "Cellular Processes") and the proportions of horizontally transferred ORFs with a single functional designation. "Metabolism" was slightly over-represented relative to "Genetic Information Processing," supported as significant

Table 1

The Proportion of ORFs in Single Copy Gene Families with 15 or More Members ("the relaxed core") at KEGG Orthology Level 1 Functional Category, Compared with the Proportion of Transferred ORFs in each Category

KEGG Orthology Level 1	All Gene Families in ≥15 Genomes (%)	HGT Events Consisting of ORFs with Single Function (%)
Metabolism	62.0	64.5
Genetic information processing	33.3	30.7
Environmental information Processing	3.9	3.5
Cellular processes	0.8	1.3

NOTE.—The differences are statistically supported by a χ^2 test.

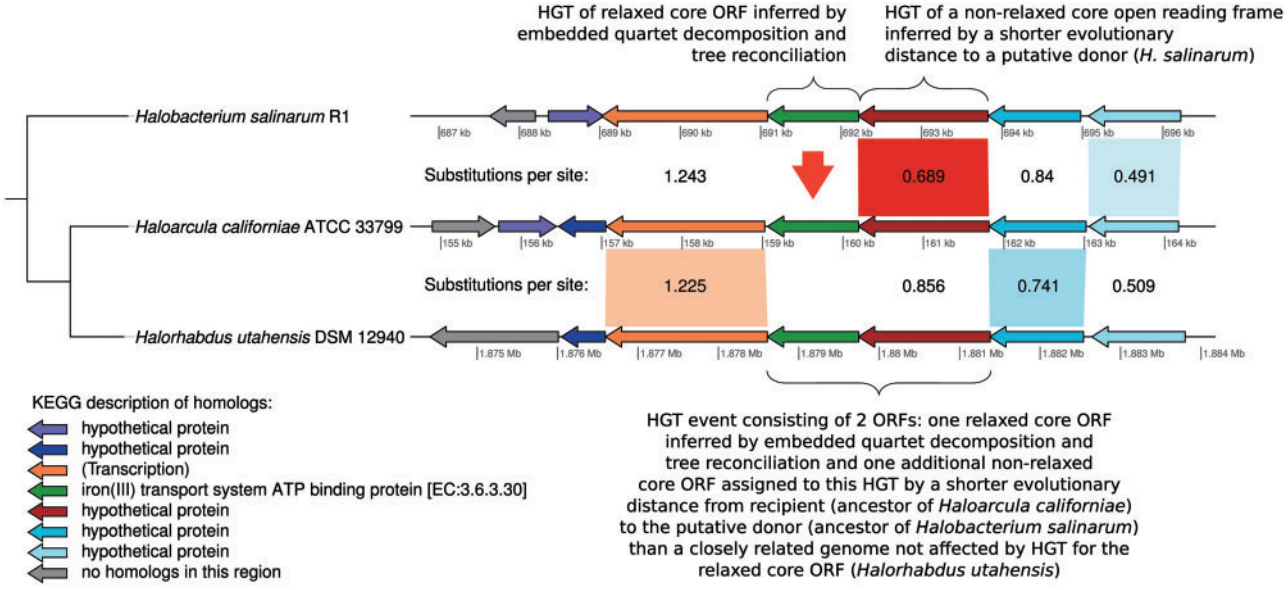

FIG. 2.—A diagram indicating a horizontal transfer of a protein coding ORF inferred by embedded quartet decomposition and tree reconciliation with an adjacent ORF inferred to have been horizontally transferred in the same event. The three horizontal lines represent regions of chromosomes from *Halobacterium salinarum* R1 (top, putative donor of transferred genetic material), *Haloarcula californiae* ATCC 33799 (middle, putative recipient), and *Halorhabdus utahensis* DSM 12940 (bottom, a reference genome). Units are megabases (Mb). Horizontal arrows represent 3′–5′ strand direction and range of protein coding regions. Shared colors indicate most recent homology except for gray, which indicates no local homology. The vertical red arrow indicates which homologs were inferred by embedded quartet decomposition and tree reconciliation to have been transferred between the ancestor of *Halobacterium salinarum* R1 and *Halorhabdus utahensis* DSM 12940 and the direction. The reference genome was selected for being more closely related to the putative recipient than donor according to the ribosomal protein phylogeny, plotted to the left side, and to have not been inferred to have been affected by HGT for the gene analyzed with embedded quartets. ML estimates of evolutionary distances measured in substitutions per site are indicated between homologous protein coding regions with the shorter distance indicated by a color.

by a χ^2 test ($\chi^2 = 4.9152$, df = 1, P value = 0.02662). One hundred twenty-six inferred transfers were between splits that did not exist simultaneously in the Bayesian sampled relaxed molecular clock ribosome phylogeny. In 79 of these transfers, the donor was older, in 26 the recipient was older, whereas in 21, the direction was ambiguous.

Characterization of Chromosomal Integration Following HGT

HR and NHI following an HGT are different processes that can cause similar patterns in evolutionary reconstruction. For instance, genetic material integrated by NHI initially results in the new and original copies coexisting. Later, if this event is followed by a loss of the preexisting original version, the process from a historical point of view is difficult to distinguish from HR, where a gene conversion-like process involving homologous recombination maintains the same number of copies by the immediate and direct replacement of the original. To differentiate between the two processes, we employed gene synteny analysis as a basis for partitioning into bins: we assume an HR would maintain the same genes (orthologs due to shared ancestry) around the transferred gene, whereas NHI would most likely go elsewhere in the recipient chromosome. NHI followed by a loss is a reasonable hypothesis for HGT between divergent organisms (e.g., different genera) due to a lesser dependence on sequence similarity, whereas rates of homologous recombination have been observed to decrease log linearly with increasing genetic distance to low frequencies, even among members of the same genus (Roberts and Cohan 1993; Vulić et al. 1997; Eppley et al. 2007).

Of the 1,610 HGT events, 206 could be tested by this approach (see Materials and Methods). Of those, we estimated that the majority (174 or 90%) was by HR, 19 were by NHI, and 13 were by either HR or NHI depending on which descendants were compared. Chromosomal rearrangements are more likely to make an ancestral HR event appear as an NHI than vice versa, so the HGTs that appeared as either process were considered as HR, raising the estimate to 91% of the testable HGTs. Of the inferred HR, 52% were identified between the five basal lineages (fig. 1), that is, between very divergent haloarchaea. The 1,414 untestable HGTs were either of ambiguous direction, lacked a close relative of the recipient, or did not share identifiably orthologous regions between putative donor, recipient, and close relative (see Materials and Methods). These characteristics determining HGT testability are dependent on the sampling of genomes and do not imply a strong sampling bias in favor of NHI or HR. In 862 of the HGTs, direction could not be inferred, as the exchange partners were within a split of being sister taxa. For this reason, there may be a bias against detection of HR as it is the process expected to be more prevalent among close relatives.

An alternative explanation for incongruent phylogenies is a paralogous duplication of an ORF or contiguous chromosome region in the common ancestor of the putative exchange partners, followed by losses of different paralogs or regions in the respective ancestors of the HGT partners ("hidden paralogy"). Except in the case of a duplication resulting in tandem repeats, one partner would keep the homolog in the original position in the chromosome, whereas the other partner would retain the duplicate elsewhere in the genome. This pattern of disrupted gene order for hidden paralogy is different to the conserved gene positions seen in the HRs via HGT inferred for the majority of statistically supported phylogenetic incongruities found in this study. Thus, the majority of phylogenetic incongruities are likely due to HGT, consistent with the analysis of Treangen and Rocha (2011) in which horizontal transfer, not paralogous duplication, was found to drive the expansion of protein families in bacteria and archaea.

The consequence of HR exchange among disparate lineages on functionally related and chromosomally adjacent genes is depicted in figure 3. A region of eight to nine homologous ORFs conserved across most of the sampled genomes and coding for proteins involved in cobalamin (vitamin B_{12}) biosynthesis was found to have at least seven distinct tree-like histories. Figure 3*A* shows the chromosomal regions in representatives of nine different genera including representatives of four of the basal lineages in the ribosomal protein reference phylogeny of the Halobacteria (fig. 1). The same reference phylogeny is plotted to the left in figure 3*A* and also in a circular form in figure 3*B* in which transfers are depicted in colored arrows illustrating the complexity of the rooted net phylogeny of this chromosomal region (sensu Williams et al. 2011 in which internal nodes represent ancestral states as opposed to a splits network representation of a nontree-like phylogenetic signal). All other inferred HGTs for these genomes are also plotted in gray in figure 3*B*. Given that the majority of HGTs are integrated by HR and all relaxed core genes families had some degree of phylogenetic incongruity, the scenario depicted in figure 3 is likely to be common in haloarchaeal evolution.

Quantifying the Relationship between Relatedness and HGT Frequency

Experimental studies analyzing rates of homologous recombination between closely related bacterial species and genera revealed an inverse log-linear relationship between frequency of recombination and degree of genomic DNA divergence (Roberts and Cohan 1993; Vulić et al. 1997). A similar relationship was inferred from natural archaeal *Ferroplasma* acid mine drainage populations (Eppley et al. 2007). Because the majority of relaxed core HGT events occurred by HR (which invokes homologous recombination), we might expect to see a similar log-linear relationship. To examine this, we estimated HGT frequency from the total inferred HGTs during the time

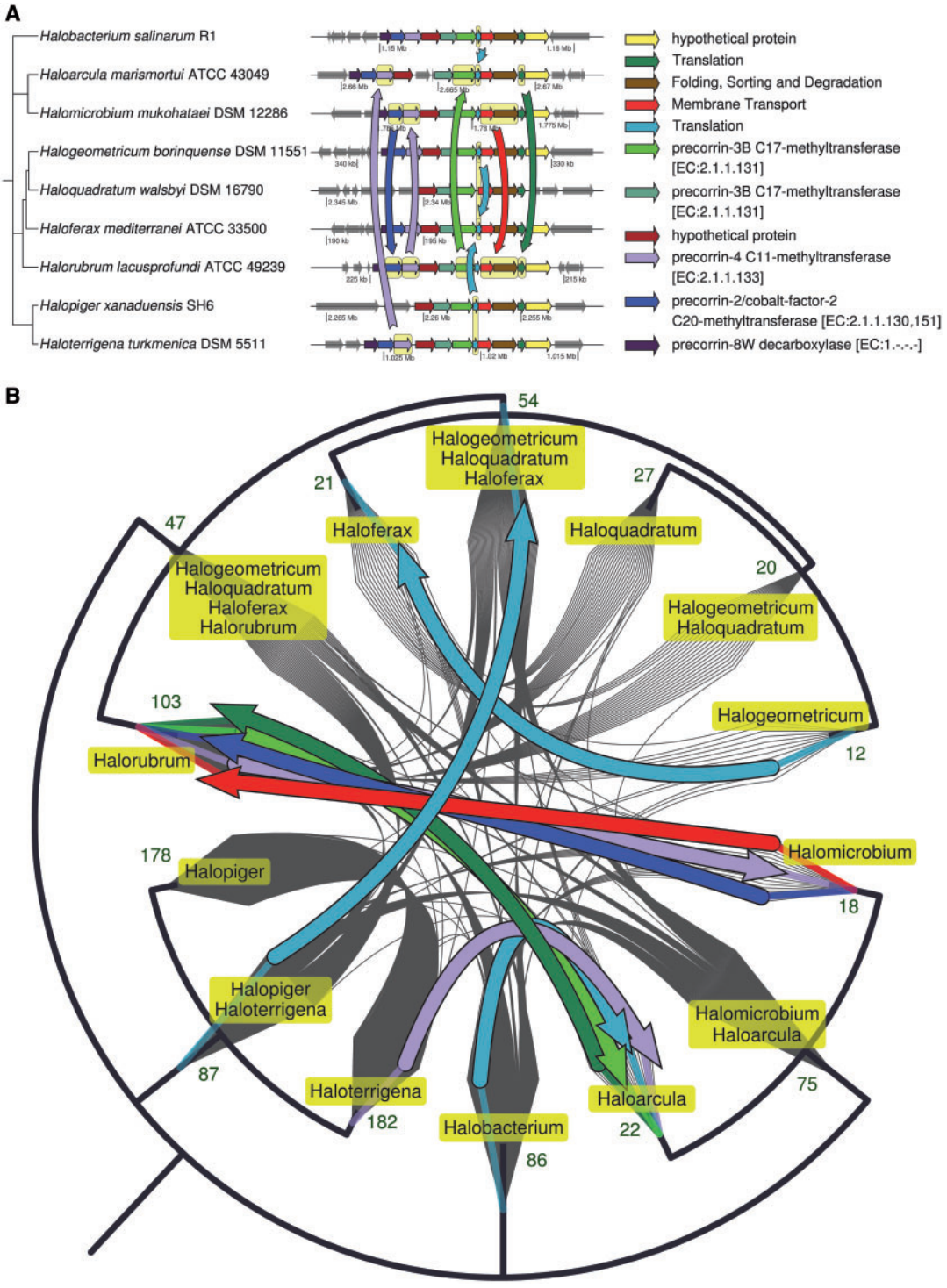

Fig. 3.—(*A*) Directed hierarchical network of extant and ancestral genomes showing HGTs of cobalamin (vitamin B_{12}) biosynthesis genes. The nodes in the network are defined by the ribosomal protein phylogeny of nine haloarchaea, a subset of the full analysis of 21 genomes. The edges of the network are the vertical arrows indicating inferred HGTs by HR and are colored by homology. The network is arranged to also show the chromosomal regions coding for the proteins involved in cobalamin (vitamin B_{12}) biosynthesis. The phylogeny edge length units are arbitrary time units; the chromosome map scale units are megabases (Mb). Horizontal arrows depict ORFs in 3′–5′ strand direction and are scaled to base pairs. Shared colors indicate most recent homology except for gray, which indicates no local homology. (*B*) Directed hierarchical network depicting the same HGTs as in figure 1(*A*) with arrows also colored by most recent homology. Gray lines indicate all other inferred HGTs, the directions of which are not indicated to maintain clarity. The green numbers indicate the total HGT events to or from that ancestral population.

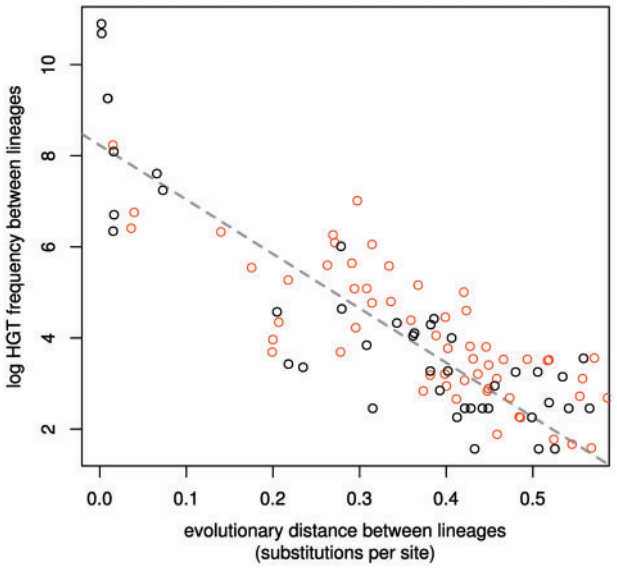

Fig. 4.—Scatter plot of frequency of HGT events between two lineages versus the distance across the ribosome phylogeny in substitutions per site. Black: HGT between terminal edges and red: HGT between internal edges. Frequency was calculated as the total inferred HGT events between two edges on the ribosome phylogeny per overlapping edge length when scaled to a relaxed molecular clock. The dashed line is a linear model fitted after natural log transformation of the frequencies between terminal edges, $R^2 = 0.78$, $P < 6 \times 10^{-15}$. Terminal edges are considered more reliable than internal edges because tips are at present day. Fitted HGT frequencies between all edges, $R^2 = 0.72$, $P < 1 \times 10^{-16}$. An inverse log-linear relationship for recombination rates versus genome similarity has also been observed in experimental studies and inferred from sequence data of closely related bacterial genera.

of coexistence and genetic distance as amino acid substitutions per site in the concatenated ribosomal protein phylogeny (see fig. 1 for an example). Figure 4 shows a scatter plot of the natural log of HGT frequency between coexisting lineages and the distance across the ribosomal protein phylogeny. A linear model fitted to these data returns an R^2 of 0.72 ($P < 1 \times 10^{-16}$) demonstrating an inverse log-linear relationship across the diversity and evolution of the class Halobacteria (*y*-axis intercept $= 8.2 \pm 0.4$, slope $= -11.9 \pm 0.6$ standard error). HGT rates estimated between internal edges of the phylogeny (such as the region labeled "t" in fig. 1*B*) will suffer from inaccuracies in the relaxed molecular clock estimation more than terminal edges. A linear model fitted to log-transformed HGT rates estimated only from the more reliable terminal edges returned an improved R^2 of 0.78 ($P < 6 \times 10^{-15}$) with almost unchanged *y*-axis intercept $= 7.9 \pm 0.25$, and slope $= -10.4 \pm 0.6$. The log linearity of this relationship is similar to that seen in other bacteria and archaea within genera: to our knowledge, this is the first time such a relationship has been witnessed over evolutionary distances that span a Class.

Quantifying Chimerism as Horizontally Transferred Amino Acid Substitutions

HR delivers amino acid substitutions into a genome that were fixed in a different population under potentially very different evolutionary conditions. For example, HGT by HR from the ancestor of *Haloarcula* and *Halomicrobium* delivers to the ancestor of *Halalkalicoccus* the sequence substitutions that evolved in the donor lineage (d3 + d4 in fig. 1*A*). These amino acid changes, referred to hereafter as "out of population" substitutions, erase those in the recipient's gene that have evolved since divergence from the donor lineage (d1 + d2 in fig. 1*A*). Even if function has been conserved between the transferred and replaced genes, the genomic, cellular, and ecological context will be different between the donor and recipient populations, such that a different evolutionary trajectory may have been followed. In the absence of HGT, the substitutions per site from the root to each tip of the ribosomal protein phylogeny would represent the total fixed mutations originating in the same population, referred to hereafter as "within population" substitutions.

The evolutionary significance of HGT-acquired "out of population" substitutions was assessed by quantification relative to the "within population" substitutions for each sampled genome since the last common ancestor of the haloarchaea. The contribution of "out of population" evolution depends on the total HGTs between all pairs of ancestors (the time of coexistence of transfer partners multiplied by the frequency of HGTs between them) and the quantity of substitutions delivered per HGT. HGT frequency was estimated using the linear model fitted above allowing HGT frequency estimates between sister lineages. The time of lineage coexistence, evolutionary distance between transfer partners, and quantity of transferred substitutions corresponded to appropriate edge lengths in the phylogenies of figure 1 (see Materials and Methods for details).

Halalkalicoccus jeotgali B3 had the least with 10.97% of its total substitutions in relaxed core genes coming via HGT and *Halopiger xanaduensis* SH-6 had the most with 20.33%. The mean was $14.9 \pm 2.4\%$ although each estimate is not phylogenetically independent. Proportions of out of population substitutions for all sampled genomes are listed in table 2. The true genome-wide diversity achieved by out-of-lineage acquisitions depends on the diversity of the donors: all from a single donor will make the recipient more like the donor but from an even distribution of donors has a greater potential for novel genetic combinations in different populations. Table 3 lists the distribution of transfers between the basal lineages to be even, allowing for the uneven sampling between lineages, so that functionally related combinations of "out of population" substitutions may often come from divergent donors as in the evolution of the cobalamin biosynthesis genes depicted in figure 3.

Table 2

Measures of Chimerism for each Sampled Haloarchaeon: the Percentage of Substitutions in Relaxed Core Genes that Occurred in Other Populations (different haloarchaeal species) but Were Subsequently Delivered by HGT to the Genome in Question during the Evolution of the Haloarchaea

Genome	Lineage	Percentage of Substitutions in Relaxed Core Genes[a] Acquired from Other Lineages
Haloarcula californiae ATCC 33799	A	13.16
Haloarcula marismortui ATCC 43049	A	13.14
Haloarcula sinaiiensis ATCC 33800	A	13.05
Haloarcula vallismortis ATCC 29715	A	13.46
Halomicrobium mukohataei DSM 12286	A	14.16
Halorhabdus utahensis DSM 12940	A	16.11
Halalkalicoccus jeotgali B3	B	10.97
Halopiger xanaduensis SH-6	B	20.33
Haloterrigena turkmenica DSM 5511	B	14.59
Natrialba magadii ATCC 43099	B	14.44
Haloferax denitrificans ATCC 35960	C	17.19
Haloferax mediterranei ATCC 33500	C	15.99
Haloferax mucosum ATCC BAA 1512	C	16.30
Haloferax sulfurifontis ATCC BAA 897	C	17.24
Haloferax volcanii DS2	C	17.11
Halogeometricum borinquense DSM 11551	C	17.06
Haloquadratum walsbyi C23	C	11.58
Haloquadratum walsbyi DSM 16790	C	11.58
Halorubrum lacusprofundi ATCC 49239	C	14.84
Natronomonas pharaonis DSM 2160	D	12.34
Halobacterium salinarum R1	E	18.09

NOTE.—The remaining substitutions were caused by mutations occurring and fixed in the population from which the genome in question was sampled (see Results and Materials and Methods for more details).

[a]Gene families in the relaxed core are defined here as single copy per genome and appearing in 15 or more of the 21 analyzed genomes.

Table 3

The Quantities of HGTs between the Basal Lineages, as Defined by the Ribosomal Protein Phylogeny, of the Halobacteria, that is, HGTs Crossing the Deepest Evolutionary Divergences in the Group

Basal Lineages	E	D	C	B
A	43[a]	75	87	121
B	67	60	161	
C	52	45		
D	0			

NOTE.—Lineages with greater sampling have more complex topologies and more opportunity to infer transfers, whereas transfers between D and E, single member sister lineages, cannot be detected by phylogenetic incongruity. Allowing for uneven phylogenetic sampling, a strong bias in HGT between certain lineages is not apparent.

[a]HGTs between lineages.

Discussion

Identifying an appropriate surrogate for the phylogeny of vertical inheritance was crucial to inferring HGT events. We selected the phylogeny inferred from a concatenation of ribosomal proteins by ML (Stamatakis 2006) and Bayesian sampling (Drummond et al. 2006) because it was largely consistent with various other reconstruction methods using multiple sets of genome characters, it could be rooted, and its edges could be scaled to substitutions per site and a relaxed molecular clock (fig. 1). Not surprisingly (e.g., see Brochier et al. 2000; Omelchenko et al. 2003; Zhaxybayeva, Doolittle, et al. 2009; Yoon et al. 2011), many haloarchaeal ribosomal proteins were transferred between numerous edges in this analysis. However, the supertree constructed from the plurality signal resulting from the quartet decomposition analysis, a distance tree based on chromosomal rearrangements, and trees inferred from the presence or absence of gene families are in general agreement with the concatenated ribosomal protein tree, suggesting this phylogeny is suitable for comparing phylogenies of other chromosomal genes. It is important to acknowledge that this is not a phylogeny of the genomes or organisms. It is, however, reasonable to assert that the concatenated ribosomal protein phylogeny is largely representative of the vertical component of the evolution of this group as it is derived from a stable, coevolving set of proteins. The fact that 97% of the gene family phylogenies were to some extent incongruent with this phylogeny supports the claim that a tree is not a realistic model for microbial evolution (Hilario and Gogarten 1993; Martin 1999). The type of directed network reconstructed here is related to that reconstructed by Popa et al. (2011; Dagan 2011) for 657 bacterial and archaeal genomes. In addition to the recent transfers analyzed by Popa et al., we inferred ancestral HGT events. This preserved transfers between nodes representing ancestral donors and recipients and their nested hierarchical relationship to extant genomes.

Although several examples of direct HR between divergent partners have been identified in Bacteria and Archaea (Brochier et al. 2000; Omelchenko et al. 2003; Yoon et al. 2011), homologous recombination is regarded as most relevant, or even limited to, a population genetics context where it modulates "vertical" genetic transfer within a population (Lan and Reeves 2001; Didelot and Maiden 2010; Lawrence and Retchless 2010). Similarly, HGT between divergent populations is often regarded as more relevant to acquisition of nonhomologous genetic material and new traits in a phylogenomic context (Lawrence and Ochman 1998; Tenaillon et al. 2010; Coscollá et al. 2011; Zhaxybayeva and Doolittle 2011). In replacing a homologous gene, HGT can involve NHI followed by subsequent loss of the original (Zhaxybayeva and Doolittle 2011). However, we infer through gene position analysis that most of the relaxed core HGTs between species in the class Halobacteria were by direct HR (i.e., homologous recombination) in a gene conversion-like process.

In contrast to detecting different topologies between adjacent sequences, Chan et al. (2009) found evidence of

recombination breakpoints within single gene sequences. Different regions of a gene were inferred to have different histories caused by homologous recombination. Of 1,462 single-copy gene families spanning Bacteria and Archaea, they detected 20% as having internal recombination breakpoints. HGT with HR affecting only regions within a coding sequence included in the present analysis is likely to go undetected: differing phylogenetic signals over different regions of a multiple sequence alignment will lead to differing topologies among those inferred from bootstrap replicate sequence alignments. This will lower embedded quartet topology bootstrap scores, so that topological conflicts indicative of HGT cannot be detected, thus total HGTs and genome chimerism are likely to have been underestimated in this analysis. If an extinct or unsampled ancestor had been involved in HGT with an ancestor of one of the sampled genomes, our algorithm would infer the extinct lineage to be at the edge where it would have joined had its ribosomal proteins been included in the reference phylogeny. Of the 126 transfers between edges in the reference phylogeny (fig. 1*B*) that did not coexist, 79 could be interpreted as having originated in an extinct or unsampled donor. In this case, the donor is mapped to the branch point between the unsampled lineage and a lineage included in our analysis. The 26 cases, in which the recipient appears older than the donor, may be artifacts due to systematic error or examples of what appears to be a single transfer is in fact two or more transfers that involve intermediate lineages not represented in our data set.

We estimated the frequency of inferred HGTs as a function of evolutionary distance by combining two versions of the reference phylogeny. Scaled to substitutions per site, the reference phylogeny provided estimates of evolutionary distance between its transfer partners (edges), while scaled to a relaxed molecular clock, estimates of duration of exchange partner coexistence were obtained. The resulting inverse log-linear relationship represents the combined rates of HGT, successful HR (in most cases), and fixation in the recipient population between exchange partners of varying divergence. The limiting factor among these three processes cannot be directly inferred from these data. The steep canonical log-linear relationship for recombination versus genetic distance observed for closely related bacteria led to the hypothesis that sequence divergence alone can act as a barrier to recombination and can form the basis for prokaryotic species and speciation (Vulić et al. 1997; Lan and Reeves 2001). Indeed, this exact relationship has been used to model speciation in computer simulations (e.g., Hanage et al. 2006; Fraser et al. 2007). Recent in vitro frequency analysis between two haloarchaeal species with ~14% nucleotide divergence showed that recombination between them was much higher than expected and that measured for Bacteria (Naor et al. 2012). The work reported here corroborates and extends those previous observations and together they imply that homologous recombination may still be a relevant process at evolutionary distances far greater than previously tested in haloarchaea and perhaps expected for either Bacteria or Archaea. Because members of Halobacteria readily form heterodiploids at high frequency between species (Naor et al. 2012) and have the capability for laboratory genetic manipulations (Cline and Doolittle 1992; Allers and Mevarech 2005), they are a good model for the genetics of the archaeal-like protoeukaryote ancestor of the eukaryotic nucleocytoplasm, even though there is no indication that the haloarchaea themselves are a phylogenetic neighbor of Eukarya.

We were able to estimate the proportion of substitutions in each genome that were originally fixed in other populations and subsequently transferred because the frequency of HGTs could be predicted by substitutions per site in the concatenated ribosomal protein phylogeny according to an inverse log-linear relationship among all sampled members of the class Halobacteria. This model provided estimates of HGT frequency between sister lineages, which is not possible by analyzing phylogenetic incongruities alone. Although HGTs between divergent partners with fixation in the recipient population are relatively rare, they deliver a relatively large number of substitutions per site, contributing to the substantial proportions of substitutions estimated to have occurred in other lineages (up to 21%). Conversely, though introducing fewer substitutions per HGT, recombination between close relatives occurs at a much higher frequency.

Is it surprising that these horizontally transferred, relaxed core genes are fixed in the recipient population often enough to contribute so many substitutions? The nearly neutral model of molecular evolution (Ohta 1973, 1992) states that substitutions, including those that are slightly deleterious, may be fixed in a population through random genetic drift. We could thus speculate that all the "out of population" substitutions delivered by HGT are in fact nearly neutral and do not contribute to the adaptive evolution of members of a population. However, prokaryotic effective population sizes are enormous (Lynch and Conery 2003) and drift only rarely fixes slightly deleterious genes when that is the case.

There are several reasons to think that HRs can be more than "slightly deleterious." The frequency of HGTs has been inferred to be higher when codon usage is similar between recipient and transferred genetic material implying mismatches in codon usage are deleterious and selected against (Medrano-Soto et al. 2004; Tuller et al. 2011). The effective number of codons used (N_c; Wright 1990) is a measure of codon bias where a maximum score of 61 signifies even usage to a minimum of 20 where one codon is used per amino acid. Even among these haloarchaeal genetic exchange partners, there is substantial codon usage bias. For the main chromosomes of genomes included in this analysis, N_c ranges from 33.556 to 52.573, whereas for taxonomically and ecologically diverse group of three Bacteria (*E. coli* O157 H7 Sakai, *Synechococcus* WH 7803, and *Salinibacter ruber* DSM 13855), N_c ranges from 38.121 to 49.542 (supplementary table S6,

Supplementary Material online; supplementary tables S7–S30, Supplementary Material online, contain the codon frequencies).

Numerous reports suggest that HGT with replacement of genes coding for functioning adaptive proteins are potentially highly disruptive. Yoon et al. (2011) inferred a recent horizontally acquired L29 ribosomal protein with contrasting codon usage to surrounding genes in *Sulfolobus solfataricus* P2. Following HR, this acquisition had apparently disrupted the transcriptome architecture, such that two mRNAs were generated instead of one. An experimental study in *E. coli* suggests that unsuitable codon usage in a newly acquired gene can be overcome if it conveys enough of an increase in fitness (Amorós-Moya et al. 2010). Coevolution among proteins in the same genome undoubtedly occurs and is the basis to predict protein–protein interactions from agreement in phylogeny (Pazos and Valencia 2001), but the HGTs with HR reported here cause many genes, possibly those involved in the same functions, to have different phylogenies. Cohen et al. (2011) found higher levels of protein–protein interactions predicted a lower propensity for transfer, yet examples such as the cobalamin synthesis pathway genes in figure 3 suggest some may have higher than expected frequencies.

Despite this potential for disruption, our evidence indicates that most of the relaxed core genes in haloarchaea have a history of HGT with replacement at a locus and fixation in the population. This seeming contradiction can be resolved either by assuming that the observed HR events are selectively neutral (the ones that are deleterious due to disrupted coevolution or different codon usage have not been fixed) or by assuming that the observed replacements convey a net adaptive benefit. In contrast to HGT of nonhomologous material providing new traits, many of the HGTs inferred here might have increased fitness in the recipient by altering existing phenotypic characters.

The products of the genes whose combined reticulate history is depicted in figure 3 are involved in the early stages of de novo synthesis of the most complex coenzyme known, coenzyme B_{12}, and are part of a long pathway involving many gene products (Warren et al. 2002). *Halobacterium* NRC-1 is able to synthesize coenzyme B_{12} de novo and to scavenge precursors from its environment (Woodson et al. 2003; Woodson and Escalante-Semerena 2004). The reaction kinetics of each enzyme in a pathway contributes to the overall efficiency of the process. Ambient cellular conditions such as pH or salinity can affect the reaction kinetics of enzymes. Amino acid substitutions that alter the folding and conformation of an enzyme can alter the kinetic response to ambient conditions, but specific predictions in the context of a complex pathway like that of coenzyme B_{12} synthesis are challenging. Nonetheless, we may speculate that amino acid substitutions following genetic mutation or HR via HGT as reported here may be in response to changes in cellular conditions caused by other evolutionary changes in the host cell or ecological factors including changes in salinity. Alternatively, when coenzyme B_{12} or its precursors are available in the environment, some or all the ORFs coding for proteins involved in its synthesis may be under relaxed purifying selection. Accumulation of deleterious mutations may subsequently be fixed by alleles acquired from other populations in which purifying selection had been maintained. This would appear to be the same allele shuffling process that Spratt (1988) observed for antibiotic resistance in human pathogens, except the exchange partners are known to be divergent and the magnitude of changes introduced in a single recombination event is much larger.

Sewell Wright (1932) introduced the concept of the location of an individual in a "fitness landscape" being determined by the combination of alleles they possess. Wright suggested that the problem of evolution is a species finding its way from lower peaks (local regions of maximum fitness) to higher peaks, across valleys of allele combinations with poor fitness (negative epistasis). The adaptive peaks themselves may also move depending on the combination of alleles possessed, so that the fitness effect of a change in any one allele depends on its genetic background, an effect known as sign epistasis (Weinreich et al. 2005). In contrast to HGT of nonhomologous material providing entirely new loci, HR of a region of chromosome with divergent genetic material, as inferred among these haloarchaeal genomes, can introduce exotic alleles into a population.

Experimental studies have characterized the effect of cumulative mutations on the fitness landscape. Lunzer et al. (2005) demonstrated that protein adaptive evolution through cumulative substitutions could get stuck in local suboptimal adaptive peaks. Weinreich et al. (2006) demonstrated that even in a single peak landscape, 105 of 120 mutational pathways toward a particular 100,000-fold increase in antibiotic resistance were inaccessible due to the intermediate combinations of amino acids providing no gain in fitness. In this study, the HGT acquired alleles that have been exchanged between haloarchaeal species have much greater sequence divergence than conventional alleles. The effect on an individual's ability to traverse between adaptive peaks may be correspondingly different. The sign epistasis caused by contrasting genetic contexts between an HGT donor and the recipient species may allow mutational pathways to be taken in the donor's allele that were inaccessible in the recipient. Thus, acquisition of an exotic allele by HGT with HR may provide access to regions of the fitness landscape that within-population mutations cannot. Conversely, such acquisitions may reverse the accumulation of deleterious mutations following periods of relaxed purifying selection.

Rates of HGT are likely to be higher where potential partners are in close proximity. Blooms of haloarchaea are common (Boujelben et al. 2012), even to the extent that their carotenoids cause hypersaline lakes to become visibly pink or orange. Community analyses of various hypersaline

environments have detected numerous different haloarchaeal members, thus opportunities for HGT would seem abundant (Boujelben et al. 2012). Furthermore, changes in haloarchaeal community composition have been observed over time (Boujelben et al. 2012), further increasing the mixing among different haloarchaea. High metabolic diversity (Falb et al. 2008) and the taxonomically patchy distribution of traits such as gas vesicle formation and flagella (Bolhuis et al. 2006; Hartman et al. 2010) among the haloarchaea is consistent with niche specialization. These differences imply that adaptive evolution has contributed to the diversification of the haloarchaea into varied niches, thus an allele acquired via HGT from a different and distantly related population and integrated via HR may be exotic in terms of ecology and genetics.

An important observation is that chromosomally adjacent and functionally related genes may be replaced by versions from different, disparate donors (e.g., the cobalamin synthesis pathway genes in fig. 3) creating diverse genetic combinations in alien genomic, cellular, and ecological contexts. This mixture of donors providing the "out of population" substitutions may augment the contribution to the adaptive evolution of any one lineage beyond what the proportions listed in table 2 imply (Omelchenko et al. 2003). The effect of this process reminds us of the mythological Chimera, a creature composed of different animals, although our observations are of a different domain of life and perhaps include thousands of different organisms.

Supplementary Material

Supplementary tables S1–S30 and figures S1–S9 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank two anonymous reviewers for valuable and constructive comments and the bioinformatics facility of the Bioservices Center of the University of Connecticut for computational facilities. The version of TNT used was that of the Willi Hennig Society. This work was supported by the National Science Foundation Grant (DEB 0830024). Additional funding was provided through the U.S.–Israel Binational Science Foundation (BSF) (award number 2007043) to R.T.P., the National Science Foundation (award number DEB0919290) to R.T.P. and NASA Astrobiology: Exobiology and Evolutionary Biology Program Element (Grant Number NNX12AD70G) to R.T.P.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.

Aivaliotis M, et al. 2007. Large-scale identification of N-terminal peptides in the halophilic archaea *Halobacterium salinarum* and *Natronomonas pharaonis*. J Proteome Res. 6:2195–2204.

Allers T, Mevarech M. 2005. Archaeal genetics—the third way. Nat Rev Genet. 6:58–73.

Amorós-Moya D, Bedhomme S, Hermann M, Bravo I. 2010. Evolution in regulatory regions rapidly compensates the cost of nonoptimal codon usage. Mol Biol Evol. 27:2141–2151.

Andam C, Harlow T, Papke RT, Gogarten JP. 2012. Ancient origin of the divergent forms of leucyl-tRNA synthetases in Halobacteriales. BMC Evol Biol. 12:85.

Anderson I, et al. 2011. Novel insights into the diversity of catabolic metabolism from ten haloarchaeal genomes. PLoS One 6:e20237.

Anderson IJ, et al. 2009. Complete genome sequence of *Halorhabdus utahensis* type strain (AX-2T). Stand Genomic Sci. 1:218–225.

Anderson IJ, et al. 2012. Complete genome sequence of *Halopiger xanaduensis* type strain (SH-6T). Stand Genomic Sci. 6:31–42.

Antón J, Llobet-Brossa E, Rodréguez-Valera F, Amann R. 1999. Fluorescence in situ hybridization analysis of the prokaryotic community inhabiting crystallizer ponds. Environ Microbiol. 1:517–523.

Arahal DR, Ludwig W, Schleifer KH, Ventosa A. 2002. Phylogeny of the family Halomonadaceae based on 23S and 16S rDNA sequence analyses. Int J Syst Evol Microbiol. 52:241–249.

Ayres DL, et al. 2011. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst Biol. 61:170–173.

Aziz R, et al. 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Baliga NS, et al. 2004. Genome sequence of *Haloarcula marismortui*: a halophilic archaeon from the Dead Sea. Genome Res. 14: 2221–2234.

Bardavid RE, Oren A. 2008. Dihydroxyacetone metabolism in *Salinibacter ruber* and in *Haloquadratum walsbyi*. Extremophiles 12:125–131.

Baum BR. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3–10.

Bolhuis H, Poele EM, Rodriguez-Valera F. 2004. Isolation and cultivation of Walsby's square archaeon. Environ Microbiol. 6:1287–1291.

Bolhuis H, et al. 2006. The genome of the square archaeon *Haloquadratum walsbyi*: life at the limits of water activity. BMC Genomics 7:169.

Boucher Y, Douady C, Sharma A, Kamekura M, Doolittle W. 2004. Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. J Bacteriol. 186:3980–3990.

Boucher Y, et al. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet. 37:283–328.

Boujelben I, et al. 2012. Spatial and seasonal prokaryotic community dynamics in ponds of increasing salinity of Sfax solar saltern in Tunisia. Antonie Leeuwenhoek. 101:845–857.

Bremer K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42:795–803.

Brochier C, Philippe H, Moreira D. 2000. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Genet. 16:529–533.

Buneman P. 1974. A note on the metric properties of trees. J Comb Theory A. 17:48–50.

Burns DG, et al. 2007. *Haloquadratum walsbyi* gen. nov., sp. nov., the square haloarchaeon of Walsby, isolated from saltern crystallizers in Australia and Spain. Int J Syst Evol Microbiol. 57:387–392.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Chan C, Darling A, Beiko R, Ragan M. 2009. Are protein domains modules of lateral genetic transfer? PLoS One 4:e4524.

Charif D, Lobry J. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences

retrieval and analysis. In: Bastolla U, Porto M, Roman H, Vendruscolo M, editors. Structural approaches to sequence evolution: molecules, networks, populations. New York: Springer-Verlag. p. 207–232.
Cline SW, Doolittle WF. 1992. Transformation of members of the genus *Haloarcula* with shuttle vectors based on *Halobacterium halobium* and *Haloferax volcanii* plasmid replicons. J Bacteriol. 174: 1076–1080.
Cock P, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423.
Cohen O, Gophna U, Pupko T. 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol. 28:1481–1489.
Coscollá M, Comas I, González-Candelas F. 2011. Quantifying nonvertical inheritance in the evolution of *Legionella pneumophila*. Mol Biol Evol. 28:985–1001.
Cuadros-Orellana S, et al. 2007. Genomic plasticity in prokaryotes: the case of the square haloarchaeon. ISME J. 1:235–245.
Dagan T. 2011. Phylogenomic networks. Trends Microbiol. 19:483–491.
Danson MJ, Hough DW. 1997. The structural basis of protein halophilicity. Comp Biochem Physiol A Physiol. 117:307–312.
DasSarma S, Capes M, DasSarma P. 2009. Haloarchaeal megaplasmids; In: Schwartz E, editor. Microbiology monographs: microbial megaplasmids. Berlin (Germany): Springer-Verlag. p. 3–30.
Didelot X, Maiden MCJ. 2010. Impact of recombination on bacterial evolution. Trends Microbiol. 18:315–322.
Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science 284:2124–2128.
Doolittle WF, Zhaxybayeva O. 2009. On the origin of prokaryotic species. Genome 19:744–756.
Drummond A, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7:214.
Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.
Dyall-Smith ML, et al. 2011. *Haloquadratum walsbyi*: limited diversity in a global pond. PLoS One 6:e20968.
Edgar R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30: 1575–1584.
Eppley J, Tyson G, Getz W, Banfield J. 2007. Genetic exchange across a species boundary in the archaeal genus Ferroplasma. Genetics 177: 407–416.
Falb M, et al. 2005. Living with two extremes: conclusions from the genome sequence of *Natronomonas pharaonis*. Genome Res. 15: 1336–1343.
Falb M, et al. 2008. Metabolism of halophilic archaea. Extremophiles 12: 177–196.
Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively mislead. Syst Zool. 27:401–410.
Foster PG. 2004. Modeling compositional heterogeneity. Syst Biol. 53: 485–495.
Fournier G, Gogarten J. 2008. Evolution of acetoclastic methanogenesis in Methanosarcina via horizontal gene transfer from cellulolytic Clostridia. J Bacteriol. 190:1124–1127.
Fraser C, Hanage W, Spratt B. 2007. Recombination and the nature of bacterial speciation. Science 315:476–480.
Gogarten J, Townsend J. 2005. Horizontal gene transfer, genome innovation, and evolution. Nat Rev Microbiol. 3:679–687.
Gogarten JP, Senejani AG, Zhaxybayeva O, Olendzenski L, Hilario E. 2002. INTEINS: structure, function, and evolution. Annu Rev Microbiol. 56: 263–287.
Goloboff PA, Farris JS, Nixon KC. 2008. TNT, a free program for phylogenetic analysis. Cladistics 24:774–786.
Grant W, Kamekura M, McGenity T, Ventosa A. 2001. Class III. Halobacteria class. nov. In: Boone DR, Castenholz RW, editors. Bergey's manual of systematic bacteriology volume 1: the archaea and the deeply branching and phototrophic bacteria. 2nd ed. New York: Springer-Verlag. p. 169.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52: 696–704.
Guy L, Kultima JR, Andersson SGE. 2010. genoPlotR: comparative gene and genome visualization in R. Bioinformatics 26:2334–2335.
Hanage W, Spratt B, Turner K, Fraser C. 2006. Modelling bacterial speciation. Philos Trans R Soc Lond B Biol Sci. 361:2039–2044.
Hartman AL, et al. 2010. The complete genome sequence of *Haloferax volcanii* DS2, a model archaeon. PLoS One 5:e9605.
Hasegawa M, Kishino H, Yano Ta. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22: 160–174.
Hedge PJ, Spratt BG. 1985. Resistance to [beta]-lactam antibiotics by re-modelling the active site of an *E. coli* penicillin-binding protein. Nature 318:478–480.
Hilario E, Gogarten J. 1993. Horizontal transfer of ATPase genes—the tree of life becomes a net of life. BioSystems 31:111–119.
Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. J Comput Graph Stat. 5:299–314.
Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Khomyakova M, Bükmez Ö, Thomas LK, Erb TJ, Berg IA. 2011. A methylaspartate cycle in haloarchaea. Science 331:334–337.
Koonin E, Puigbo P, Wolf Y. 2011. Comparison of phylogenetic trees and search for a central trend in the "forest of life." J Comput Biol. 18: 917–924.
Kosakovsky Pond S, Frost S, Muse S. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679.
Lan R, Reeves P. 2001. When does a clone deserve a name? A perspective on bacterial species based on population genetics. Trends Microbiol. 9: 419–424.
Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 7(Suppl 1):S4.
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 21:1095–1109.
Lawrence J, Retchless A. 2010. The myth of bacterial species and speciation. Biol Philos. 25:569–588.
Lawrence JG, Ochman H. 1998. Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci U S A. 95:9413–9417.
Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol. 25:1307–1320.
Legault B, et al. 2006. Environmental genomics of "*Haloquadratum walsbyi*" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. BMC Genomics 7:171.
Lin Y, Rajan V, Moret B. 2011. Fast and accurate phylogenetic reconstruction from high-resolution whole-genome data and a novel robustness estimator. J Comput Biol. 18:1131–1139.
Lunzer M, Miller SP, Felsheim R, Dean AM. 2005. Evolution: the biochemical architecture of an ancient adaptive landscape. Science 310: 499–501.
Lynch EA, et al. 2012. Sequencing of seven haloarchaeal genomes reveals patterns of genomic flux. PLoS One 7:e41389.
Lynch M, Conery J. 2003. The origins of genome complexity. Science 302: 1401–1404.

Makarova KS, et al. 1999. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res. 9:608–628.
Malfatti S, et al. 2009. Complete genome sequence of *Halogeometricum borinquense* type strain (PR3T). Stand Genomic Sci. 1: 150–159.
Mao F, et al. 2012. Quartet decomposition server: a platform for analyzing phylogenetic trees. BMC Bioinformatics 13:123.
Martin W. 1999. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. Bioessays 21:99–104.
Medrano-Soto A, Moreno-Hagelsieb G, Vinuesa P, Christen J, Collado-Vides J. 2004. Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. Mol Biol Evol. 21:1884–1894.
Minegishi H, et al. 2010. Further refinement of the phylogeny of the Halobacteriaceae based on the full-length RNA polymerase subunit B′ (rpoB′) gene. Int J Syst Evol Microbiol. 60: 2398–2408.
Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35:W182–W185.
Müller J, Creevey CJ, Thompson JD, Arendt D, Bork P. 2010. AQUA: automated quality improvement for multiple sequence alignments. Bioinformatics 26:263–265.
Mylvaganam S, Dennis PP. 1992. Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaebacterium *Haloarcula marismortui*. Genetics 130:399–410.
Naor A, Lapierre P, Mevarech M, Papke RT, Gophna U. 2012. Low species barriers in halophilic archaea and the formation of recombinant hybrids. Curr Biol. 22:1444–1448.
Narasingarao P, et al. 2012. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J. 6:81–93.
Nelson KE, et al. 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399:323–329.
Normand P, et al. 2007. Genome characteristics of facultatively symbiotic Frankia sp. strains reflect host range and host plant biogeography. Genome Res. 17:7–15.
Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96–98.
Ohta T. 1992. The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst. 23:263–286.
Omelchenko M, Makarova K, Wolf Y, Rogozin I, Koonin E. 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biol. 4:R55.
Oren A. 2008. Microbial life at high salt concentrations: phylogenetic and metabolic diversity. Saline Syst. 4:2.
Papke R, et al. 2007. Searching for species in haloarchaea. Proc Natl Acad Sci U S A. 104:14092–14097.
Papke R, et al. 2011. A multilocus sequence analysis approach to the phylogeny and taxonomy of the Halobacteriales. Int J Syst Evol Microbiol. 61:2984–2995.
Papke RT, Koenig JE, Rodreguez-Valera F, Doolittle WF. 2004. Frequent recombination in a saltern population of *Halorubrum*. Science 306: 1928–1929.
Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng. 14:609–614.
Pearson T, et al. 2009. Phylogeographic reconstruction of a bacterial species with high levels of lateral gene transfer. BMC Biol. 7:78.
Pérez F, Granger BE. 2007. IPython: a system for interactive scientific computing. Comput Sci Eng. 9:21–29.
Perler FB. 2002. InBase: the intein database. Nucleic Acids Res. 30: 383–384.
Pfeiffer F, et al. 2008. Evolution in the laboratory: the genome of Halobacterium salinarum strain R1 compared to that of strain NRC-1. Genomics 91:335–346.
Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res. 21:599–609.
Poptsova M, Gogarten JP. 2007. BranchClust: a phylogenetic algorithm for selecting gene families. BMC Bioinformatics 8:120.
Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.
Purdy K, et al. 2004. Isolation of haloarchaea that grow at low salinities. Environ Microbiol. 6:591–595.
Ragan M. 1992. Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol. 1:53–58.
Rambaut A, Drummond A. 2007. Tracer v1.4.1. Available from: http://beast.bio.ed.ac.uk/Tracer, last accessed January 16, 2011.
Rice P, Longden L, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16:276–277.
Roberts M, Cohan F. 1993. The effect of DNA sequence divergence on sexual isolation in bacillus. Genetics 134:401–408.
Roh SW, et al. 2010. Complete genome sequence of Halalkalicoccus jeotgali B3(T), an extremely halophilic archaeon. J Bacteriol. 192: 4528–4529.
Rosenshine I, Tchelet R, Mevarech M. 1989. The mechanism of DNA transfer in the mating system of an archaebacterium. Science 245: 1387–1389.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.
Salaün L, et al. 1998. Panmictic structure of *Helicobacter pylori* demonstrated by the comparative study of six genetic markers. FEMS Microbiol Lett. 161:231–239.
Saunders E, et al. 2010. Complete genome sequence of *Haloterrigena turkmenica* type strain (4kT). Stand Genomic Sci. 2:107–116.
Siddaramappa S, et al. 2012. A comparative genomics perspective on the genetic content of the alkaliphilic haloarchaeon *Natrialba magadii* ATCC 43099T. BMC Genomics 13:165.
Spratt BG. 1988. Hybrid penicillin-binding proteins in penicillin-resistant strains of *Neisseria gonorrhoeae*. Nature 332:173–176.
Stajich J, et al. 2002. The BioPerl toolkit: Perl modules for the life sciences. Genome Res. 12:1611–1618.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.
Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.
Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal *Escherichia coli*. Nat Rev Microbiol. 8: 207–217.
Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O. 2001. Towards a reliable objective function for multiple sequence alignments. J Mol Biol. 314:937–951.
Thompson JD, Thierry JC, Poch O. 2003. RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics 19: 1155–1161.
Thompson JR, et al. 2005. Genotypic diversity within a natural coastal bacterioplankton population. Science 307:1311–1313.
Tindall BJ, et al. 2009. Complete genome sequence of *Halomicrobium mukohataei* type strain (arg-2T). Stand Genomic Sci. 1:270–277.
Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 7: e1001284.

Tuller T, et al. 2011. Association between translation efficiency and horizontal gene transfer within microbial communities. Nucleic Acids Res. 39:4743–4755.
Vreeland R, et al. 2002. Halosimplex carlsbadense gen. nov., sp. nov., a unique halophilic archaeon, with three 16S rRNA genes, that grows only in defined medium with glycerol and acetate or pyruvate. Extremophiles 6:445–452.
Vulić M, Dionisio F, Taddei F, Radman M. 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci U S A. 94:9763–9767.
Walsh DA, Bapteste E, Kamekura M, Doolittle WF. 2004. Evolution of the RNA polymerase B′ subunit gene (rpoB′) in Halobacteriales: a complementary molecular marker to the SSU rRNA gene. Mol Biol Evol. 21: 2340–2351.
Weinreich D, Watson R, Chao L. 2005. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. Evolution 59:1165–1174.
Weinreich DM, Delaney NF, DePristo MA, Hartl DL. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312:111–114.
Welch RA, et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A. 99:17020–17024.
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.
Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM. 2007. Of clades and clans: terms for phylogenetic relationships in unrooted trees. Trends Ecol Evol. 22:114–115.
Williams D, et al. 2011. A rooted net of life. Biol Direct. 6:45.
Warren MJ, Raux E, Schubert HL, Escalante-Semerena JC. 2002. The biosynthesis of adenosylcobalamin (vitamin B12). Nat Prod Rep. 19: 390–412.
Woodson JD, Escalante-Semerena JC. 2004. CbiZ, an amidohydrolase enzyme required for salvaging the coenzyme B12 precursor cobinamide in archaea. Proc Natl Acad Sci U S A. 101:3591–3596.
Woodson JD, Peck RF, Krebs MP, Escalante-Semerena JC. 2003. The cobY gene of the archaeon *Halobacterium* sp. strain NRC-1 is required for de novo cobamide synthesis. J Bacteriol. 185:311–316.
Wright F. 1990. The "effective number of codons" used in a gene. Gene 87:23–29.
Wright S. 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Proceedings of the VI International Congress of Genetics; Ithaca, NY.
Yoon SH, et al. 2011. Parallel evolution of transcriptome architecture during genome reorganization. Genome Res. 21:1892–1904.
Zhaxybayeva O. 2009. Detection and quantitative assessment of horizontal gene transfer. Methods Mol Biol. 532:195–213.
Zhaxybayeva O, Doolittle WF. 2011. Lateral gene transfer. Curr Biol. 21: R242–R246.
Zhaxybayeva O, Doolittle WF, Papke RT, Gogarten JP. 2009. Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus*. Genome Biol Evol. 2009:325–339.
Zhaxybayeva O, Gogarten J, Charlebois R, Doolittle W, Papke R. 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. 16: 1099–1108.
Zhaxybayeva O, Gogarten JP. 2003. An improved probability mapping approach to assess genome mosaicism. BMC Genomics 4:37.
Zhaxybayeva O, et al. 2009. On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proc Natl Acad Sci U S A. 106:5865–5870.

Associate editor: Martin Embley