Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea

Shijulal Nelson-Sathi, Tal Dagan, Giddy Landan, Arnold Janssen, Mike Steel, James O. McInerney, Uwe Deppenmeier, and William F. Martin

Institute of Molecular Evolution, Institute of Genomic Microbiology, Mathematisches Institut, Heinrich Heine University, 40225 Düsseldorf, Germany; Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand; Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland; and Institute of Microbiology and Biotechnology, University of Bonn, 53115 Bonn, Germany

Edited* by W. Ford Doolittle, Dalhousie University, Halifax, NS, Canada, and approved October 25, 2012 (received for review May 29, 2012)

Archaeal halophiles (Haloarchaea) are oxygen-respiring heterotrophs that derive from methanogens—strictly anaerobic, hydrogen-dependent autotrophs. Haloarchaeal genomes are known to have acquired, via lateral gene transfer (LGT), several genes from eubacteria, but it is yet unknown how many genes the Haloarchaeae acquired in total and, more importantly, whether independent haloarchaeal lineages acquired their genes in parallel, or as a single acquisition at the origin of the group. Here we have studied 10 haloarchaeal and 1,143 reference genomes and have identified 1,089 haloarchaeal gene families that were acquired by a methanogenic recipient from eubacteria. The data suggest that these genes were acquired in the haloarchaeal common ancestor, not in parallel in independent haloarchaeal lineages, nor in the common ancestor of haloarchaeans and methanomarcinales. The 1,089 acquisitions include genes for catabolic carbon metabolism, membrane transporters, menaquinone biosynthesis, and complexes I–IV of the eubacterial respiratory chain that functions in the haloarchaeal membrane consisting of diphtyanyl isoprene ether lipids. LGT on a massive scale transformed a strictly anaerobic, chemolithoautotrophic methanogen into the heterotrophic, oxygen-respiring, and bacteriorhodopsin-photosynthetic haloarchaeal common ancestor.

Haloarchaea always branch well within the methanogens (2–4). For 414 of these families were reconstructed using maximum likelihood inference and eubacterial homologs (Fig. 1 f). Of 1,479 trees, 1,089 (73%) uncovered Haloarchaea as monophyletic and rooting within (or branching next to) eubacterial genomes—Haloarcula marismortui (22), Halobacterium salinarum (23), Halobacterium sp. (20), Halomicrobium mukohatae (24), Haloquadratum walsbyi (25), Haloarcula walsbyi (26), Halobacterium salinarum (27), Natrohalis magadii (28), Natronomonas pharaonis (29), and Haloterrigena turkmenica (30)—in the context of 65 other archaeabacterial and >1,000 eubacterial reference genomes.

Results and Discussion

We first clustered the 172,531 proteins encoded in the chromosomes of 75 archaeabacterial genomes into families using the standard Markov cluster (MCL) procedure (31) yielding 16,061 protein families. Comparison with 1,078 completely sequenced eubacterial genomes delivered 1,479 protein families that are present in at least two Haloarchaeae and contain archaeabacterial and eubacterial homologs (Fig. 1f). Gene trees for the protein families were reconstructed using maximum likelihood inference (Methods).

Of 1,479 trees, 1,089 (73%) uncovered Haloarchaeae as monophyletic and rooting within (or branching next to) eubacterial rather than archaeabacterial homologs (Fig. 1B). For 414 of these trees, no homologs at all were detected in nonarchaeal...
archaeabacteria. An additional 538 families had only very distant homologs (E values >10^{-10} or amino acid identity <30%) in some nonhalophilic archaeabacteria, together we designate these 952 cases as “acquisitions.” An additional 137 genes yielded trees in which Haloarchaea branch within eubacteria to the exclusion of readily detectable archaeabacterial homologs, we designate these genes as “replacements”; acquisitions and replacements we designate collectively as “imports” (Fig. 1 B). The 390 cases of Haloarchaea nonmonophyly included 76 trees in which one haloarchaeon branched deviantly and 105 trees in which the Haloarchaea were split into two groups of two or more species. Because LGT is common in prokaryotes (32, 33), among haloarchaeans in particular (21), these 181 gene trees could well depict secondary transfers into or from the Haloarchaea.

**Single Ancestral Acquisition.** Are the 1,089 eubacterial imports in haloarchaeal genomes the result of a single ancestral acquisition or multiple parallel acquisitions? Monophyly alone does not completely decide the issue, because it is possible that a bacterial gene could be acquired recently in one haloarchaeal lineage and then passed around to other Haloarchaea by LGT. Such a process could, in principle, also generate monophyly for imported genes in a phylogenetic tree. However, in that case, individual gene trees for imported genes would be very different from one another as opposed to the case of single acquisition, where trees for imports should be the same due to vertical inheritance from the haloarchaeal common ancestor. Moreover, trees for ancestrally acquired eubacterial imports should not only be similar to each other, they should also be similar to trees for endogenous haloarchaeal genes that are shared only with other archaeabacteria, which we call recipient genes. There are 364 haloarchaeal recipient genes that are present as single copies in all 10 Haloarchaea sampled and 109 haloarchaeal imports that are present as single copies in all 10 Haloarchaea (Fig. 2 A), providing comparable tree sets.

Comparing the distributions of phylogenetic splits observed in the 364 recipient trees and the 109 imported trees containing all 10 (condensed to 9) Haloarchaea shows that the two sets exhibit a very similar phylogenetic signal (Fig. 2 B). The six most common splits in the two sets of trees are identical and comprise 51% and 46% of the splits in the two sets, respectively. Moreover, these six splits exactly define the haloarchaeal phylogeny.
generated by 56 universally distributed archaeobacterial genes (Fig. 1A, Left). To test the statistical significance of this evidence in favor of single acquisition, we used a goodness-of-fit test to compare the distributions of topologies in the two tree sets. The null hypothesis is that the two samples of trees are drawn from the same distribution, whereas the alternative hypothesis is that the two samples differ in their distribution (Methods). The test’s P value was 0.543, meaning that the null hypothesis could not be rejected (Fig. 2B, Inset). To complement this result, we examined two additional sets of trees. One set consisted of 109 random trees, and the second consisted of the observed 109 single-copy imported genes subject to one LGT rearrangement (one random prune and graft operation) each. The latter case of one LGT rearrangement represents the slightest possible LGT-induced deviation from the null hypothesis of single acquisition in the haloarchaeal common ancestor followed by vertical evolution. Both sets were tested against the recipient trees and found to be significantly different (P values <<10^{-16}; Fig. 2B, Inset), strongly rejecting the one LGT rearrangement per gene case.

When we include the 53 multiple copy genes that are present in all 10 genomes, the one LGT rearrangement per gene is also excluded, although the significance (P values <10^{-8}; see SI Text) drops. That drop is expected, however, because horizontal transfer, not duplication, drives the expansion of gene families in prokaryotes (34), hence the inclusion of multicycopy genes preferentially includes genes for which LGT is more prevalent. We note that the goodness-of-fit test does not exclude one LGT for each gene, up to 34% of the 109 single-copy recipient trees can accept a single random prune and graft operation without the test rejecting the one LGT rearrangement case for the recipient set as a whole. However, for the 162 genes that are present in all 10 genomes, the possibility that the majority of imported genes are monophyletic because of import into one of the haloarchaeal lineages and subsequent passing around of the same gene can be excluded.

For the imports present in eight or fewer haloarchaeal genomes, excluding the (perhaps unlikely) possibility that monophyly is not due to acquisition in the haloarchaeal common ancestor but to lineage-specific acquisition and subsequent spread, is more difficult, mainly for reasons of sample size. The goodness-of-fit test based on split distributions cannot be used because few comparisons yield identical leaf sets for import vs. recipient trees. For the ≤8-species cases, we therefore developed a less direct test, comparing the sets of recipient and import trees via their phylogenetic compatibility with the recipient trees for the 10 haloarchaeal species (Methods). Here, too, the null hypothesis of common
ancestry for import and recipient genes could not be rejected in any of the ≤5-species cases, although the acquire-and-spread scenario was also not rejected for the 4-, 5-, and 6-species single and multiple copy cases (268 imports total; SI Text). Given that (i) the conventional interpretation of monophyly is presence in the common ancestor, that (ii) the 151 eight and seven species cases reject the acquire-and-spread scenario (SI Text) as an alternative explanation of monophyly, and that (iii) the data that most directly address the acquire-and-spread scenario—the 162 eubacterial imports present in all 10 genomes—most strongly reject it (Fig. 2B), the simplest interpretation of monophyly for the 1,089 imports is that their origin traces to a single acquisition in the haloarchaea common ancestor followed mainly by vertical descent and widespread differential loss, with some subsequent LGT among haloarchaea (21, 32, 33), notably for multicopy genes (34), not being excluded.

**Methodogens Are Affine for Eubacterial Genes.** As seen in Fig. 1A, not only the 10 Haloarchaea, but also the five Methanosarcinaceae (Ms), the two Methanococcales (Mc), and the five Methanomicrobiales (Mm) sampled share many genes with eubacteria, raising the question of when these imports entered these methanogen lineages. Repeating our phylogenetic analyses for these groups (Fig. 2C) reveals that merely four eubacterial imports (three predicted membrane proteins and a glycosyl transferase) can be traced to their common ancestor, and that these are present in at most 6 of the 22 descendant genomes. Whereas 124 imports can be traced to the Ms/Mc/Mm common ancestor, these imports are also sparsely distributed, with only two (COG1032, an FeS-oxidoreductase and COG1387, histidinol phosphatase) being present in all 12 descendant methanogens. This contrasts to the 1,089 haloarchaeal imports that are specific to the haloarchaea lineage, 162 of which (15%) have been retained in all 10 haloarchaeans sampled. The Ms, Mc, and Mm lineages have—like the haloarchaea—independently acquired hundreds of eubacterial genes, but the crucial observation is that they have remained strict anaerobes, and they have furthermore remained obligatory methanogenic (5–7). In stark contrast, the halophiles became aerobic heterotrophs and lost methanogenesis altogether. Collectively, the data point to a very different nature of the gene acquisition process in the halophiles and methanogens sampled here.

**Donor Lineages.** The acquisition of >1,000 genes is reminiscent of massive gene acquisitions surrounding the origin of mitochondria (35, 36) or plastids (37, 38). From what donor were these genes acquired? Because bacterial chromosomes undergo gene influx and gene export over time, it is unlikely that any one contemporary bacterial lineage would emerge as the donor of all eubacterial genes in haloarchaeal chromosomes (36, 39). All of the higher level taxa sampled appear as the sole sister group to the haloarchaeal gene or appeared in a sister group of mixed phylogenetic composition, as one might expect due to frequent LGT among bacteria (Figs. S1 and S2A). The most frequent apparent donor lineage was the actinobacteria with 131 occurrences as the sole taxon in the sister group to Haloarchaeae and 169 occurrences in the mixed sister group cases, followed by α-proteobacteria (88 sole plus 97 mixed), γ-proteobacteria (51 sole plus 111 mixed), and δ-proteobacteria (53 sole plus 100 mixed).

**Function of Imported Genes.** Trees generated from 56 recipient genes present as a single copy in all archaea bacteria place the Haloarchaeae branching from within the methanogens, but not specifically as sisters to the Methanosarcinaceae (Fig. 1). Rather, the Haloarchaeae appear to have emerged from simpler and more primitive methanogens, ones that lack both cytochromes and methanophenazine (5). Methanogens that lack cytochromes and methanophenazine are capable only of H₂-dependent methanogenesis, and have a single coupling site in their energy metabolism (5, 40). Haloarchaeae have a respiratory chain with several coupling sites (1). Methanogens are strict autotrophs and strict anaerobes (5), whereas Haloarchaeae are heterotrophs and can use O₂ as their terminal acceptor. Thus, the essential metabolic functional units for transforming a methanogen into the haloarchaeal common ancestor are (i) membrane transporters for reduced carbon compounds; (ii) a heterotrophic carbon metabolism that directs the oxidation of organic substrates to support carbon and energy metabolism; (iii) a respiratory chain for terminal oxidation and chemiosmotic ion pumping; and (iv) genes for the synthesis of any additional cofactors required, for example menaquinone, the quinone universally present in all halophiles (41). Those four essential functional units are very clearly represented within the eubacterial imports in haloarchaeal genomes.

Among the 1,089 haloarchaeal imports from eubacteria almost half (482, 44%) of the imports are related to metabolism, with amino acid transport and metabolism (114) and energy conversion (95) being the most abundant classes, followed by inorganic ion transport and metabolism (86) (Table S1; Fig. S2 B and C). Whereas methanogens without cytochromes grow on gases, which traverse membranes freely without transporters, Haloarchaeae abound in eubacterial transporters: 157 of the acquired families are annotated as permease, importer, or transporter. Although the true substrate spectrum of these transporters is yet unknown, 49 trace to amino acid or carbohydrate metabolism (Tables S1 and S2), and they operate in a membrane consisting of typical archaebacterial lipids (1).

Methanogens cannot use exogenous carbohydrates for growth (5, 42); their sugar synthetic pathways are anabolic, whereas carbon metabolism in Haloarchaeae runs in the catabolic direction. For a methanogen to become heterotrophic, it needs to acquire the enzymes underpinning the heterotrophic lifestyle from a heterotrophic donor (43). Among the eubacterial genes imported into Haloarchaeae are pyruvate kinase, glucose-6-phosphate isomerase, phosphoglyceromutase, 6-phosphogluconate dehydrogenase, the eubacterial type fructose 1,6-bisphosphatase, as well as genes for 2-keto-3-deoxy-6-phosphogluconate aldolase of the Entner–Doudoroff pathway. Eubacterial enzymes of pyruvate breakdown were also found, including two copies of pyruvate:ferredoxin oxidoreductase, and genes for pyruvate dehydrogenase complex E1 and E2 components.

Earlier studies showed that five haloarchaeal respiratory chain components are eubacterial acquisitions in two Haloarchaeae (15). Fig. 3 shows that most of the 11 subunits of NADH dehydrogenase (complex I) are present in all 10 Haloarchaeae. Complexes I–III require quinones. Haloarchaeae possess the naphthoquinone menaquinone (41) and several of the imported genes are involved in menaquinone biosynthesis, including menA. Finally, among the imported genes, 26 are annotated as transcriptional regulators and 8 are annotated as chaperones, including members of the DnaJ family.

**Conclusion**

Were these 1,000 genes accrued in the haloarchaean ancestor one by one or in a single mass acquisition? The former possibility is unlikely, because in the absence of corresponding interaction partners to form functional complexes, individual protein subunits of catalytic carbon metabolism, the respiratory chain, or cofactor biosynthesis lack selectable function, which would allow them to become fixed in a methanogenic recipient. This argues in favor of mass transfer of genes for the entire pathways and complexities over a short period of evolutionary time. The origin of Haloarchaeae was thus an evolutionary leap that transformed a methanogenic host into an oxygen-respiring heterotroph—the founder haloarchaeon. A possible context of that cellular association is anaerobic syntrophy (44, 45), that is, a H₂-producing heterotrophic bacterial donor in association with a H₂-
dependent methanogenic recipient. Anaerobic syntrophy is common in nature and has been suggested as the selective force at the origin of eukaryotes (43, 46). If similar processes underlie the origin of Haloarchaea and eukaryotes, why did Haloarchaea remain prokaryotic, whereas eukaryotes became complex? The main physiological difference between Haloarchaea and eukaryotes concerns the location of the bioenergetic membrane. In Haloarchaea it is the archaeabacterial plasma membrane (1). In eukaryotes it is the mitochondrial inner membrane—the key to eukaryote genome complexity (47). Mitochondria afforded ancient eukaryotes many orders of magnitude more energy per gene than their prokaryotic ancestors. That boost surmounted the energetic constraints imposed by reliance upon the cytoplasmic membrane as the source of chemiosmotic potential, thus allowing eukaryotic genomes to expand freely, resulting in eukaryotic cell complexity (47). The Haloarchaea have long figured into issues of early microbial evolution (48). From the standpoint of genome chimerism, they now appear to have undergone the very same physiological transformation as the eukaryotes, and the kind of gene transfer involved—from symbionts to the host chromosomes—is still ongoing in eukaryotic cells today (49). Haloarchaea remained prokaryotic because they failed to preserve a genome-containing bioenergetic organelle.

**Methods**

Data. Completely sequenced genomes of 1,153 microbial species were downloaded from the National Center for Biotechnology Information (NCBI) website (www.ncbi.nlm.nih.gov). This includes 75 archaeal genomes (version April 2010) and 1,078 eubacterial genomes (version September 2010). Taxonomic classification of the species was downloaded from the NCBI Taxonomy database (www.ncbi.nlm.nih.gov/Taxonomy/).

Clusters of Homologous Proteins. Clusters of homologous proteins were reconstructed from a total of 172,531 proteins encoded within the archaeal chromosomes. An all-against-all BLAST (50) yielded 147,071 reciprocal best BLAST hits (BBH) (51) using E value <10~10 and ≥30% amino acid identity as a threshold. Protein pairs were globally aligned using the Needleman–Wunsch algorithm with needle program (EMBOSS package) (52). A total of 137,022 protein pairs having global amino acids identities ≥30% were clustered into protein families using the MCL algorithm (31) with default parameters. This yielded a total of 16,061 archaeal protein families of ≥2 proteins. The remaining 35,509 proteins were classified as singletons. Eubacterial homologs to archaeal proteins were found using an rBBH analysis as described above, which yielded 8,451 archaeal protein families having one or more eubacterial homologs. The functional classification of protein families was based on the archaeobacterial orthologous groups database (KOG) database (53). Protein families that overlapped with KOG clusters were annotated to the same function as the matching KOG. The remaining protein families were manually classified by sequence similarity to the COG database (www.ncbi.nlm.nih.gov/COG) (32, 33), the KOG database (53) (see also Table S4).

**Fig. 3.** Eubacterial respiratory chain components in Haloarchaea. Green boxes indicate presence of the gene in the corresponding Haloarchaea genome and that the gene is more similar to eubacterial than to archaeobacterial homologs in the corresponding phylogenetic trees. Gray boxes indicate that homologs can be detected in the corresponding genome by BLAST searches, but that the clustering procedure did not include them within the 16,061 archaeal clusters. White boxes indicate that no homolog was detected. (A) Haloarchaeal nuo sequences are monophyletic but an additional paralogous copy is present in Halobutus. (B) Salinibacter has acquired a copy of ndh from Haloarchaea, which are otherwise monophyletic. (C) Haloarchaeal sdhA sequences are monophyletic but additional paralogous copies of eubacterial origin are present in several genomes (see also Table S4).

Comparison of Tree Sets. Two sets of trees were compared using MAFFT (multiple alignment using fast Fourier transform) (54), and trees were reconstructed using Phyml (55) with the best fitting model in individual trees as inferred by ProtTest (56) using the AIC measure. An archaeabacterial reference tree was reconstructed from a weighted concatenated alignment of archaeal single copy universal genes using Phyml with the LG+G+I model, which was the most frequent best fitting model, rooted using Nanaarchaeota and Kaoarchaeota as an outgroup. Trees of recipient genes were reconstructed from sequences of all 10 Haloarchaea and one nonhaloarchaeal sequence using the same procedure. For polarizing the direction of gene transfers, the root of Jain et al. (33) was used.

Reconstruction of Lateral Gene Transfer Events. Eubacterial acquisitions within halophilic archaeal genomes were identified by presence absence pattern (PAP) analysis and BLAST protein sequence similarity searches. Of the total 8,451 bacterial-like protein families in archaebacteria 1,479 had ≥2 Halarchaea species. Of these, 952 do not possess other nonhaloarchaeal homologs in the same families and correspond to unique acquisitions within Haloarchaea from eubacterial species. Archaeabacterial xenologous genes that were replaced by a eubacterial acquisition are expected to be more similar to their eubacterial ancestors than to their orthologues in other archaeabacterial species (57). Putative replaced halophilic proteins were identified by comparing the E value of their BBHs within eubacterial and archaeabacterial genomes. Proteins having a eubacterial BBH of lower E value than that of the archaeabacterial BBH were classified as putative acquisitions from eubacteria, corresponding to 527 protein families. All 1,479 protein families were aligned with their eubacterial homologs including the three best eubacterial hits per archaebacterial protein (but excluding redundant eubacterial sequences), and phylogenies were reconstructed as described above. The trees were classified into groups by the branching topology of Haloarchaea and eubacteria using an in-house PERL script. A group is considered as monophyletic for Haloarchaea if there exists a bipartition (branch) in the tree that splits between Haloarchaea and the rest. Single eubacterial sequences branching with the haloarchaeal clade, and vice versa were tested manually. In each tree, the branch connecting the monophyletic Haloarchaea clade to the eubacteria serves to split the eubacteria clade into two groups, the nearest neighbor of Haloarchaea was assigned as described in Thiagarajan et al. (36).

Phylogenetic Compatibility with a Reference Set. Two sets of trees were compared using a J2 goodness-of-fit test (58), operating on a 2χm contingency table. The m cells were defined in an adaptive procedure as follows. The two samples were pooled together into a single set of size n, and the n trees converted into splits. Each split was ranked according to its frequency in the pooled split sets. Each tree was labeled by its lowest ranking split, and the pooled tree set was sorted by this label. Cells were defined as a collection of split ranks by sequential addition of split ranks from the sorted list, and creation of a new cell when the current cell included at least √n cells, resulting in m ≤√n cells. In the last step, trees from the two sets were added to a 2χm contingency table based on their least ranked split. We have studied the adaptive cell procedure and goodness-of-fit testing in a series of permutation analyses, and the resulting J2 test proved to be an unbiased χ-level test (SI Text, Table S5, and Figs. S3 and S4).

Phylogenetic Compatibility with a Reference Set. Two sets of trees were compared using their compatibility with a reference set of trees. Each n taxon tree was decomposed into its (n-3) splits, and each split was scored by the
fraction of splits in the reference set that are phylogenetically compatible with it. The (r-2) split compatibility scores were averaged to produce a tree compatibility score. The distributions of the tree compatibility scores for the two sets of trees was compared using the Kolmogorov-Smirnov test (58) (SI Text).

ACKNOWLEDGMENTS. We thank Martin Embley and Dan Graur for critical comments on an earlier version of the manuscript. We thank the central computing resources of the University of Düsseldorf for technical support. G.L. acknowledges financial support from the University of Düsseldorf rectorate, W.F.M. and T.D. are funded by the European Research Council and the German Ministry of Science and Education, J.O.M. is funded by the Science Foundation of Ireland, M.S. is funded by the Allan Wilson Centre and the Alexander von Humboldt Foundation, and A.J. and U.D. are funded by the German Research Foundation.