Modularized Evolution in Archaeal Methanogens Phylogenetic Forest

Jun Li[1,†], Chi-Fat Wong[1,†], Mabel Ting Wong[1,4], He Huang[2], and Frederick C. Leung[1,3,*]

[1]School of Biological Sciences, Faculty of Science, The University of Hong Kong, China

[2]Center for Marine Environmental Studies, Ehime University, Japan

[3]Bioinformatics Center, Nanjing Agricultural University, People's Republic of China

[4]Present address: Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON, Canada

*Corresponding author: E-mail: fcleung@hkucc.hku.hk.

[†]These authors contributed equally to this work.

Accepted: November 17, 2014

Abstract

Methanogens are methane-producing archaea that plays a key role in the global carbon cycle. To date, the evolutionary history of methanogens and closely related nonmethanogen species remains unresolved among studies conducted upon different genetic markers, attributing to horizontal gene transfers (HGTs). With an effort to decipher both congruent and conflicting evolutionary events, reconstruction of coevolved gene clusters and hierarchical structure in the archaeal methanogen phylogenetic forest, comprehensive evolution, and network analyses were performed upon 3,694 gene families from 41 methanogens and 33 closely related archaea. Our results show that 1) greater than 50% of genes are in topological dissonance with others; 2) the prevalent interorder HGTs, even for core genes, in methanogen genomes led to their scrambled phylogenetic relationships; 3) most methanogenesis-related genes have experienced at least one HGT; 4) greater than 20% of the genes in methanogen genomes were transferred horizontally from other archaea, with genes involved in cell-wall synthesis and defense system having been transferred most frequently; 5) the coevolution network contains seven statistically robust modules, wherein the central module has the highest average node strength and comprises a majority of the core genes; 6) different coevolutionary module genes boomed in different time and evolutionary lineage, constructing diversified pan-genome structures; 7) the modularized evolution is also closely related to the vertical evolution signals and the HGT rate of the genes. Overall, this study presented a modularized phylogenetic forest that describes a combination of complicated vertical and nonvertical evolutionary processes for methanogenic archaeal species.

Key words: network analysis, co-evolution network, HGT, LGT, phylogenetic forest, modularized evolution.

Introduction

The biosynthesis of methane is a ubiquitous, defining characteristic of methanogens (Ferry 1994). Via the process of methanogenesis, these methanogens play key roles in carbon cycle by producing 900 million tons of methane annually, contributing to 16% of total emission of global warming gases (Schlesinger 1997; Elizabeth and Dina 2006; Hedderich and Whitman 2006). In terms of ecology, these strict anaerobes find residence in sediment, wetland, rice paddy, as well as anthropogenic sites such as anaerobic digesters and biogas plants (Liu and Whitman 2008). Phylogenetically speaking, methanogens are classified under six taxonomic orders in the *Archaea* domain: *Methanopyrales*, *Methanococcales*, *Methanocellales*, *Methanobacteriales*, *Methanomicrobiales*, and *Methanosarcinales* (Garcia et al. 2000; Liu and Whitman 2008; Sakai et al. 2008). These methanogens are distinguished from other archeons by the possession of a unique complex biochemistry for methane synthesis as part of their energy metabolism, forming a nonmonophyletic cluster juxtaposed with nonmethanogenic taxa (Deppenmeier 2002). Herewith, the interesting taxonomic position of methanogens intrigued both ecologists and phylogenists regarding its evolutionary history (Reeve et al. 1997; Bapteste et al. 2005; Luo et al. 2009).

With the use of different genetic markers, contradicting phylogenetic relationships between methanogens and closely

related nonmethanogenic species are yielded, which indicate pervasive events of horizontal gene transfers (HGTs). One unresolved classification resides in the ambiguous relationship among *Methanomicrobiales*, *Methanosarcinales*, *Methanocellales*, and a closely related nonmethanogenic order *Halobacteriales*, which remains debated to date (Boone et al. 1993; Garcia et al. 2000; Bapteste et al. 2005; Brochier et al. 2005; Wright 2006; Yarza et al. 2008; Kelly et al. 2011; Nelson-Sathi et al. 2012). The disputable phylogenetic position of *Methanopyrales* serves as another example—this order was considered to be distantly related to all other methanogen lineages (Burggraf et al. 1991; Rivera and Lake 1996; Brochier et al. 2004), whereas other reports proposed a close phylogenetic linkage between *Methanopyrales*, *Methanococcales*, and *Methanobacteriales* (Nolling et al. 1996; Slesarev et al. 2002; Brochier et al. 2004; Bapteste et al. 2005; Gao and Gupta 2007; Luo et al. 2009). The major cause of the incongruence in methanogen genealogy is the rampant episodes of HGT in prokaryotic evolution, distributing genes across prokaryotes with distant genetic relationship (Dagan and Martin 2006; Fraser et al. 2007; Brilli et al. 2008; Dagan et al. 2008; Boucher and Bapteste 2009; Norman et al. 2009; Schliep et al. 2011; Treangen and Rocha 2011). Such cellular mechanisms created species-independent evolutionary modules, whereby a majority of the genes possess distinct evolutionary history, and only around 1% (or less) of the gene trees share an identical topology with the species tree (Dagan and Martin 2006; Bapteste et al. 2008). As a result, the clarification of phylogenetic relationship in methanogens requires a robust classification based on the complexity of the whole phylogenetic forest, composed of trees from all orthologous gene families.

With the development of comparative genomics in studying multiple stains within single species, the concept of "pan-genome" has been proposed to allow an inclusion of both "core genome" (genes shared by all strains) and "dispensable genome" (genes shared by specific strains) (Medini et al. 2005). The concepts of core-genome and pan-genome have been extended to metagenomes and higher taxonomic level (Lawrence and Hendrickson 2005; Segata and Huttenhower 2011; Droge and McHardy 2012). According to the complexity hypothesis, informational genes (involved in transcription, translation, and related critical pathway) are usually necessary and less prone to be transferred horizontally. Meanwhile, operational genes involved in housekeeping, are commonly horizontally transferred (Jain et al. 1999). To allow an inclusion of these new concepts, various approaches combining both vertical and horizontal evolution analyses have been developed recently (Huson and Bryant 2006; Leigh et al. 2008; Schliep et al. 2011). Among these methods, phylogenetic networks have promising trends in describing both vertical and nonvertical evolution in given gene sets, although the results may not be interpreted easily (Huson and Bryant 2006). In service of this, Leigh et al. (2008) proposed a hierarchical clustering method which identifies the congruently evolved gene families and reconstructs the super-matrix tree for each cluster. Concurrently, Schliep et al. (2011) adopted clanistics analysis on prokaryotic phylogenetic forest to reveal the prevalent incongruence among the phylogenetic trees, potential HGTs, and evolutionary pattern related functional traits. Over the past decades, the network analysis approach has been widely applied in biological evolution and ecology studies (Proulx et al. 2005). With the rapid expansion in biological data, especially in the fields of genomics, transcriptomics, proteomics, and other "omics," it is anticipated that biological network analyses would become relied upon heavily.

Previous studies have illustrated that clustering genes based on their topology incongruence levels is very useful in identifying modules with identical or similar evolutionary history, and in filtering vertically inherited genes (coherently evolved core genes) before the construction of phylogenomic trees (Susko et al. 2006; Puigbo et al. 2009, 2012; Leigh et al. 2011). Albeit likelihood-based clustering methods are more statistically robust than topology-distance based methods in principle, the analyses on real data set often yield an excessive number of clusters (Leigh et al. 2008; Leigh et al. 2011). A notable example would be the formation of over ten clusters in 102 nearly universal trees in the work of Leigh et al. (2008, 2011). This kind of decomposition based on likelihood method in the phylogenetic forest is indeed helpful in studying the intricate phylogenetic incongruence or system and random errors in phylogeny reconstruction. Nonetheless, blurred cluster methods, such as fuzzy clustering, are more useful discovering vertical and nonvertical evolutionary processes and revealing major modules of genes sharing similar evolutionary history. For this purpose, a pioneering work using boot-split distance-based classical multidimensional scaling analysis was developed by Puigbo et al. (2009). Overall, the development of distance-based clustering methods for coevolved genes recognition remains an incomplete avenue. To derive an accurate inference of methanogen phylogeny, it is necessary to employ in-depth evolutionary analyses with sophisticated approaches that prevent these limitations and plight.

In this study, comprehensive evolutionary and network analyses were performed in an effort to reveal congruent and conflicting evolutionary signals, coevolved gene clusters, and global and local features in the archaeal methanogen phylogenetic forest. Although both congruent and incongruent evolutionary signals between genes were revealed in the incongruence tests analysis of the whole phylogenetic forest, the clanistics analysis discovered a scrambled evolutionary relationship among methanogens and certain methanogenesis-adaptive genes. Furthermore, a comprehensive network analysis based on topology similarity was implemented, in order to elucidate the hierarchical structure and global and local features in the archaeal methanogenic phylogenetic forest. In addition, the combined analyses of HGT,

origin of the gene families and simulated evolution of genes families, revealed ubiquitous vertical inheritance evolution, as well as the variable phylogenetic depth (origins time) of gene families and variable HGT rate. Altogether, the results contribute to the modularized phylogenetic forest for methanogens and related species.

Materials and Methods

Acquisition of Genomic Sequences, Taxonomy, and Lifestyle Information

Seventy-four archaeal genomes, encompassing 41 methanogen genomes and 33 closely related archaeal genomes, were downloaded from NCBI Genbank database (November 2012). Twenty-seven additional archaeal and bacterial genomes were included as outgroup species to evaluate the influence of root position on the phylogenetic relationships among methanogens (see supplementary tables S1 and S2, Supplementary Material online, for more details). Taxonomic information of 101 prokaryotic genomes was retrieved from NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/, last accessed November 2012), whereas the corresponding phenotypic information was obtained from NCBI Genome Information (http://www.ncbi.nlm.nih.gov/genome/browse/, last accessed November 2012). Further on, genetic information related to the methanogenic pathway was gathered from the *Bergey's Manual of Systematic Bacteriology* (Boone et al. 2001). For species that were not listed in the *Bergey's Manual of Systematic Bacteriology*, their information was extracted manually from respective literature.

Function and Pathway Annotation for Genes and Ortholog Gene Families

The functional category of COG (Tatusov et al. 1997) for each protein was assigned using NCBI RPS-BLAST at 1e-5 cutoff. For each ortholog gene family, the COG category was defined by majority of the functional category among the member proteins. Based on this definition, functional categories of most (>80%) ortholog clusters are consistent with at least 90% of their family members. The KEGG (Tatusov et al. 1997) pathway annotation for each family was achieved using similar procedure described above.

Construction of Ortholog Families Using 101 Prokaryotic Genomes

All-versus-all NCBI-BLAST (BLASTp) (Altschul et al. 1990) for 254,439 proteins from the 101 genomes was performed with parameters "-e 1e-6 -m8 -F T." To reduce the number of clusters containing partially homologous proteins, protein sequences shorter than 50% of the longer protein sequence were removed from the pair-wise alignments.

Markov clustering (MCL) method with OrthoMCL (Li et al. 2003) was used to identify ortholog gene families (clusters) based on the all-versus-all BLASTp results. In the MCL algorithm, inflation parameter I (an indicator of the tightness of gene families) has a critical influence on the final clustering result. For example, a smaller inflation would produce less number of clusters with larger average cluster size. Considering the consistency of enzyme commission numbers in each ortholog group, an inflation parameter of range 1.5–2.5 was described previously as the best fit for empirical data sets (Li et al. 2003). Meanwhile, inflation parameters 1.5, 2, and 4 have extensively studied previously (Li et al. 2003; Chen et al. 2007; Salichos and Rokas 2011). To determine best suitable inflation parameter I for our study, I of 1.5, 2, and 4 were all evaluated, and $I=2$ was chosen for further analysis due to its balanced performance. The influence of inflation parameter on the consistency of protein function in each family was described in supplementary materials, Supplementary Material online. Via these procedures, 27,858 ortholog clusters were defined, with 9,650 clusters containing at least four proteins in the 101 prokaryotic genomes. Several rounds of filtering were further carried out to ensure that only full-length orthologs were clustered, and no outparalogs retained in the ortholog clusters (see supplementary materials, Supplementary Material online, for more details). Finally, 3,694 ortholog clusters with minimum alignment length of 100 amino acids and at least four taxa in 74 methanogen-related species were achieved.

Thirteen genes were defined as prokaryotic core genes, which present one and only one copy in all 101 prokaryotic species. Similarly, 92 genes were defined as archaeal core genes. Lastly, the archaeal near core genes were defined as the genes represented in at least half (on average 81%) of the methanogen.

Reconstruction of Phylogenetic Tree and Ultrametric Tree

Protein sequences in each clusters were first aligned independently using MUSCLE v3.7 (Edgar 2004) with the parameter maxiters = 30. The nucleotide alignments were constructed based on the corresponding protein alignments. Gblocks v0.91 (Castresana 2000) was adopted to refine the original alignments with parameters "b1 = half number of the sequences in each cluster, b2 = b1, b3 = 10, b4 = 6, and b5 = n." And this combination of parameters enabled us to filter both ambiguously aligned and saturated substation sites, hence generating credible alignments.

Seventy-four species in *Euryarchaeota* were chosen to show the evolutionary relationship between methanogenic species and their related neighbors. Prottest (Abascal et al. 2005) was adopted to evaluate the substitution models. The result showed that "Whelan and Goldman (WAG) + Gamma + I" provided the best fit for 51.4% of the genes, thus this model was applied in downstream analysis.

The maximum likelihood (ML) and Bayesian trees for each gene family and concatenated core genes were reconstructed using RAxML (Stamatakis et al. 2005) and Phylobayes (Stamatakis et al. 2005), respectively. Root positions in the trees in figure 1 and supplementary figure S1, Supplementary Material online, was determined by the four species in archaeal phylum *Crenarchaeota*.

To detect HGT events and estimate the phylogenetic depth, a reference ultrametric species tree was constructed upon Archaea core genes using the function of "chronopal" in ape (Paradis et al. 2004) with parameter lambda 500 (an empirical optimum value based on Sanderson [2002]). To be specific, the distances between root (common ancestor of all species used) and the tips (current species) were all adjusted to 1 in this ultrametric tree. Using AnGST (David and Alm 2011), the common ancestor (birth time), evolutionary process and the relative time of origin (birth process) were achieved simultaneously for each gene family.

Incongruence Tests

Approximately unbiased (AU) tests (Shimodaira 2002) were performed using Consel (Shimodaira and Hasegawa 2001) to test the incongruence level between trees. For the incongruence test among species trees (phylogenomic trees), the likelihood values of the phylogenetic trees in concatenated 92 archaeal core genes, as well as their individual genes, based on different topologies were calculated using RAxML. The statistical significance of difference among trees was further evaluated using Consel. In the whole phylogenetic forest of 3,694 gene trees, when two genes contain different taxa, the congruent taxa and sequences were used for further analysis. As to the taxa limited AU-tests (lower triangle matrix in fig. 1*b*), the number of taxa for each comparison was limited to at most 20. When the number of total congruent taxa between the two genes exceeded 20, 20 common taxa were selected randomly.

Seq-gen (Rambaut and Grassly 1997) was applied to generate sequences with totally congruent gene trees (shown in supplementary fig. S2, Supplementary Material online). Tree IV in figure 1 was adopted as the reference species tree for each gene family with WAG + Γ4 (gamma value was retrieved from the original RAxML gene tree) substitution model. The sequence length and taxa composition in each simulated family were adjusted to be the same as those in the original gene families.

Detection of HGT and Phylogenetic Depth

AnGST (David and Alm 2011) was used to deduce potential horizontally transferred genes. The reference species tree was reconstructed based on the aforementioned species tree III

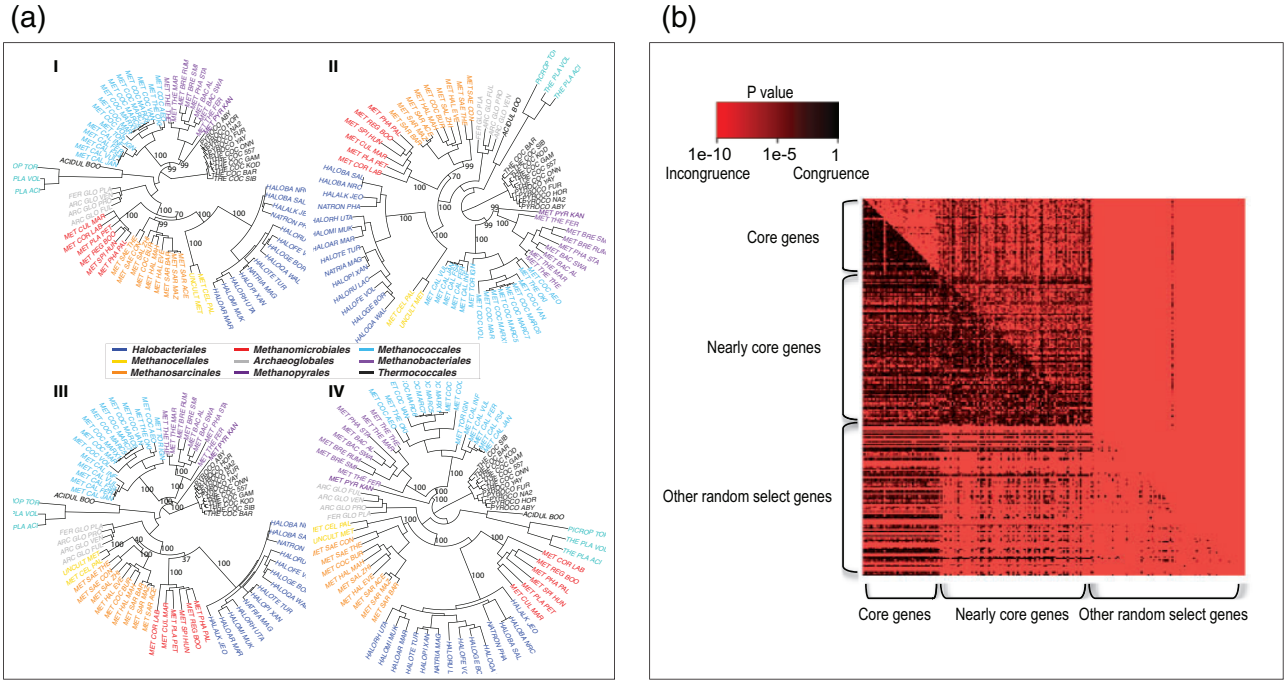

Fig. 1.—Phylogenomic trees and incongruence AU tests among genes. (*a*) phylogenomic trees constructed based on I) 13 prokaryotic core genes using ML method; II) 13 prokaryotic core genes using Bayesian method; III) 92 archaeal core genes using ML method and; IV) 92 archaeal core genes using Bayesian method. (*b*) Incongruence AU tests among genes among 250 randomly selected genes. The upper triangular matrix refers to AU tests using original data. The lower triangular matrix refers to AU tests with at most 20 taxa. Core genes refer to 50 randomly selected archaeal core genes. Nearly core genes refer to 100 randomly selected genes, which contain at least half (on average 81%) of the methanogen genomes.

shown in figure 1*a*. Assigning a proper penalty score for HGT detection is an uneasy task because the gene families are highly variable using different pipelines retrieving them (David and Alm 2011). Here, a broad range of event penalties (HGT: 3–8, duplication: 2–5, and loss: 1–3) on 92 core gene families were tested to suit the following criteria: 1) the reconciled root positions for the majority of the core families should be right (according to the rooted ultrametric species tree); 2) the total number of deduced events (HGT, duplication, and loss) should be minimized. The chosen penalty score H7-D4-L2 were very similar to the optimized penalty score (H3-D2-L1) in Mowgli (Doyon et al. 2010) and AnGST (David and Alm 2011), and the scores functioned in minimizing the genome size variation in related genomes (Salichos and Rokas 2011). Unlike common HGT detection methods, in which only the uncertainty in gene trees is considered, uncertainties in species tree were also considered in this study. Potential HGT events were identified by the ultrametric species tree with bootstrap = 100 for gene trees and subsequently shortlisted by the requirement of at least 50% support by the bootstrap reference species trees. Based on this method, 26,013 potential HGTs (without bootstrap species tree support) were detected and 4,466 confirmed HGTs (supported by over half of the bootstrap reference trees) were retained for downstream analyses.

Network Analysis

Distances between different trees were defined by RPHD, a modified version of Penny and Hendy's topology distance, which evaluates the number of internal branches with different bipartitions between trees (Penny and Hendy 1985). The procedures to calculate RPHD were as follow: 1) Penny and Hendy's topology distance d_{ij} was first calculated for each pairwised gene trees i and j using ape package in R. If two trees contain different taxa composition, only the subtree containing the congruent taxa would be used for d_{ij} calculation; 2) a topology distance d_{ij0} between two random trees with the same size as tree i and j, respectively, was calculated by Penny and Hendy's topology distance method; 3) the normalized tree distance was calculated as $D'_{ijn} = \min(1, d_{ij}/d_{ij0})$, ranging from 0 to 1; 4) the final topology distance RPHD was defined as $D_{ij} = \frac{\sum_{n=0}^{100} D'_{ijn}}{100}$, where D'_{ijn} is the normalized tree distance from step (3) for the nth bootstrap trees in the tested genes. The similarity between two gene trees was defined as $S_{ij} = 1 - D_{ij}$, and this measurement was used in further coevolution network analysis. Each node in coevolutionary network represented one gene tree and the edge denotes the topology similarity S_{ij} defined above. The edge strength in this weighted network ranges from 0 to 1 according to the definition of RPHD and is suitable for qualitative description of the topology similarity/congruence between two gene trees. It should be noted that the branch length information was neglected in this modified Penny and Hendy distance, because the topology congruence level was positioned as the primary subject in this study.

Because random trees could have nonzero RPHD, filtering out the trivial similarity among topologically unrelated genes (unnecessary edges in the network) would increase the signal-to-noise ratio and reduce the computational burden in downstream analyses. We therefore evaluated the distribution of topology similarity using 500,000 random trees with the same size distribution observed in the real data. The result shows that less than 0.5% of pairwised topology similarities in random trees are over 0.1. So the connections (edges) with topology similarity less than 0.1 (minimum edge weight = 0.1) in the phylogenetic network for real data were removed. This filtering process would ensure the final edges in the phylogenetic forest reflect nonrandom connections.

To reveal potential causes for the heavy tail pattern in node strength distribution, we created five simulation tree sets from the original 3,694 gene trees. In each set of simulated trees, HGT events were randomly introduced by grafting branches from different phylogenetic positions using an in-house perl script. The frequency of HGT ranges from 0 (fully congruent) to infinite (fully incongruent) by adjusting the times of grafting.

The network modules were identified using hierarchical clustering and dynamic branch cut methods (Langfelder et al. 2008) based on topological overlap matrix (TOM), which is intrinsically a smoothed out matrix from original adjacency matrix (Dong and Horvath 2007; Yip and Horvath 2007; Langfelder and Horvath 2008). Compared with the traditional tree cutting method with constant height, this clustering process with dynamic branch cut provide more flexible, automatic, and accurate clusters results, especially for nested clusters and outliers (Langfelder et al. 2008). After the first round of clustering, sporadic genes (with constitutive length <5) embedded in single clusters were further merged with the clusters they reside in. This merging procedure was critical because without it, certain statistically weak and fragmentary tiny modules would retain due to random error and other intrinsic defect in hierarchical clustering. Function of "walktrap.community" in igraph package (http://igraph.sourceforge.net/) in R was adopted to implement the community identification based on random walks. Function of "modularity" in igraph was used to evaluate the community structure in the network.

Tnet (Opsahl 2009), igraph and in-house Perl and R scripts were used for the weighted network analysis. Best fit degree distribution (ML estimation) was implemented using MASS package (Venables and Ripley 2002), and in-house R scripts. Measure of closeness was calculated using function "closeness_w" in tnet with "alpha = 0.5 and gconly = FALSE" (in the whole network even if the components are not connected). Measure of betweenness was calculated

using function "betweenness" in igraph with normalized method.

Clanistics Analysis

Clanistics analysis was performed using Phangorn package (Schliep 2011) in R (http://www.r-project.org). Ten taxonomy (*Thermococcales*, *Methanococcales*, *Methanobacteriales*, *Thermoplasmatales*, *Archaeoglobales*, *Methanosarcinales*, *Halobacteriales*, *Methanomicrobiales, Methanococcales*, and *Methanopyrales*), eight phenotypic/ecological trait categories (anaerobic, aerobic, mesophilic, thermophilic, hyperthermophilic, hydrogenotrophic, acetoclastic, and methylotrophic), and ten additional assumed taxonomic categories were used to delineate the evolutionary patterns in the phylogenetic forest. Specifically, these ten extra taxonomic categories were used to test whether two or more taxonomic categories (at order level) belong to larger monophyletic groups. The groups encompass 1) taxonomic class *Methanomicrobia*; 2) assumed category of *Methanosarcinales* and *Methanomicrobiales* as a monophyletic group; 3) assumed category of *Halobacteriales* and *Methanomicrobiales* as a monophyletic group; 4) assumed category of *Halobacteriales* and *Methanosarcinales* as a monophyletic group; 5) assumed category of *Halobacteriales*, *Methanomicrobiales* and *Methanosarcinales* as a monophyletic group; 6) assumed category of *Methanomicrobiales* and *Methanosarcinales* as a monophyletic group; 7) assumed monophyletic class I methanogen, including *Methanococcales*, *Methanobacteriales*, and *Methanopyrales*; 8) assumed monophyletic class I methanogen, including *Methanococcales* and *Methanobacteriales*, assumed category of *Methanococcales*, *Methanobacteriales*, *Methanopyrales*, and *Thermococcales* as a monophyletic group; 9) assumed category of *Methanococcales*, *Methanobacteriales*, *Methanopyrales*, *Thermococcales*, and *Thermoplasmatales* as a monophyletic group; 10) assumed category of *Methanococcales*, *Methanobacteriales*, *Methanopyrales*, *Thermococcales*, *Thermoplasmatales*, and *Archaeoglobales* as a monophyletic group.

When dissecting the unrooted phylogenetic tree into clans and slices according to native and intruder categories, some recurring patterns can be observed (shown in schematic supplementary fig. S3, Supplementary Material online) (Schliep et al. 2011). Pattern A1 (supplementary fig. S3*a*, Supplementary Material online) reflects that only native OTUs can be observed in the tree. In contrast, pattern A2 (supplementary fig. S3*b*, Supplementary Material online) reflects that a tree contains only intruder OTUs. Pattern B1 (supplementary fig. S3*c*, Supplementary Material online) and B2 (supplementary fig. S3*d*, Supplementary Material online) reflect that a tree can be separated into a slice with single intruder or single native OTU in it, respectively. Pattern C (supplementary fig. S3*e*, Supplementary Material online) reflects two perfect clans, separating the intruder and native OTUs. Pattern D1 (supplementary fig. S3*f*, Supplementary Material online) and D2 (supplementary fig. S3*g*, Supplementary Material online) reflect that a tree contains a perfect slice with native or intruder OTUs in it, respectively. Pattern E reflect a tree with a mixture of native and intruder OTUs. These patterns have different biological meanings between phenotype and taxonomy categories. For taxonomy categories, a monophyletic group (clade) could be defined on perfect clan (Pattern A1 or C), if further information confirms that the root is outside of the target clan. Pattern B (B1 and B2) indicate single potential nonvertical event happened in one-specific lineage (lineage-specific nonvertical event) in the evolution history. In contrast, pattern D (D1 and D2) indicates at least one nonvertical event (sometimes ancestral event) in certain closely related lineages happened in the evolution process. Furthermore, frequent and independent nonvertical events could cause pattern E in a tree. It should be noted that although nonvertical events usually provide a most parsimonious explanation for these coherent patterns for pattern B, D, and E, complicated gene duplication and gene loss could also lead to the same pattern. For phenotypic or ecological categories (such as anaerobic as native and aerobic as intruder), pattern A1, B1, C, and D1 are usually regarded as coherent patterns and pattern E is mixed pattern (Schliep et al. 2011). These phenotypic coherent patterns are extremely useful to discover adaptive genes to specific environment.

Results and Discussion

Phylogenomic Trees

To resolve the phylogeny between methanogens and closely related organisms, species/phylogenomic trees based on 13 prokaryotic core genes (one and only one copy in 101 prokaryotic genomes) and 92 archaeal core genes (one and only one copy in 74 archaeal genomes) were reconstructed using ML and Bayesian methods. Previous results revealed inconsistent interrelationship among *Methanomicrobiales*, *Methanosarcinales*, *Methanocellales*, and *Halobacteriales* (Bapteste et al. 2005; Brochier et al. 2005; Yarza et al. 2008; Kelly et al. 2011). Such intricate relationships have also been shown in our results. First, the position of *Methanocellales* is uncertain in phylogenies using different markers. *Methanocellales* is closer to *Halobacteriales* in tree I and II in figure 1*a* with high confidence level (bootstrap or Bayesian posterior probability >90%); but in tree III and IV, it is closer to *Methanosarcinales*, again with high confidence level. Second, *Methanomicrobiales* is closer to *Methanosarcinales* than to *Halobacteriales* in tree II with high confidence level, whereas tree I, III, and IV support that *Methanomicrobiales* is closer to *Halobacteriales*. Therefore, the taxonomic class *Methanomicrobia* (containing orders *Methanomicrobiales*, *Methanosarcinales*, and *Methanocellales*) is not consistently monophyletic due to the ambiguous positions of

Methanocellales and *Methanomicrobiales* among different species trees. Although the statistically best species trees III and IV (see the following part for the criteria) show that *Methanocellales* is closer to *Methanosarcinales*; and *Methanomicrobiales* is closer to *Halobacteriales*, the hypothesis is not supported with high confidence level (with bootstraps values below 50% in tree III). This suggests that different regions in 92 archaeal core genes might not have exactly the same evolutionary history in these lineages. Besides these, the four phylogenomic trees also show that *Methanopyrales* is either closer to *Methanobacteriales* based on prokaryotic cores genes (in tree I and II) or closer to the common ancestor of *Methanococcales* and *Methanobacteriales* (in tree III and IV). This reveals that although the accurate phylogenetic position of *Methanopyrales* is still vague in prokaryotic or archaeal core gene trees, the monophyletic group composed of *Methanococcales*, *Methanobacteriales*, and *Methanopyrales* is consistent in all phylogenomic trees.

In addition, statistical tests with AU test (Shimodaira 2002) were used to compare topological consistency in trees in figure 1*a* to identify the best tree. Five additional trees were also constructed for comparisons, and they are 1) three trees of 16S rRNA using neighbor joining (NJ), ML, and Bayesian methods; and 2) two trees using NJ method on above two core gene sets (See supplementary fig. S1, Supplementary Material online, for the full description of the nine trees). First, the likelihood was recalculated on archaeal core gene set (92 genes) with nine alternative topologies in figure 1*a* and supplementary figure S1, Supplementary Material online. The result shows that ML tree (tree III in fig. 1*a* or supplementary fig. S1*e*, Supplementary Material online) and Bayesian trees (tree IV in fig. 1*a* or supplementary fig. S1*f*, Supplementary Material online) based on archaeal core gene set, 1) have the highest likelihood; 2) are statistically no different (AU-test, Bonferroni *P*-value > 0.01) with each other; 3) fit the data significantly better than other seven topologies (AU-test, Bonferroni *P*-value $<$ 1e-10); and 4) over 91% of the individual genes trees of the archaeal core genes is congruent (AU-test, Bonferroni *P*-value > 0.01) with tree III and IV. The same tests were performed on prokaryotic core gene sets (13 genes) and results show that the topology in tree I–IV (in fig. 1*a*) are statistically no different (AU-test, Bonferroni *P*-value > 0.01) and significantly better than five other alternative topologies (AU test, Bonferroni *P*-value $<$ 1e-10). To check whether the empirical substitution models for amino acid have significant influence on the topology estimation in the phylogenomic tree, substitution model JTT and LG were also implemented in RAxML and Phylobayes. These trees show concordant topologies with original trees using the WAG model (Bonferroni adjusted *P*-value > 0.01 in AU test). Lastly, nine supertrees using Clann (Creevey and McInerney 2005) based on both prokaryotic, archaeal core gene sets or multiple copy universal genes (species-specific duplicated genes) using Bayesian method. Various substitution models (WAG, LG, or JTT) were evaluated with the same procedure. The topologies in these supertrees did not fit the data better than tree III or IV (AU test, Bonferroni *P*-value $<$ 1e-10). We also applied the AU-test on the reference topologies from 92 archaeal core genes to the whole phylogenetic forest (3,694 trees). The result shows that found that tree III and IV (in fig. 1*a*) are still best fit trees.

To evaluate the influence of an arbitrarily chosen outgroup (*Crenarchaeota*) to the root position of the phylogenomic trees, we adopted other distantly related archaeal (e.g., *Nanoarchaeota*) or bacterial genomes to root these phylogenies. The result shows that over 95% of the interorder relationships among six orders of methanogens, and the closely related *Halobacteriales* are still the same for the same data set using identical reconstruction methods.

Incongruence among Gene Trees

To decipher the consistent and conflicting evolutionary signals in the phylogenetic forest, incongruence levels among 250 randomly selected genes were measured via the AU test (Shimodaira 2002). The result is illustrated using heatmap depictions (the upper triangular matrix in fig. 1). The results show a surprising pattern in that 1) 55% (51 out of 92) of the archaeal core genes are in topological discordance with either archaeal core genes or other genes; 2) nearly core genes (represented in at least half of the methanogen genomes) show higher (56%) topology congruency with all genes than with archaeal core genes; and 3) other genes show very distinct evolutionary processes and few of them have congruent topologies.

Our further regression analysis revealed that the AU test *P*-value has a negative correlation with taxa number (supplementary fig. S2*a*, Supplementary Material online) for methanogen ortholog families, suggesting that taxa number has a nonnegligible impact on the congruence results with AU tests. Additional correlation analysis based simulation data (see Materials and Methods for more details), which have totally congruent topology and same taxonomy composition as real data. This confirmed that with the increase of taxa number, the proportion of false rejection of congruent topology would get increased (supplementary fig. S2*b*, Supplementary Material online). To correct this family size effect, the number of taxa for each comparison was limited to at most 20 (taxa were selected randomly when total congruent taxa between the two genes exceeded 20). The new incongruence test result shows in lower triangular matrix of fig. 2 that 1) the core genes have more consistent topologies than nearly core genes and other genes, indicating strong vertical inherited signals within the core gene trees; 2) the nearly core genes show moderately congruent topologies with all genes, suggesting that their vertical inherited signal were disrupted by episodes of HGT, duplication/loss, or other nonvertical

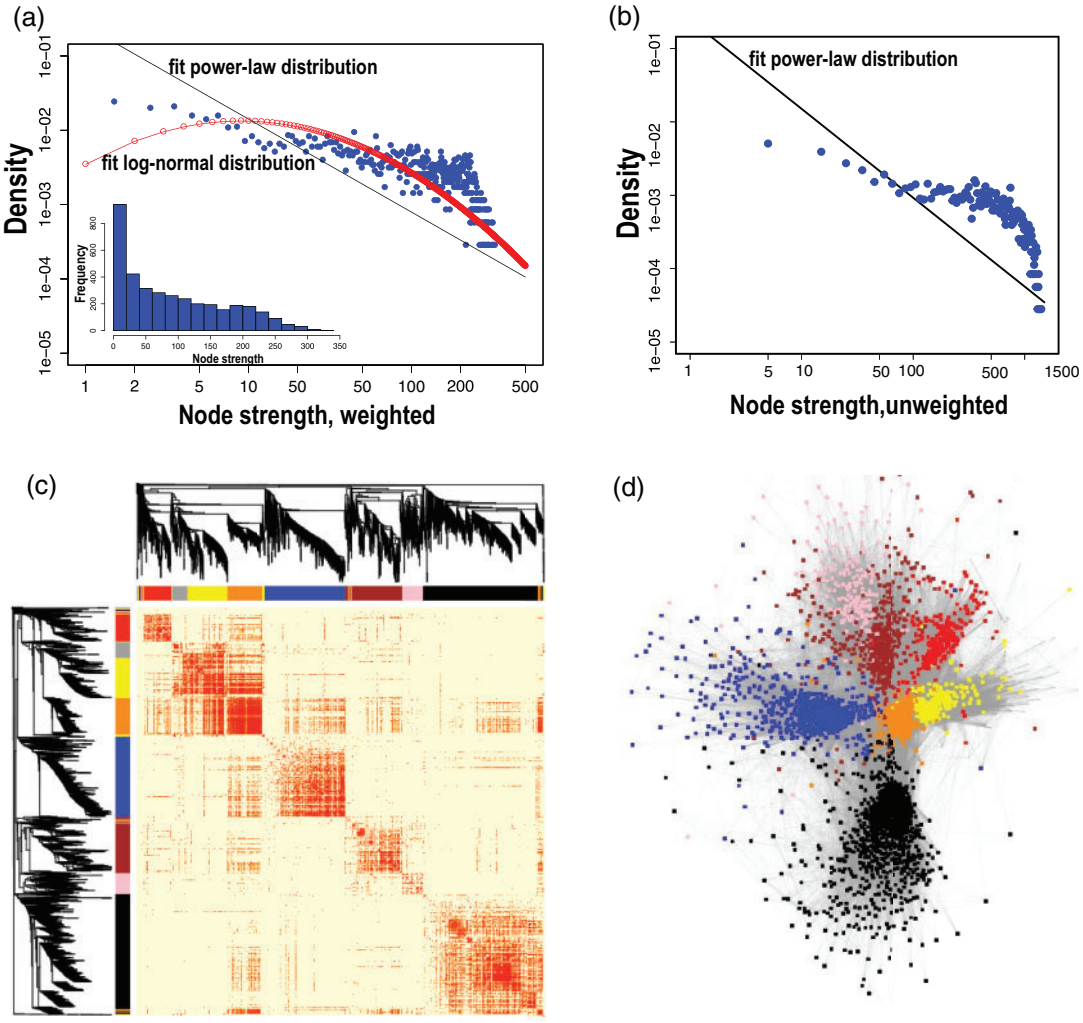

Fig. 2.—Summary of properties in coevolutionary network. (*a*) Distribution of weighted node degree (strength). The straight line denotes the fit power-law distribution and the red circles denote the fit log-normal distribution. (*b*) Distribution of unweighted node degree. (*c*) Module definition using hierarchical clustering in the coevolution network. Seven major modules were defined with colors: dark orange, yellow, blue, black, red, brown, and pink. Gray color denotes the unclassified genes. (*d*) Visualization of the network based on force-directed layout. All nodes are colored according to the module classification in figure 2*c*.

evolutionary processes; 3) most genes (other genes) are incongruent with others overall, implying that HGTs and other nonvertical evolutionary process were rampant. A similar pattern could be observed if the size of the gene family was limited to 10. All these results provide evidence to the fact that the archaeal methanogenic core genes have a much stronger consistency regarding the vertical evolutionary signals than other genes, notwithstanding the ubiquitous HGT episodes. The strong evolutionary signal of vertical inheritance among core genes is also supported by phylogenomic tree results, as a high confidence level encompassed most branches for tree I–IV (fig. 1*a*–*d*).

It should be noted that the discrepancies between trees usually arise due to two fundamental reasons: real biological incongruences and estimation errors. Real biological incongruences could be attributed to complex nontree-like evolution, such as HGT (most abundant nonvertical evolutionary processes), gene conversion, and other nonvertical evolutionary process, etc. (Morrison 2011; Puigbo et al. 2012). On the other hand, estimation errors are usually caused by 1) inappropriate data, including insufficient (random errors), low-quality or missing data, laboratory artifact, and inclusion of overly divergent outgroup; 2) misspecified ortholog or paralog relationships; 3) systematic errors caused by improper models or parameters in phylogeny reconstruction; 4) artifacts in phylogeny reconstruction, such as long branch attraction caused by rapid substitution rates. Attributed by these factors, an example would be the inconsistency of the phylogenetic position

of *Methanopyrus kandleri* based on different transcription and ribosomal genes (Brochier et al. 2004). Another example is the convoluted phylogenetic relationship within the rapid diverging of class II methanogen groups and *Halobacterials* (Brochier-Armanet et al. 2011).

To circumvent these issues, we exhausted almost all available complete genomes and applied a thorough evaluation of ortholog families in this study. These procedures effectively prevented the detrimental problems, pertaining missing data, insufficient information, or misspecified orthologs. Furthermore, we also evaluated different reconstruction methods (ML, Bayesian, and NJ) with various substitution models (WAG, JTT, and LG). The result shows that the overall incongruence patterns are extremely similar to those described in figure 1. Nonetheless, some local topologies remained difficult to be recovered due to unusual evolutionary processes in certain trees (even in some phylogenomic tree).

Clanistics Analysis

Recently, an innovative evolutionary analytical method named clanistics analysis was developed. This method analyzes the intricate evolutionary units from thousands of unrooted gene trees to reveal the evolutionary patterns and ecological relationships (Lapointe et al. 2010). When the unrooted phylogenetic tree is dissected into clans and slices according to native and intruder categories, useful insights on the evolutionary history could be unveiled by recurring patterns. In this study, clanistics analysis was first performed on 3,694 gene trees using ten taxonomic order categories. The result shows that genes could be classified into three groups according to their coherent patterns in ten taxonomy categories (patterns A–E are shown in supplementary fig. S3, Supplementary Material online, and their biological meanings are illustrated in supplementary materials, Supplementary Material online). These groups are I) 107 universal vertically evolved genes (with at least six patterns of A1 or C and without any nonvertical patterns); II) 1,359 genes with at least one potential nonvertical evolutionary event among taxonomic orders (with at least one pattern of B1, B2, D1, D2 or E); and III) 2,228 nearly lineage-specific genes (with at most five patterns of A1 or C and without any other nonvertical patterns). Among the 107 genes in group I, 31 were previously defined as archaeal core genes. COG functional enrichment analysis using Fisher's exact test reveals that categories J (translation, ribosomal structure and biogenesis) and K (transcription) of group I genes are significantly overrepresented when compared with all 3,694 genes (Bonferroni adjusted P-value < 0.01). For group III genes (lineage-specific genes), 95% (2,113 out of 2,228) contains no more than two taxonomic orders, suggesting that these gene families are much younger than core genes. Functional enrichment analysis suggests that none of the functional categories is significantly overrepresented or underrepresented in group III. For group II genes, COG functional categories L (replication, recombination and repair) and M (cell wall/membrane/envelope biogenesis) are the only two significantly enriched groups (Bonferroni adjusted P-value < 0.01). This observation reveals that the potential interorder HGTs in methanogens are significantly related to the genes involving in the integration of alien gene into host genomes.

To reveal the phylogenetic relationship among six methanogen orders, clanistics analysis was performed on the whole phylogenetic forest (3,694 trees) with additional taxonomy categories (see Materials and Methods for more details). The result shows that most trees (57%) displays adjacent relationship among the class I orders of *Methanococcales*, *Methanobacteriales*, and *Methanopyrales* (pattern A1 and C). In contrast, most of the other trees reveal that nonmethanogen species (mainly *Thermococcales*, *Thermoplasmatales*, or *Archaeoglobales*) is adjacent to at least one class I methanogen order. These results illustrate that although average phylogenetic signals support the monophyly structure of class I methanogen in the most of the phylogenomic trees (fig. 1*a*), evolutionary history for individual gene is variable and inconsistent. We further investigated the relationship between class II methanogen and nonmethanogenic order *Halobacteriales.* The result shows that *Methanomicrobiales* is closer to *Halobacteriales* (supported by 46% of genes) than to *Methanosarcinales* (supported by 23% of genes).

One attractive feature of clanistics analysis is that the potential phenotypic/environmental adapted genes could be identified from the coherent patterns in a phylogenetic forest (Schliep et al. 2011). A gene cluster is defined as potential phenotype adaptive/related one when it displays 1) at least one perfectly coherent phenotypic pattern (pattern A1) and 2) no coherent taxonomic pattern could be observed. Our analysis for the coherent patterns in methanogenesis phenotypes (hydrogenotrophic, acetoclastic, and methylotrophic) reveals that 114 genes are affiliated with hydrogenotrophic functionalities, 1 with acetoclastic, and 4 with methylotrophic. Further functional enrichment analysis shows that genes of COG category M (Cell wall/membrane/envelope biogenesis) are overrepresented significantly (Bonferroni adjusted P-value <0.05) for hydrogenotrophic phenotype adaptation. It is interesting to note that genes in COG category M are not only related to hydrogenotrophic phenotype adaptation, but are also overrepresented in genes which experienced HGT (coherent group II genes defined above). Among all 119 methanogenesis adaptation-related genes, we noticed one interesting gene belongs to COG3276 group, which encodes a translation factor involved in selenoprotein biosynthesis, a process that is directly related to methanogenesis. We also observed a series of genes (COG0348, COG1032, COG0535, COG1242) related to ferredoxin or Fe-S redox reactions. These genes are potential participants in numerous cellular redox reactions, including methanogenesis. These methanogenesis-related genes would provide useful

information for a better understanding of the formation and adaptation of methanogenesis pathways.

A Coevolution Network in Archaeal Methanogen Phylogenetic Forest

In this study, a comprehensive network analysis based on topology similarity was implemented. The aim is to decompose the hierarchical structure, as well as the global and local features in the archaeal methanogenic phylogenetic forest. Pairwised phylogenetic topology similarities among gene trees were calculated to construct the adjacent matrix in the weighted coevolutionary network by RPHD, a modified version of Penny and Hendy's topology distance (Penny and Hendy 1985) (see Materials and Methods for more details).

First, to reveal the basic characteristic of the coevolutionary network, node strength distribution was calculated and shown in figure 2*a*. The distribution (blue points in fig. 2*a*) follows neither power-law distribution nor log-normal distribution using Pearson χ^2 test (P-value < 0.001). Previous studies showed that in weighted networks, the log–log frequency distribution of the node strength follows a power-law distribution (Pareto distributions) or log-normal distribution, and both patterns reflect a specific processes of network growth (Barabási et al. 1999; Redner 2005; Bhattacharya et al. 2008). In this coevolutionary network, tails are heavier in both raw node strength distribution (histogram in left-bottom corner of fig. 2*a*) and log–log transformation (blue points in fig. 2*a*) than that in best fitting log-normal and power-law distribution. In addition, the same pattern of heavy tail can still be observed when the unweighted node strength was adopted (fig. 2*b*).

To reveal the potential causes for the heavy tail pattern in the node strength distribution, topology similarities and node strengths of simulated data sets with varying HGT rates were calculated (see Materials and Methods for more details). The five simulated data sets share the same taxonomy distribution and gene size as the experimental data, with HGT rates ranging from 0 (fully congruent) to infinite (fully incongruent). With the increasing HGT rate in each family (fig. 3), it is observed that the node strength distributions display a clear transition from strong fat tail (fully congruent) to weak fat tail (sporadic HGT), then to nonfat tail (fully incongruent). The result suggests that the ubiquitous vertical inheritance signals in phylogenetic tree could be denoted by the heavy tail pattern in node strength distribution.

Network Properties and Functional Bias in Coevolutionary Modules

The major objectives of this network analysis are to identify coevolved gene clusters (modules) within the phylogenetic forest, and to identify the potential driving force behind the complex coevolutionary relationships. In this study, the network modules were identified using hierarchical clustering and dynamic branch cut methods (Langfelder et al. 2008) based on TOM, a smoothed out matrix from original adjacency matrix (Dong and Horvath 2007; Yip and Horvath 2007; Langfelder and Horvath 2008). Seven coevolution

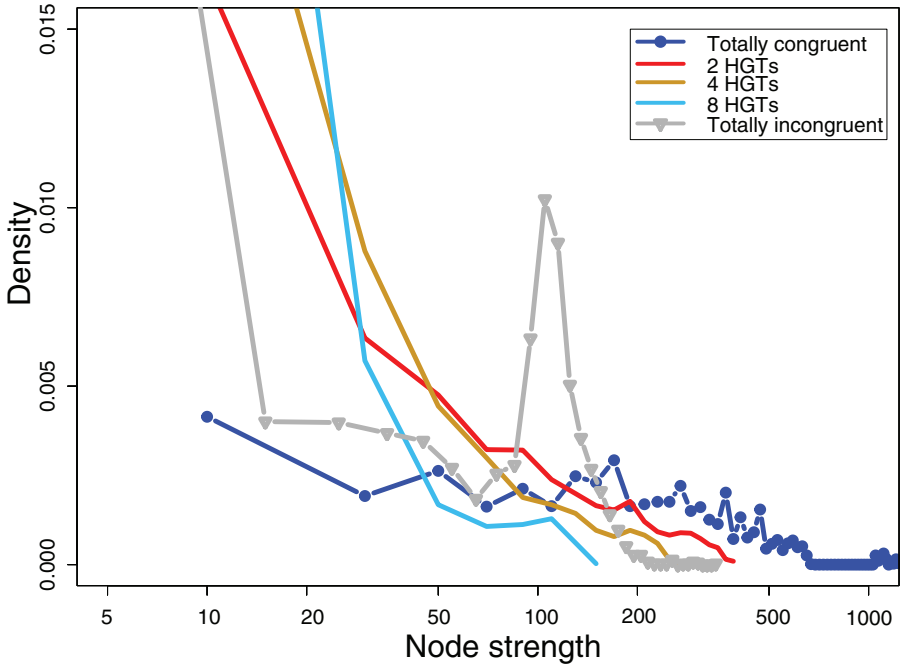

Fig. 3.—Distribution of node strength for simulated data with different HGT rates. Five simulation data sets have same taxonomy distribution and gene size as the real data but different number of HGT events, from 0 (totally congruent) to 2, 4, 8, and infinite (totally incongruent between random topologies).

modules were detected and shown in figure 2*c*. To justify the module classification, the coevolution network was visualized with force-directed layout, in which all nodes were treated as intrinsically repelling beads, and edges acted as springs to pull beads together (fig. 2*d*). The network structure as displayed by force field layout coincides well with the module definition (fig. 2*c*), as modules are located in distinct regions while the members are tightly clustered within modules. The reproducibly analysis for these seven modules (see supplementary materials, Supplementary Material online) revealed that there modules are statistically robust.

Besides the node strength and modularity, we also adopted various measures to investigate the characteristics and roles of different modules in the coevolution network. These measurements, including intramodular connectivity, closeness, betweenness, and local clustering coefficients, reveal that genes in dark orange and yellow modules have high intramodular connection and play key role in the coevolutionary network. More detailed description and discussion about these network measures can be found in the supplementary materials, Supplementary Material online.

Here, we also performed the enrichment analysis based on the functional annotation of all coevolution modules against KEGG (supplementary table S5, Supplementary Material online) and COG databases (supplementary fig. S6, Supplementary Material online). The COG functional categories H (coenzyme transport and metabolism), J (translation, ribosomal structure, and biogenesis), and F (nucleotide transport and metabolism) were found to be overrepresented in the dark orange module (P-value < 0.01 with χ^2 test). We also discovered translation-related genes to be overrepresented according to KEGG annotation. Genes in the dark orange module are mainly core genes, which are the indispensable part of all taxa being studied. This suggests that there exists strong selective pressure on these core functions to maintain their consistent (and vertical) evolutionary signals. Annotation in the yellow module is rather eye catching. As for the genes in yellow module, vital functions (e.g., DNA replication, methane metabolism, etc.) of methanogens were overrepresented (P-value < 0.01 with χ^2 test) according to KEGG annotation. Given the intimate relationship between yellow and dark orange modules, there is a possibility that the yellow module gene families participate in vital biological processes of methanogens and have experienced similar evolutionary force as genes in dark orange module. In addition, an overrepresentation of glutamine and peptidoglycan metabolism pathways was identified in the red module, whereas transporters pathways were identified in the pink module. Meanwhile, it is worthwhile to note that methane metabolism and ribosome pathway genes were found underrepresented in the black module. A similar trend was observed in COG-based functional annotation, wherein functional categories I (lipid transport and metabolism), G (carbohydrate metabolism and transport), V (defense mechanisms), T (signal transduction mechanisms), and M (cell wall/membrane/envelope biogenesis) were overrepresented in the black, blue, brown, pink, and red modules, respectively. Hence, it was of our interest to identify the driving force behind the modularized coevolution in the phylogenetic forest. Herewith, coevolution is a well-established concept in host–pathogen interaction (Buckling and Rainey 2002; Gagneux 2012). In brief, such events could be interpreted as the evolution of certain genes under a common evolutionary driving force. However, coevolutionary events have not been well studied in correlation to biological functions. From our results, there appears to be a trend for gene families of similar or related biological functions to experience common evolutionary pressures. It was hence postulated that such biological function-biased pressures could be one of the factors leading to the modularized evolution of the gene families.

Patchy Taxonomic Structure and Origin of Different Coevolutionary Modules

To investigate the composition of the coevolutionary modules, the taxonomy abundance distributions were calculated. The taxonomy abundance evaluates the proportion of genes from certain taxonomy in given gene families. For example, 20% of *Methanomicrobia* means 20% of OTUs in the tested gene families belongs to *Methanomicrobia*. The results show that all modules have distinct taxonomy composition at either class (fig. 4*a*) or family level (data not shown). Almost all the modules have distinct dominant OTUs at the taxonomic class level, except for the dark orange module. Because the dark orange module has very similar composition with all genes, their genes are expected in almost all species and have similar taxonomy composition as the whole gene pool. On the other hand, the pink and brown modules both have similar dominant OTUs of *Methanomicrobia* at taxonomy class level, whereas their compositions are quite different at taxonomy family level. The intriguing taxonomic distinction in coevolutionary modules raises the question of whether these modules are defined by a similar taxonomy or a similar evolutionary history (topology). To answer this question, we randomized the tree topology with the same taxonomy composition. The results show that although genes from the same module defined in the real data tend to be closer with each other due to similar taxonomy composition (supplementary fig. S7*a*, Supplementary Material online), no similar and statistical robust modules can be identified in the whole network using the same procedures (supplementary fig. S7*b*, Supplementary Material online). Hence, it is reassured that the coevolutionary modules represent the reassembly of genes with similar topologies, instead of similar taxonomy compositions.

To investigate potential causes of the patchy structure and distinct taxonomic compositions among the coevolutionary modules, the phylogenetic depth of family birth (relative

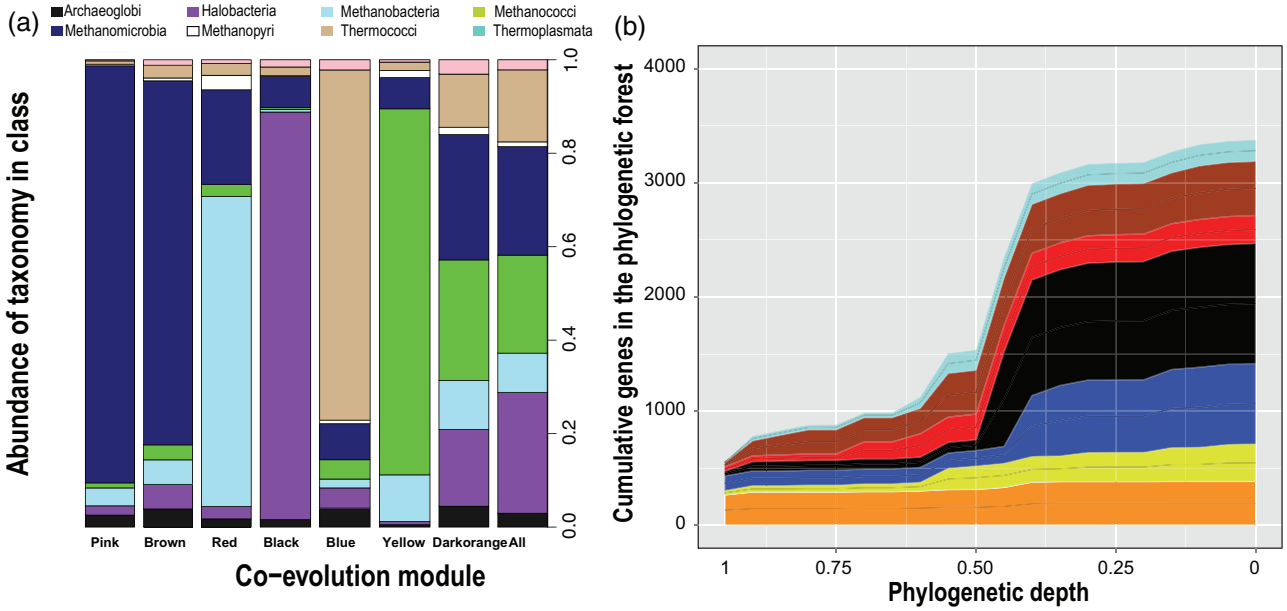

Fig. 4.—Difference of taxonomy composition and phylogenetic depth among modules. (*a*) Distribution of taxonomy abundance in class level. Module names followed the definition in figure 2. (*b*) Relationship between cumulative gene number and phylogenetic depth. *x* axis of phylogenetic depth reflects the relative time of origin for the gene families.

time of family origin) was calculated for each family. The result shows that with the exception of the dark orange module, all other modules genes bloomed at different time periods (fig. 4*b*). Most of the genes in dark orange module originate in the common ancestor of 74 archaeal species. This result provides one possible explanation for the origin of modularized evolution in the phylogenetic forest that the origin of genes burst out in different time period and in distinct evolutionary lineages. Assuming that different gene families originate from different evolutionary lineages, distinct patchy taxonomic structures among different modules, leading to the pan-genome structures, can be reasonably explained.

Quantitative Characterization of HGTs in the Phylogenetic Forest

To further investigate 1) the trends and barriers of HGT quantitatively in archaeal methanogen-related genomes; 2) their influence on the scrambled relationships among methanogen orders; and 3) the contribution of HGT (from archaeal species) to the current genome contents, we employed a recently developed method with explicit evolutionary model. The model incorporates gene birth, speciation, duplication, loss, and horizontal transfer (David and Alm 2011). The distribution of total HGT events in each family (shown in supplementary fig. S5, Supplementary Material online) shows that only around 37% of gene families (ortholog clusters) are not affected by HGT, and 63% of gene families experienced at least one HGT during the entire evolutionary history. This HGT frequency (number of HGT events in the family) is similar to a previous estimation by Kloesges et al. (2011) in a study of *proteobacteria*. In the whole phylogenetic forest, 48% of gene families experienced only 1–2 times HGT events. This suggests that although most gene families experienced HGT in their whole life period (since origin), the HGT frequency in each gene family remains low.

To correct the effect of family size on HGT measurement, the HGT frequency was further normalized with family size by defining HGT rate = HGT frequency/family size. The distributions of normalized HGT rate for seven modules (shown in fig. 5*a*) reveal that genes in dark orange module have the lowest normalized HGT rate, followed by the yellow module; and other module genes have much higher HGT frequencies. For the 45 genes directly involved in methanogenesis pathways, their normalized HGT rate (mean: 0.097 and median: 0.083) is very similar to that of other genes (mean: 0.099 and median: 0.071). For the 119 genes related to methanogenesis pathway, they have a slightly higher average HGT rate (mean: 0.129 and median: 0.125) than other genes. One interesting phenomenon is that there is a significant negative correlation (*P*-value <2e-16) between normalized HGT rate and global node strength in the coevolution network. The correlation coefficient square (R^2) is not very high (0.118) due to the influence of some extreme HGT rate outliers (value of 0). If the correlation is calculated base on genes with at least one HGT event, the negative correlation becomes much stronger ($R^2 = 0.352$) (shown in fig. 5*b*). The results from this correlation analysis suggest that besides the time of origin of

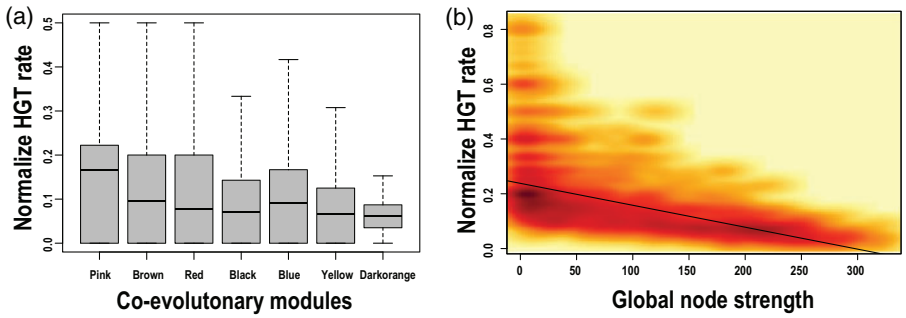

Fig. 5.—Characteristics of normalized HGT rate in coevolutionary network. (*a*) Normalized HGT rate in each coevolution module. (*b*) Linear regression between normalized HGT rate and global node strength in coevolution network. Density of the dots reflects the density of the genes.

gene families, HGTs could also influence the ortholog cluster relationships and the formation of modules in the coevolutionary network. Genes in the central part of the network have much lower HGT rate, whereas genes in the peripheral part of the network could experience more frequent HGTs. This reveals that composite evolutionary processes may contribute to the modularized evolution in the archaeal methanogen phylogenetic forest.

Furthermore, to evaluate the influence of HGT on current genomes in terms of different proteome size, the HGT recipient rate (the ratio of the number of horizontally transferred genes to the total number of genes tested in each genome) was calculated. This rate measures the long-term influence of HGT from archaeal species on the proteomes in each genome. The result shows that the median HGT recipient rate ranges from 0.23 to 0.42 for different taxonomy orders, suggesting at least 20% of the genes in archaeal genomes were transferred from other archaeal genomes. In addition, most of the nonmethanogen species (taxonomy order *Thermococcales*, *Halobacteriales*, *Thermoplasmatales*, and *Archaeoglobales*) acquired fewer foreign genes than methanogen species. One exception is *Methanopyrales*, which has the second lowest HGT recipient rate (~0.23).

Finally, to reveal the tendency and barriers of HGT in the phylogenetic forest, the HGT frequency of donor-recipient pairs was calculated. The result shows that 52% of HGTs are intraorder events and 42% are interorder events. The remaining 6% occurred between more distantly related groups. Genes in the genome of *Halobacteriales*, *Methanomicrobiales*, and *Methanosarcinales* have the highest cumulative rate as HGT donors. The results in figure 6 show that *Archaeoglobales-Archaeoglobales* and *Methanomicrobiales–Methanomicrobiales* are the most frequent HGT paths (donor-recipient). One nonmethanogen order *Thermococcales* has a high HGT rate as donor but a low HGT rate as recipient. This order mainly interacts with other nonmethanogen species, suggesting its special role as a "generous" genetic material reservoir. *Methanomicrobiales* is regarded as the most active order involved in HGTs either as donor or recipient. Among the 15 most frequent HGT paths (marked with stars in fig. 6), seven are related to *Methanomicrobiales*. In contrast, four paths in the ten least frequent donor-recipient paths involve *Halobacteriales* (*Methanococcales–Halobacteriales*, *Halobacteriales–Methanococcales*, *Methanomicrobiales–Halobacteriales*, and *Methanobacteriales–Halobacteriales*). These patterns imply that the scrambled relationship between class II methanogen and related species was caused mainly by intermethanogens HGTs in class II, instead of HGTs between methanogens and related nonmethanogen species (*Halobacteriale*).

Compared with recipient, the donor of HGT is much harder to be detected because the extinct, undiscovered, un-sampled or unsequenced genomes will make the reference species tree based estimation miss the biological direct donor (Ge et al. 2005). In this study, we integrated almost all available complete methanogen-related archaeal genomes to ensure the accuracy of the deduced the archaeal donor-recipient HGT relationship. To evaluate the influence of reference species tree and accuracy of phylogenetic tree over HGT detection, species trees I–III (fig. 1*a*) were tested with multiple substitution models (JTT, WAG, or LG). The results showed that all the elements aforementioned (e.g., HGT rates, HGT recipient rates, and HGT path frequency) were reproduced perfectly, with less than 5% variation in value when compared with the original estimated values. In summary, the robustness of the patterns was confirmed.

Furthermore, HGT rates and donor-recipient relationships were evaluated in methanogens and related archaeal species in this study, providing a reflection of the HGT tendencies (e.g., HGT recipient rate and frequent HGT paths, etc.) among the archaeal species herein studied. Previous studies revealed that when bacteria species were taken into consideration, about one-third of genes in *Methanosarcina mazei* could have been transferred from bacteria horizontally (Deppenmeier et al. 2002). In addition, Nelson-Sathi et al. (2012) discovered that over 70% gene families (containing at least one bacteria homolog) in *Halobacterials* have acquired genes from bacteria. As a result, the effort to quantify HGTs in

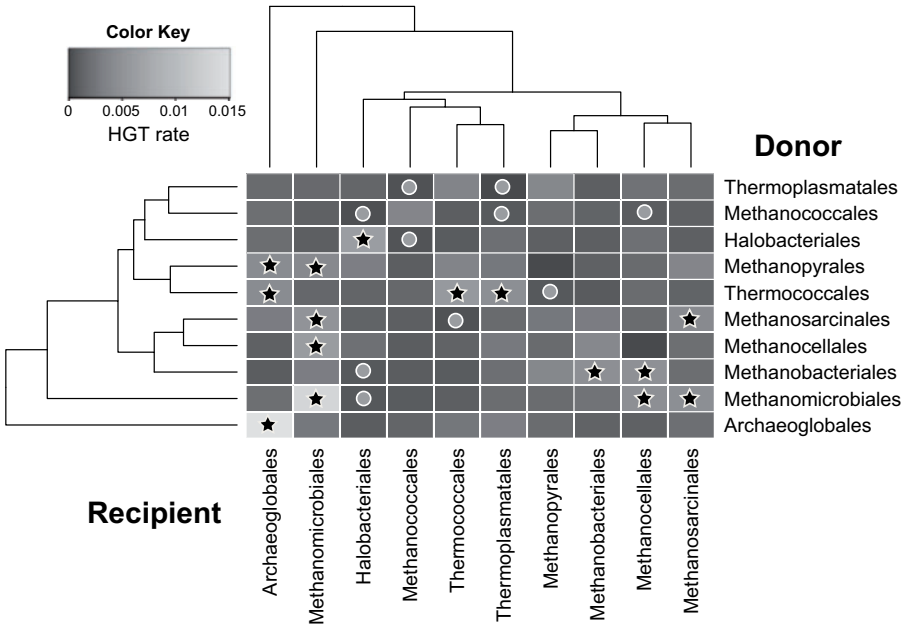

Fig. 6.—Heatmap of donor-recipient frequency among different taxonomy orders. Darker color refers to lower frequency transfer path and lighter color refers to higher frequency transfer path. The top 15 highest and 10 lowest frequent paths were marked with stars and circles, respectively.

all major methanogen-related archaeal lineages in our study provides complementary and useful information regarding more recent gene flow histories within major archaeal lineages.

From a functional genomic perspective, we found categories of J (Translation, ribosomal structure, and biogenesis) and K (Transcription) have significantly (Bonferroni adjusted *P*-value < 0.01 in Mann–Whitney *U* test) lower HGT rates than other categories. This is consistent with complexity hypothesis (Jain et al. 1999). In contrast, the COG categories of M (cell wall/membrane/envelope biogenesis), T (signal transduction mechanisms), and V (defense mechanisms) have significantly higher HGT rates than other categories. Among the genes in COG V category (defense mechanisms) with HGT, COG1131 (ABC-type multidrug transport system) were most frequently observed. These evolutionary patterns in methanogen-related species are concordant with some previous finding that some HGTs could strengthen the defense systems in some prokaryotic species (Godde and Bickerton 2006).

In summary, highly frequent and specific HGT paths among methanogen orders may lead to their scrambled phylogenetic relationships; and the pattern of modularized evolution in the phylogenetic forest is also related to HGTs bias in these gene families.

Conclusions

This study revealed the intrinsic reasons of conflicting phylogenetic signals in previous literatures. The modularized evolutionary pattern in the phylogenetic forest of methanogen-related species would deepen our understanding of the mode and processes of Archaeal origin and evolution, which pave the way for in-depth evolutionary and functional genomics studies in the future. At the same time, the frequently observed horizontally transferred genes between methanogens with COG category M (cell wall/membrane/envelope biogenesis) and V (defense mechanisms) would inspire further investigation to the factors rendering the unique life style and biochemical traits of methanogens.

Supplementary Material

Supplementary materials are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors appreciate the help from Dr Ivan Tasovski in editing the language. The authors also thank three anonymous reviewers for helping us improving the manuscripts. This work was partially supported by Initiative on Clean Energy and Environment, The University of Hong Kong (HKU-ICEE).

Literature Cited

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

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Bapteste E, Brochier C, Boucher Y. 2005. Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. Archaea 1: 353–363.

Bapteste E, et al. 2008. Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. Mol Biol Evol. 25:83–91.

Barabási A-L, Albert R, Jeong H. 1999. Mean-field theory for scale-free random networks, Physica A 272: 173–187.

Bhattacharya K, Mukherjee G, Saramaki J, Kaski K, Manna SS. 2008. The International Trade Network: weighted network analysis and modelling. J Stat Mech. 2008:P02002.

Boone D, Castenholz R, Garrity G. 2001. Bergey's manual of systematic bacteriology. New York: Springer.

Boone DR, Whitman WB, Rouviere P. 1993. Methanogenesis. New York: Chapman and Hall Co.

Boucher Y, Bapteste E. 2009. Revisiting the concept of lineage in prokaryotes: a phylogenetic perspective. Bioessays 31:526–536.

Brilli M, et al. 2008. Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. BMC Bioinformatics 9:551.

Brochier C, Forterre P, Gribaldo S. 2004. Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox. Genome Biol. 5:R17.

Brochier C, Forterre P, Gribaldo S. 2005. An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. BMC Evol Biol. 5:36.

Brochier-Armanet C, Forterre P, Gribaldo S. 2011. Phylogeny and evolution of the Archaea: one hundred genomes later. Curr Opin Microbiol. 14:274–281.

Buckling A, Rainey PB. 2002. Antagonistic coevolution between a bacterium and a bacteriophage. Proc R Soc Lond Ser B Biol Sci. 269: 931–936.

Burggraf S, Stetter KO, Rouviere P, Woese CR. 1991. *Methanopyrus kandleri*: an archaeal methanogen unrelated to all other known methanogens. Syst Appl Microbiol. 14: 346–351.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17: 540–552.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2:e383.

Creevey CJ, McInerney JO. 2005. Clann: investigating phylogenetic information through supertree analyses. Bioinformatics 21:390–392.

Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A. 105:10039–10044.

Dagan T, Martin W. 2006. The tree of one percent. Genome Biol. 7:118.

David LA, Alm EJ. 2011. Rapid evolutionary innovation during an Archaean genetic expansion. Nature 469:93–96.

Deppenmeier U. 2002. The unique biochemistry of methanogenesis. Prog Nucleic Acid Res Mol Biol. 71:223–283.

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

Dong J, Horvath S. 2007. Understanding network concepts in modules. BMC Syst Biol. 1:24.

Doyon J-P, et al. 2010. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. Proceedings of the 2010 International Conference on Comparative Genomics; Ottawa, Canada: Springer-Verlag. p. 93-108.

Droge J, McHardy AC. 2012. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. Brief Bioinform. 13:646–655.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Elizabeth AS, Dina K. 2006. Global anthropogenic methane and nitrous oxide emissions. Energy J. 27:33–44.

Ferry JG 1994. Methanogenesis: ecology, physiology, biochemistry & genetics. New York, London: Springer.

Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. Science 315:476–480.

Gagneux S. 2012. Host–pathogen coevolution in human tuberculosis. Philos Trans R Soc B Biol Sci. 367:850–859.

Gao B, Gupta RS. 2007. Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. BMC Genomics 8:86.

Garcia JL, Patel BK, Ollivier B. 2000. Taxonomic, phylogenetic, and ecological diversity of methanogenic Archaea. Anaerobe 6: 205–226.

Ge F, Wang LS, Kim J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3:e316.

Godde J, Bickerton A. 2006. The repetitive DNA elements Called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J Mol Evol. 62:718–729.

Hedderich R, Whitman W. 2006. Physiology and biochemistry of the methane-producing Archaea. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E, editors. The prokaryotes. New York: Springer Verlag. p. 1050–1079.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A. 96: 3801–3806.

Kelly S, Wickstead B, Gull K. 2011. Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. Proc Biol Sci. 278:1009–1018.

Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Mol Biol Evol. 28: 1057–1074.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.

Langfelder P, Zhang B, Horvath S. 2008. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24:719–720.

Lapointe FJ, Lopez P, Boucher Y, Koenig J, Bapteste E. 2010. Clanistics: a multi-level perspective for harvesting unrooted gene trees. Trends Microbiol. 18:341–347.

Lawrence JG, Hendrickson H. 2005. Genome evolution in bacteria: order beneath chaos. Curr Opin Microbiol. 8:572–578.

Leigh JW, Lapointe FJ, Lopez P, Bapteste E. 2011. Evaluating phylogenetic congruence in the post-genomic era. Genome Biol Evol. 3:571–587.

Leigh JW, Susko E, Baumgartner M, Roger AJ. 2008. Testing congruence in phylogenomic analysis. Syst Biol. 57:104–115.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Liu Y, Whitman WB. 2008. Metabolic, phylogenetic, and ecological diversity of the methanogenic archaea. Ann N Y Acad Sci. 1125: 171–189.

Luo H, et al. 2009. Gene order phylogeny and the evolution of methanogens. PLoS One 4:e6069.

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Curr Opin Genet Dev. 15:589–594.

Morrison DA. 2011. Introduction to phylogenetic networks. [cited 2013 Mar]. Available from: http://www.rjr-productions.org/Networks/index.html.

Nelson-Sathi S, et al. 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc Natl Acad Sci U S A. 109:20537–20542.
Nolling J, et al. 1996. Phylogeny of *Methanopyrus kandleri* based on methyl coenzyme M reductase operons. Int J Syst Bacteriol. 46:1170–1173.
Norman A, Hansen LH, Sorensen SJ. 2009. Conjugative plasmids: vessels of the communal gene pool. Philos Trans R Soc Lond B Biol Sci. 364: 2275–2289.
Opsahl T. 2009. Structure and evolution of weighted networks. London: University of London (Queen Mary College).
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.
Penny D, Hendy MP. 1985. The use of tree comparison metrics. Syst Zool. 34:75–82.
Proulx SR, Promislow DE, Phillips PC. 2005. Network thinking in ecology and evolution. Trends Ecol Evol. 20:345–353.
Puigbo P, Wolf YI, Koonin EV. 2009. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. J Biol. 8:59.
Puigbo P, Wolf YI, Koonin EV. 2012. Genome-wide comparative analysis of phylogenetic trees: the prokaryotic forest of life. Methods Mol Biol. 856:53–79.
Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci. 13:235–238.
Redner S. 2005. Citation statistics from 110 years of physical review. Phys Today. 58:49–54.
Reeve JN, Nolling J, Morgan RM, Smith DR. 1997. Methanogenesis: genes, genomes, and who's on first? J Bacteriol. 179:5975–5986.
Rivera MC, Lake JA. 1996. The phylogeny of *Methanopyrus kandleri*. Int J Syst Bacteriol. 46:348–351.
Sakai S, et al. 2008. Methanocella paludicola gen. nov., sp nov., a methane-producing archaeon, the first isolate of the lineage 'Rice Cluster I', and proposal of the new archaeal order Methanocellales ord. nov. Int J Syst Evol Microbiol. 58:929–936.
Salichos L, Rokas A. 2011. Evaluating ortholog prediction algorithms in a yeast model clade. PLoS One 6:e18755.
Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol. 19: 101–109.
Schlesinger WH. 1997. Biogeochemistry: an analysis of global change. San Diego (CA): Academic Press.
Schliep K, Lopez P, Lapointe FJ, Bapteste E. 2011. Harvesting evolutionary signals in a forest of prokaryotic gene trees. Mol Biol Evol. 28: 1393–1405.
Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27: 592–593.
Segata N, Huttenhower C. 2011. Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. PLoS One 6:e24704.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51:492–508.
Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17: 1246–1247.
Slesarev AI, et al. 2002. The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. Proc Natl Acad Sci U S A. 99:4644–4649.
Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–463.
Susko E, Leigh J, Doolittle WF, Bapteste E. 2006. Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the gamma-proteobacteria. Mol Biol Evol. 23:1019–1030.
Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637.
Treangen TJ, Rocha EP. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 7: e1001284.
Venables WN, Ripley BD 2002. Modern applied statistics with S. New York, London: Springer.
Wright AD. 2006. Phylogenetic relationships within the order Halobacteriales inferred from 16S rRNA gene sequences. Int J Syst Evol Microbiol. 56:1223–1227.
Yarza P, et al. 2008. The all-species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol. 31:241–250.
Yip AM, Horvath S. 2007. Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics 8:22.

Associate editor: Bill Martin