# Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution

**Tal Dagan[†‡], Yael Artzy-Randrup[§¶], and William Martin[†]**

[†]Institut für Botanik III, Heinrich-Heine Universität Düsseldorf, Universitätsstrasse 1, 40225 Düsseldorf, Germany; and [§]Biomathematics Unit, Department of Zoology, Faculty of Life Sciences, Tel Aviv University, Ramat-Aviv 69978, Israel

Edited by W. Ford Doolittle, Dalhousie University, Halifax, NS, Canada, and approved May 6, 2008 (received for review January 23, 2008)

**Lateral gene transfer is an important mechanism of natural variation among prokaryotes, but the significance of its quantitative contribution to genome evolution is debated. Here, we report networks that capture both vertical and lateral components of evolutionary history among 539,723 genes distributed across 181 sequenced prokaryotic genomes. Partitioning of these networks by an eigenspectrum analysis identifies community structure in prokaryotic gene-sharing networks, the modules of which do not correspond to a strictly hierarchical prokaryotic classification. Our results indicate that, on average, at least 81 ± 15% of the genes in each genome studied were involved in lateral gene transfer at some point in their history, even though they can be vertically inherited after acquisition, uncovering a substantial cumulative effect of lateral gene transfer on longer evolutionary time scales.**

community structure | molecular phylogeny | microbial genomes

Over evolutionary time, prokaryotic genomes undergo lateral gene transfer (LGT), the mechanisms of which entail acquisition through conjugation, transduction, transformation, and gene transfer agents (1, 2) in addition to gene loss (3). This leads to different histories for individual genes within a given prokaryotic genome and networks of gene sharing across chromosomes among both closely and distantly related lineages (4–9). In genome comparisons, LGT is traditionally characterized in terms of conflicting gene trees (10, 11) or aberrant patterns of nucleotide composition (12). Networks should, in principle, be able to more fully uncover the dynamics of prokaryotic chromosome evolution (9). Networks are currently used to model various aspects of biological systems such as gene regulation (13), metabolic pathways (14), protein interactions (15), conflicting phylogenetic signals (16), and ecological interactions (17). A network analysis of gene distributions across prokaryotic genomes should provide new insights into the contribution of LGT to microbial evolution.

A network is a graphical representation of a set of "agents," or vertices, linked by edges that represent the connections or interactions between these agents. The degree of any given vertex is defined as the total number of edges attached to it (for a glossary of network terms, see ref. 18). A network of $N$ vertices can be fully defined by matrix, $A = [a_{ij}]_{N*N}$, with $a_{ij} = a_{ji} \neq 0$ if a link exists between node $i$ and $j$, and $a_{ij} = a_{ji} = 0$ otherwise. In the study of biological networks, the vertices might represent genes or neurons and the links might represent regulation pathways or synaptic connections. In the case of prokaryotic genome evolution, each genome is represented by a vertex, $i$, whereas the elements of the matrix, $A$, correspond to the number of shared genes between genome pairs, $a_{ji}$. Gene sharing can result either from vertical inheritance or from LGT.

## Results

**Modules and Community Structure in Networks of Shared Genes.** To obtain matrices of all shared genes, we used standard clustering procedures to assort the 539,723 proteins encoded among 181 sequenced prokaryotic genomes into groups of shared sequence similarity that we designate as protein families (see *Materials and Methods*). At the 25% amino acid identity threshold ($T_{25}$), clustering yields 54,349 families containing 431,492 individual genes, with 108,231 singletons that were not considered further. Higher sequence similarity thresholds yield larger numbers of less inclusive families for fewer numbers of more highly conserved proteins (Table 1).

Each sequence identity threshold delivers a binary matrix of presence or absence for each family that is readily assorted into a 181 × 181 matrix-represented gene-sharing network of vertices (genomes) and edges (number of shared genes). There are 16,290 possible edges in the network, all of which have weight ≥1 at clustering thresholds ≤40%, meaning that all of the genomes in the network of shared genes share at least one gene family, and therefore are interconnected with each other, thereby forming a complete network, or a "clique" in network terms (19). But the clique property is not attributable to universally distributed genes only, because the use of higher similarity thresholds reduces the size of protein families and the number of edges (Table 1). Only six families are present in all genomes at $T_{25}$, only two are present in all genomes at $T_{30}$, one at $T_{35}$, and none are present in all genomes at $T_{40}$ and higher. Rather, the clique results from the high connectivity of gene-sharing patterns for 54,349 to 66,118 ($T_{25}$ to $T_{40}$) families distributed among 181 genomes ranging in size from 307 to 4,820 families each, with a mean of 2,133 ± 1,252 at $T_{30}$.

Unlike metabolic networks (13) or the Internet (20), the network of shared genes contains no "hubs" (20), that is, a few genomes that are far more connected than all others. However, some groups of genomes are more strongly interconnected among themselves than with others in the network, thereby forming communities (21–24). We examined the community structure in the network by a division into modules (23): for each possible bipartition of the network, a modularity function is defined as the number of edges within a community minus the expected number. Maximizing this modularity function by using the leading eigenvector of the matrix form of this function yields the modules of the network (23).

If little or no lateral gene transfer existed in the present genome data, and if the taxonomic groups shown were natural in terms of a hierarchical classification (9), we would expect modules to divide the network strictly along recognized taxonomic boundaries. But the converse is observed (Fig. 1*A*), as a few examples illustrate. The mosaicism among proteobacteria that is well documented in extensive gene phylogeny studies (25) and whose mechanisms involve gene transfer agents (2) is

Author contributions: T.D., Y.A.-R., and W.M. designed research; T.D., Y.A.-R., and W.M. performed research; T.D. and Y.A.-R. analyzed data; and T.D., Y.A.-R., and W.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[‡]To whom correspondence should be addressed. E-mail: tal.dagan@uni-duesseldorf.de.

[¶]Present address: Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109.

This article contains supporting information online at www.pnas.org/cgi/content/full/0800679105/DCSupplemental.

**Table 1. Number of protein families (excluding singletons), edges, and modules in the shared gene network for different protein similarity cutoffs**

| Cutoff | No. families | No. proteins | No. edges | No. modules | Families within modules, % | Edges within modules |
|---|---|---|---|---|---|---|
| 25 | 54,349 | 431,492 | 16,290 | 4 | 73 | 5,398 (33%) |
| 30 | 57,670 | 412,427 | 16,290 | 5 | 80 | 4,658 (29%) |
| 35 | 61,981 | 391,664 | 16,290 | 4 | 79 | 6,136 (38%) |
| 40 | 66,118 | 367,651 | 16,290 | 6 | 85 | 4,041 (25%) |
| 45 | 68,906 | 334,381 | 16,275 | 6 | 86 | 4,222 (26%) |
| 50 | 71,013 | 308,172 | 16,260 | 6 | 92 | 3,981 (24%) |
| 55 | 71,569 | 280,315 | 15,936 | 8 | 90 | 2,493 (16%) |
| 60 | 70,639 | 252,952 | 14,311 | 11 | 92 | 2,126 (15%) |
| 65 | 68,311 | 225,878 | 13,305 | 13 | 95 | 2,197 (15%) |
| 70 | 64,714 | 199,700 | 12,488 | 13 | 97 | 2,116 (17%) |
| 75 | 60,000 | 174,415 | 9,585 | 21 | 97 | 1,665 (17%) |
| 80 | 54,358 | 149,511 | 3,293 | 32 | 98 | 1,328 (40%) |
| 85 | 47,982 | 125,488 | 1,874 | 48 | 98 | 735 (39%) |
| 90 | 41,023 | 102,223 | 924 | 68 | 98 | 578 (63%) |

evident within the gene-sharing network. The $\alpha$-, $\beta$-, and $\gamma$-proteobacteria form a nearly discrete module at the 25% amino acid identity threshold ($T_{25}$), with $\alpha$-proteobacteria representing a discrete module at $T_{50}$, the network of which comprises a smaller number of more highly conserved proteins. Some $\gamma$-proteobacteria form a module with all $\beta$-proteobacteria at $T_{55}$, but the two modules do not correspond to the rRNA-based taxonomic framework. By contrast, some of the $\delta$- and $\varepsilon$-proteobacteria sampled tend to cluster with firmicutes, a group of Gram-positive bacteria encompassing bacilli, clostridia, and mollicutes. The methanogens—some of which also possess gene transfer agents (2)—tend to cluster with sulfate-reducing $\delta$-proteobacteria, possibly reflecting similar gene collections by virtue of similar habitats (26), in agreement with the ≈30% eubacterial genes found in *Methanosarcina* genomes (27), which, however, went undetected in LGT analyses based on tree comparisons (28). Cyanobacterial gene phylogenies uncover mosaicism (6), as do modules in the gene-sharing network. At $T_{30}$, the cyanobacteria form a module with some $\alpha$-proteobacteria (Fig. 1*A*), as seen in the networks showing only the edges within modules (Fig. 1*B*), whereas at $T_{40}$ (Fig. 1*C*) the same module includes the chlamydias. Phylogenies suggest that photosynthetic eukaryotes might have acquired ≈20 genes from the *Chlamydia* lineage (29), the modules show that gene exchanges among prokaryotes could produce the same result. One actinobacterium in our sample, *Symbiobacterium thermophilum*, falls within the module of Gram-positive bacteria for all thresholds, congruent with analyses of overall gene content (30). The present networks show that gene sharing across lineages is a substantial component of natural variation among microbes (4, 28).

Fig. 1*B* depicts the five modules and all 4,658 within-module edges for $T_{30}$. Vertex radius in the figure is not scaled to genome size, but instead to centrality, also known as community centrality (23), that is, the level to which each genome contributes to the overall modularity of the network (23). Small vertices have low centrality, are less connected within the module, and have little contribution to modularity; the converse is true for large vertices. Fig. 1*C* shows the six modules at $T_{40}$ and all 4,041 within-module edges. Because the complete gene-sharing networks form cliques, their graphical representations are dense (supporting information (SI) Fig. S4). Although it is possible to generate bifurcating trees from such patterns of shared genes (31, 32), it is clear that no single tree of whatever topology could adequately account for complete pattern of gene sharing among these genomes in the fully represented network.

**Cumulative Impact of LGT During Prokaryote Evolution.** So far, we have considered all shared genes, whether vertically or laterally inherited. How many of these shared genes reflect vertical inheritance from a common ancestor, how many reflect LGT, and how many reflect commonly inherited acquisitions? Genes that are infrequently shared across broad taxonomic boundaries are said to have patchy distributions (33). They provide an objective criterion for discriminating between LGT and vertical inheritance, because if one attributes all patchy occurrences to differential loss only, then the sizes of the inferred ancestral genomes underpinning those losses become untenably large (34). That constraint can be used to obtain a lower bound estimate for LGT frequency, if we embrace three simplifying assumptions: (*i*) that the gene tree within each protein family is completely compatible with a reference tree, (*ii*) that all genes are orthologous, and (*iii*) that gene loss is not penalized (35). Starting with a "genome of Eden" (34) harboring 57,670 genes and reasoning that ancestral genome sizes were not fundamentally different in the past from those observed today, incremental allowance of LGT to account for patchy distributions specifies the minimum amount of LGT that is required to bring the distribution of inferred ancestral genome sizes into agreement with the distribution of 181 modern genome sizes. The LGT amount so specified is a minimum because no LGT events are inferred from conflicting gene trees (35). In the present data for the inclusive $T_{30}$ threshold, the only accepted model ($P = 0.79$ using the Wilcoxon test; Fig. S5) allows up to three LGTs per gene family (35), and results in an average of 1.06 LGT events per gene family. As the reference tree, we use an ML tree of the rRNA operon (Fig. 2*A*) with monophyly of all taxonomic groups. This approach attributes as many gene distribution patterns as possible to vertical inheritance and hence delivers a far-too-conservative lower bound for LGT frequency, recalling that all gene trees are assumed to be congruent (35). Those gene distributions that do not map exactly onto the 361 vertical edges, with losses unpenalized and LGT constrained by ancestral genome size only, constitute the minimal lateral network (MLN). The MLN consists of 361 vertices, of which 181 are contemporary genomes and 180 are ancestral genomes (internal nodes in the reference tree). The vertices are interconnected either by the branches of the reference tree that represent vertical inheritance or by lateral edges that represent lateral inheritance.

For genes that have undergone more than one LGT, the number of edges in the MLN exceeds the minimum number of LGTs required to account for the distribution. To address network properties for the MLN, 1,000 replicates were therefore generated in which the number of lateral edges and the minimum number of LGTs for genes transferred more than once exactly

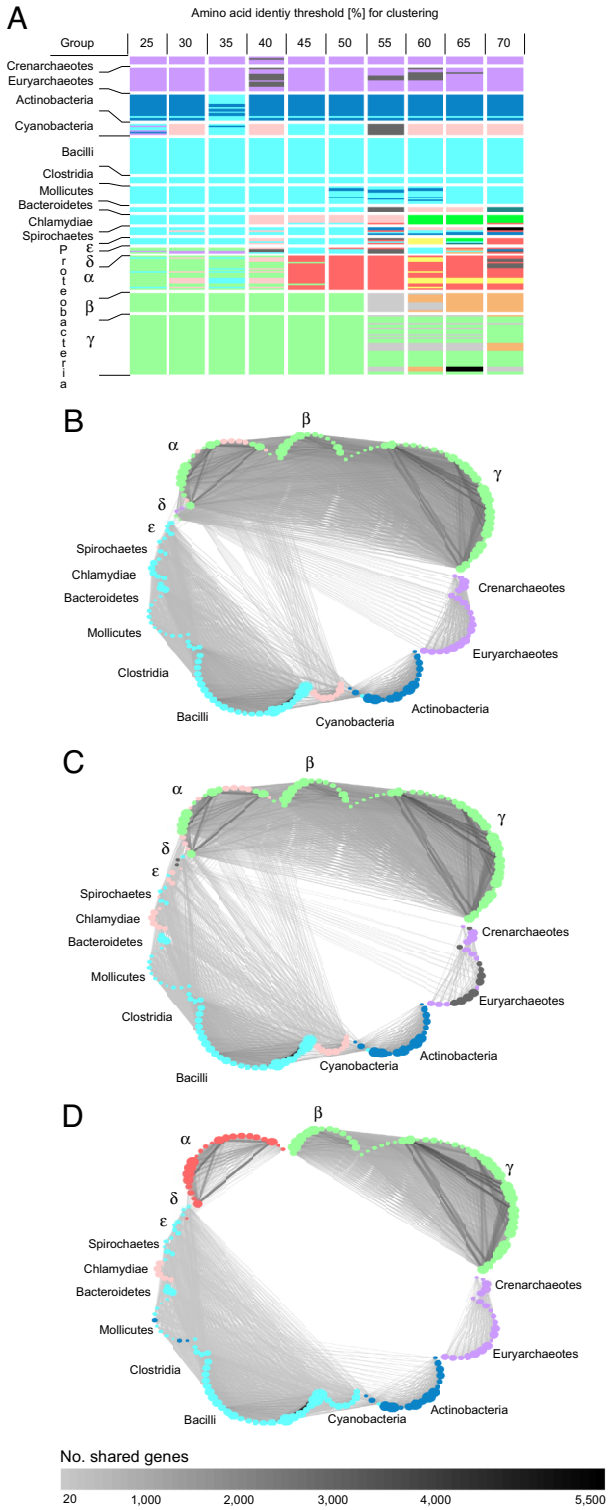


**Fig. 1.** Modules in networks of shared genes. (*A*) Modules detected (see *Materials and Methods*) are shown as colored boxes within columns for thresholds from $T_{25}$ to $T_{70}$. Currently recognized higher-level taxonomic groups are indicated in rows for comparison. For example, for the network at $T_{25}$ all but one actinobacteria and the cyanobacterium, *Thermosynechococcus elongatusform*, form one module, which is dark blue. An expanded version of the panel containing all species names is given in Figs. S1–S3. (*B*) Modules in the gene-sharing network at $T_{30}$. Only edges connecting within modules are shown, edge shading is proportional to the number of shared genes per edge (see scale). Vertices (genomes) are colored according to their module as in *a*, vertex radius is linearly scaled to centrality (see text). (*C*) Modules in the gene-sharing network at $T_{40}$. (*D*) Modules in the gene-sharing network at $T_{50}$.

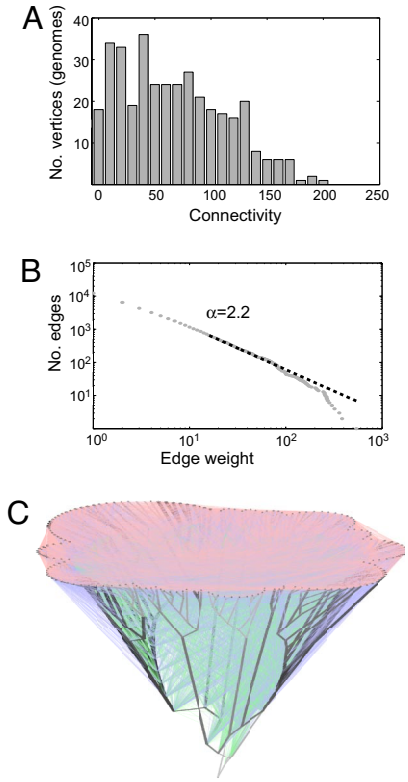


**Fig. 2.** Properties of the minimal LGT network. Properties are shown for a randomly selected replicate. The coefficient of variation (CV) for the whole data were ≪1% (Fig. S6). (*A*) Distribution of connectivity, the number of one-edge-distanced neighbors for each vertex, in the MLN. Note the absence of vertices that are far more highly connected than others (hubs). (*B*) Frequency distribution of edge weight in the lateral component of the MLN. (*C*) A three-dimensional projection of the MLN. Edges in the vertical component are shown in the same grayscale as in Fig. 3. Vertices inferred as gene origin in the same protein family are connected by a lateral edge. Lateral edges are classified into three groups according to the types of vertices they connect within the vertical component: 4,040 external-external edges (red), 5,862 internal-external edges (blue), and 2,345 internal-internal edges (green).

correspond (see *Materials and Methods*). The internal and external vertices of the MLN for the broad sample of genes at $T_{30}$ are linked by 12,262 ± 32 lateral edges. There are no hub genomes with exceptional connectivity (number of edges per vertex) in the MLN. Connectivity ranges between 0 and 191–213 edges per genome among the 1,000 replicates with a mean of 67–69 and a median of 59–64 edges (Fig. 2*A*). The Clustering Coefficient (36) of the MLN ranges between 0.43 and 0.44, which is significantly higher ($P < 0.05$) than expected for a random network with the same connectivity (37) per genome. The mean shortest path of the MLN ranges between 2.09 and 2.17 edges. Combined with the high level of clustering, this means that the MLN forms a small world network (19, 20). LGTs involving one or few genes comprise the majority of the MLN. The number of genes shared between each pair of genomes has a mean of 2.09–2.17 and follows a power law fit in all MLN replicates with $\hat{\alpha} = 2.08$–$2.35$ at the 95% confidence interval (Fig. 2*B*) by using a maximum likelihood test (38). In biological terms, the power law fit means that small numbers of genes are transferred far more often than large numbers of genes and that the relationship between edge weight and edge frequency is log linear (Fig. 2*B*). Because the method of LGT inference is robust with respect to tree topology and rooting (35), the same basic network properties are obtained for the MLN inferred by using a neighbor-joining (NJ) reference tree for comparison (Fig. S7).

EVOLUTION

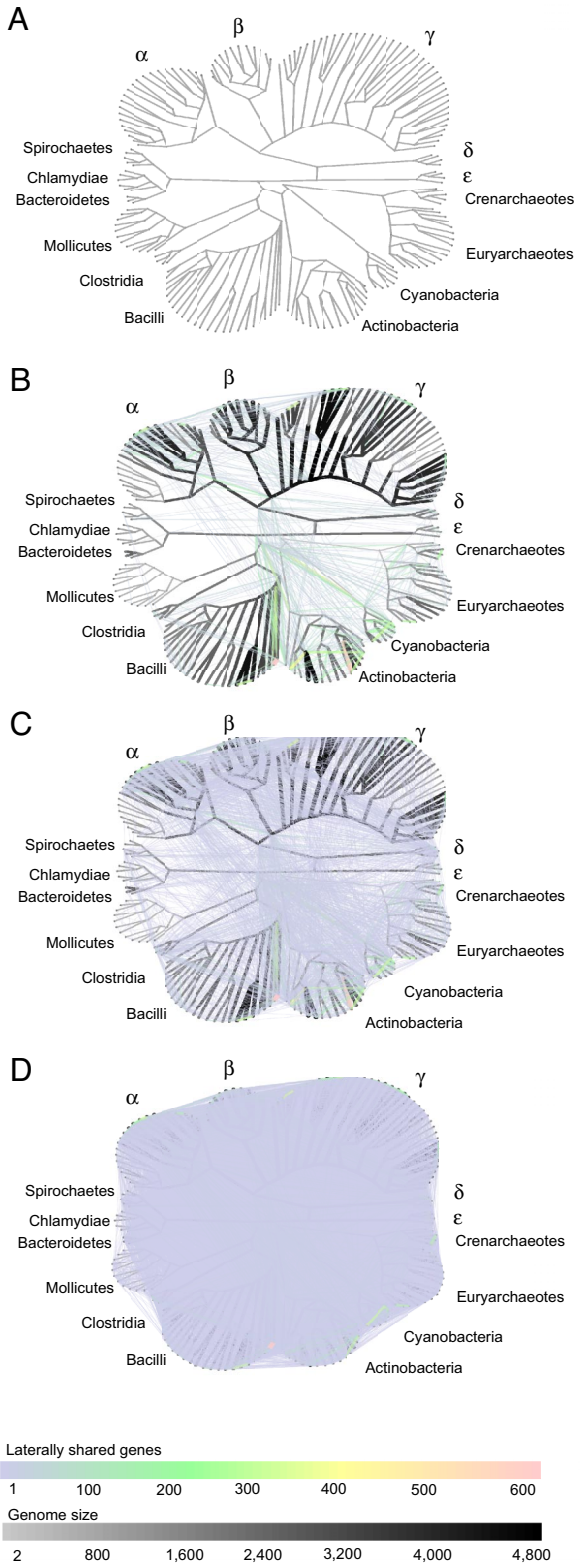


**Fig. 3.** A minimal LGT network for 181 genomes. (*A*) The reference tree used to ascribe vertical inheritance for inference of the MLN (see *Materials and Methods*). (*B*) The network showing only the 823 edges of weight ≥20 genes. Vertical edges are indicated in gray, with both the width and the shading of the edge shown proportional to the number of inferred vertically inherited genes along the edge (see the scale). The lateral network is indicated by edges that do not map onto the vertical component, with number of genes per edge indicated in color (see the scale). (*C*) The MLN showing only the 3,764 edges of weight ≥5 genes. (*D*) The MLN showing all 15,127 edges of weight ≥1 gene in the MLN.

**Table 2. Average ± SD percent of genes involved in LGT per genome across lineages**

| Group | % acquired in genome | % acquired in lineage | Mean genome size |
|---|---|---|---|
| Epsilonproteobacteria | 18 ± 8 | 75 ± 6 | 1,157 ± 60 |
| Deltaproteobacteria | 34 ± 2 | 98 ± 1 | 1,694 ± 222 |
| Gammaproteobacteria | 11 ± 7 | 90 ± 6 | 2,984 ± 1,197 |
| Betaproteobacteria | 12 ± 10 | 86 ± 9 | 3,345 ± 1,020 |
| Alphaproteobacteria | 13 ± 11 | 83 ± 13 | 2,177 ± 1,346 |
| Spirochaetes | 13 ± 16 | 60 ± 25 | 1,001 ± 1,28 |
| Chlamydiae | 4 ± 7 | 49 ± 15 | 850 ± 61 |
| Bacteroidetes | 8 ± 2 | 57 ± 10 | 2,185 ± 646 |
| Mollicutes | 11 ± 6 | 72 ± 12 | 429 ± 46 |
| Clostridia | 24 ± 4 | 89 ± 5 | 1,891 ± 83 |
| Bacilli | 14 ± 11 | 87 ± 9 | 2,498 ± 966 |
| Actinobacteria | 21 ± 19 | 82 ± 12 | 2,227 ± 1,283 |
| Cyanobacteria | 27 ± 20 | 79 ± 11 | 1,582 ± 447 |
| Euryarchaeota | 19 ± 16 | 69 ± 13 | 1,403 ± 539 |
| Crenarchaeota | 25 ± 12 | 70 ± 14 | 1,234 ± 563 |
| All | 15 ± 13 | 81 ± 15 | 2,133 ± 1,252 |

The MLN can be represented in three dimensions (Fig. 2*C*) to highlight the frequency of gene sharing that cannot be attributed to vertical inheritance as constrained by ancestral genome size. Of the 12,262 ± 32 lateral edges, 33 ± 0.13% connect external nodes of the reference tree only (red), corresponding to genes with the most patchy distributions. The 48 ± 0.16% edges that connect external nodes to internal nodes (blue) correspond to genes shared by a group and an outlier, whereas the 19 ± 0.13% that connect internal nodes (green) correspond to genes patchily shared by two or more groups. The plotting threshold for edge weight decisively influences the degree of connectivity among genomes that is implied in the network graph. Only 493 ± 6 (4 ± 0.05%) edges carry 20 genes or more (Fig. 3*B*), 2,529 ± 17 (20 ± 0.15%) carry five genes or more (Fig. 3*C*), whereas 5,773 ± 44 (47 ± 0.3%) carry only one. The densely connected network showing all edges is shown in Fig. 3*D*.

Lateral edges connected to external nodes correspond to comparatively recent inferred acquisitions, and the average proportion (% ± SD) thereof is 15 ± 13% of the genes across all 181 genomes (Table 2). For some groups with small genomes, such as chlamydias (4 ± 7%) or mollicutes (11 ± 6%), recent transfers are inferred to be rare. There is a weak but significant correlation ($r = -0.08$, $P < 0.05$) between genome size and recent acquisitions, meaning that the former can account for $\ll$1% of variation in the latter. The estimated proportion of ≈15% recent acquisitions per genome obtained here from gene distributions is consistent with values inferred from analysis of nucleotide patterns (12) and codon bias (39).

More heavily debated than recent acquisitions is the cumulative role of LGT over longer evolutionary time scales (4, 40). For each genome, we therefore calculated the percentage of genes that were connected by lateral edges at any point in their history as inferred from the MLN. The result indicates that on average, 81 ± 15% of the genes in each genome were involved in LGT at some point in their history, with 61 of the 181 individual values exceeding 90% (Table S1) and the averages for each group given in Table 2. Once acquired, genes can be vertically inherited within a group (39, 40), and the MLN suggests that this has occurred for the vast majority of genes, and probably all, given that we have inferred no LGT events from conflicting gene trees, during prokaryote genome evolution. Methods of LGT inference other than those used here, such as gene tree conflicts (28) or nucleotide frequency (12), could also be used to construct networks of vertical and lateral inheritance.

**Table 3. Lateral edge (LE) frequencies between and within groups in the MLN**

| Group | *n** | Normalized LE frequency† | | Median LE weight‡ | |
|---|---|---|---|---|---|
| | | int | ext | int | ext |
| Epsilonproteobacteria | 4 | 0.99 ± 0.01 | 1.1 ± 0.02 | 13–38 | 1–1 |
| Deltaproteobacteria | 4 | 2.0 ± 0 | 2.1 ± 0.02 | 14–28 | 2–2 |
| Gammaproteobacteria | 39 | 12.5 ± 0.1 | 12.1 ± 0.1 | 2–3 | 1–1 |
| Betaproteobacteria | 13 | 5.1 ± 0.1 | 5.9 ± 0.04 | 5–7 | 2–2 |
| Alphaproteobacteria | 22 | 5.6 ± 0.1 | 7.1 ± 0.04 | 3–4 | 2–2 |
| Spirochaetes | 5 | 1 ± 0 | 1.3 ± 0.02 | 2–2 | 1–2 |
| Chlamydiae | 6 | 1.4 ± 0.1 | 0.5 ± 0.01 | 1–3 | 1–1 |
| Bacteroidetes | 3 | 0.4 ± 0 | 1.4 ± 0.02 | 25–29 | 1–1 |
| Mollicutes | 12 | 3.9 ± 0.1 | 0.6 ± 0.02 | 2–2 | 1–1 |
| Clostridia | 4 | 1 ± 0 | 2.1 ± 0.03 | 11–21 | 1–2 |
| Bacilli | 24 | 9.7 ± 0.1 | 7 ± 0.05 | 3–4 | 1–1 |
| Actinobacteria | 17 | 7.2 ± 0.1 | 7.1 ± 0.05 | 5–6 | 1–2 |
| Cyanobacteria | 7 | 2.8 ± 0.05 | 3.3 ± 0.03 | 20–34 | 1–2 |
| Euryarchaeota | 16 | 6.4 ± 0.1 | 4.8 ± 0.04 | 2–3 | 1–1 |
| Crenarchaeota | 5 | 1.6 ± 0 | 1.5 ± 0.02 | 7–12 | 1–2 |

*Number of genomes within the group

†For internal edges (int), number of internal edges per no. of nodes within the group; for external edges (ext), number of external edges per no. of nodes outside the group.

‡Range of median number of genes per lateral edge in the 1,000 MLN replicates

Networks can also address the issue of whether genes are exchanged more frequently within than between groups (5, 25). The number of edges between taxonomic groups in the MLN is anywhere from 3 to 300 times higher than the number of edges within groups (Table 3, Table S2), but the differences dissipate after normalization for the number of vertices with which edges can connect in the MLN (i.e., the number of vertices within the compared groups, sample sizes of which vary). However, the median number of genes per edge is 4–20 times higher for lateral edges that connect within groups than between groups, indicating that fixation after gene sharing within groups occurs either more frequently, or that transfers within groups involve larger numbers of genes per event than transfers between groups, or both.

## Discussion

Traditional approaches to characterizing prokaryote genome evolution focus on the component of the genome that fits the metaphor of a tree. The issue is how large that component is over the fullness of evolutionary time (9). Although there can be little doubt that a considerable component of prokaryote genome evolution over recent evolutionary time scales is fundamentally treelike in nature (12, 39), differences in gene content exceeding 30% among individual strains of *E. coli* (42) demonstrate that LGT has substantial impact on genome evolution even at the species level. Our findings indicate that, over long evolutionary time scales, the cumulative role of LGT leaves almost no gene family among prokaryotes untouched. That conclusion is consistent with the findings of Sorek *et al.* (43) who showed that *E. coli* accepts virtually all prokaryotic genes offered to it in the laboratory, indicating that genuine barriers to LGT are low in that model organism.

The conservative lower bound nature of our method for inferring LGT among prokaryotes indicates that evolution by lateral transfer affects the vast majority of gene families, and probably all, but possibly at a low rate. This results in a modest proportion of recently acquired genes in contemporary genomes, but a cumulative impact that snowballs over evolutionary time. When all genes and genomes are considered, the tree paradigm fits only a small minority of the genome at best (27, 44); hence, more realistic computational models for the microbial evolutionary process are needed. By accounting for all genes, including the many that are patchily distributed across broad taxonomic boundaries, networks uncover a view of microbial genome evolution that incorporates LGT as a quantitatively important mechanism of natural variation among prokaryotic genomes. In contrast to trees, networks thus present a means of reconstructing microbial genome evolution that accommodates the incorporation of foreign genes, hence, more realistically modeling the process as it occurs in nature.

## Materials and Methods

**Gene-Sharing Network.** Proteomes from sequenced genomes of 22 archaebacteria and 159 eubacteria were downloaded from the National Center for Biotechnology Information web site (http://www.ncbi.nlm.nih.gov/; August 2005version). For each species, only the strain with the largest number of genes was used. All proteins were clustered by similarity into gene families by using the reciprocal best BLAST hit (BBH) approach (45). Each protein was BLASTed against each of the genomes. Pairs of proteins that resulted as reciprocal BBHs of E-value $<1^{-10}$ were aligned by using ClustalW (46). Protein pairs with above the sequence identity threshold (25–90%) where clustered into protein families of ≥2 members by using the MCL algorithm to set the inflation parameter, *I*, to 2.0 (47). Gene distribution in genomes is highly nonrandom (35). Previous work has shown that *I* has little influence in nonrandom networks (48). When we clustered with *I* set to 1.8 or 2.2, the gene family size distributions did not differ significantly from that of $I = 2.0$ ($P = 0.09$ and $P = 0.12$, respectively, by using Wilcoxon test), indicating that *I* has little influence in the present analysis. The number of shared genes between each genome pair was calculated as the number of protein families in which both genomes are present.

A division of the network into modules, or communities, is based on maximizing a modularity function defined as the number of edges within a community minus the expected number of edges. Initially an optimal division into two components is found by maximizing this function over all possible divisions by using spectral optimization, which is based on the leading eigenvector of the matching modularity matrix. To further subdivide the network into more than two communities, additional subdivisions are made, each time comparing the contribution of the new subdivision with the general modularity score of entire network. This process is carried out until there are no additional subdivisions that will increase the modularity of the network as a whole (23).

**Lateral Network.** For the reference tree, rRNA operon (16S, 23S, and 5S) sequences were first aligned (46) for each of the groups shown in Table 2. From the concatenated alignments, gapped sites were removed and a maximum likelihood tree of each group was inferred by using dnaml (49) with the default parameters or neighbor with Kimura 2 parameters. From each group alignment, a consensus sequence was constructed by concatenating the most abundant nucleotide in each alignment column into a single sequence. The consensus sequences were used to infer the tree of groups with dnaml and to

EVOLUTION

root each neighboring group subtree; leaves in the tree of groups were replaced with each rooted group subtree. Presence and absence of protein families were superimposed on the reference tree and LGTs inferred to yield gene presence or absence for all protein families at internal nodes as described in ref. 35. Edges connecting the same two nodes for different protein families are joined to form an edge that is weighted according to the number of protein families in which it appears.

**Network Analysis.** The number of genes shared by each pair of genomes was fitted by a power law distribution by using discrete maximum likelihood estimators along with a goodness-of-fit-based approach to estimate the lower cutoff for the scaling region (38). The distribution of laterally shared genes according to the ML reference tree had an exponent of $\hat{\alpha} = 2.31 \pm 0.11$, with an estimated lower bound of $\hat{x}_{\min} = 16$, the distribution for the network using the NJ reference tree gave an exponent of $\hat{\alpha} = 2.11 \pm 0.17$, with an estimated lower bound of $\hat{x}_{\min} = 6$, calculated as described in ref. 38. Although a Kolmogorov–Smirnov test (38) rejected the hypothesis that the distributions of edge weights (number of genes shared between each pair of genomes) are strictly power law, a moving-tail test showed that there is a higher likelihood that these distributions follow a power law rather than an exponential. In this moving-tail test, both probabilistic models are confronted with different subsets of the data, giving Akaike information criterion (AIC) weights that determine the likelihood of the data fitting either distribution. Figures were plotted by using Matlab.

The clustering coefficient (CC) is defined as the probability that two genomes laterally sharing genes with a third genome will also laterally share genes with each other (36). To test the significance of the high CC found in the binary network of laterally shared genes (that is, a network in which a link exists if two genomes laterally share at least one gene), we generated a random ensemble of 10,000 networks by switching the pairs of links between genomes, thus conserving the degree of connectivity of each genome. The samples were created sequentially, separated by 1,000 such switches, and the Add Method (37) was used to fix any potential biases that could arise from nonuniform sampling.

**ACKNOWLEDGMENTS.** We thank E. Bapteste, J. O. McInerney, M. Lercher, and L. Stone for discussions and F. Bartumeus for advice on the moving-tail test. This work was supported by the Deutsche Forschungsgemeinschaft (W.M.), the German-Israeli Foundation for scientific research and development (T.D.), the Horowitz Center for Complexity Science, and the James S. McDonnell Foundation (Y.A.-R.).

1. Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* 3:711–721.
2. Lang AS, Beatty JT (2007) Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol* 15:54–62.
3. Moran NA (2007) Symbiosis as an adaptive process and source of phenotypic complexity. *Proc Natl Acad Sci USA* 104:8627–8633.
4. Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science* 284:2124–2128.
5. Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* 19:2226–2238.
6. Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE (2002) Whole-genome analysis of photosynthetic prokaryotes. *Science* 298:1616–1620.
7. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39:309–338.
8. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA (2005) The net of life: Reconstructing the microbial phylogenetic network. *Genome Res* 15:954–959.
9. Doolittle WF, Bapteste E (2007) Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci USA* 104:2043–2049.
10. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 6:361–375.
11. Ciccarelli FD, *et al.* (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.
12. Nakamura Y, Itoh T, Matsuda H, Gojobori T (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* 36:760–766.
13. Alon U (2007) Network motifs: Theory and experimental approaches. *Nat Rev Genet* 8:450–461.
14. Pal C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 37:1372–1375.
15. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411:41–42.
16. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254–267.
17. Rezende EL, Lavabre JE, Guimaraes PR, Jordano P, Bascompte J (2007) Non-random coextinctions in phylogenetically structured mutualistic networks. *Nature* 448:925–928.
18. Proulx SR, Promislow DE, Phillips PC (2005) Network thinking in ecology and evolution. *Trends Ecol Evol* 20:345–353.
19. Burt RS (1980) Models of network stucture. *Annu Rev Sociol* 6:79–141.
20. Albert R, Jeong H, Barabási AL (1999) Internet diameter of the world-wide web. *Nature* 401:130–131.
21. Guimera R, Nunes Amaral LA (2005) Functional cartography of complex metabolic networks. *Nature* 433:895–900.
22. Palla G, Derenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435:814–818.
23. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys Rev* E *7* 4:036104.
24. Gallos LK, Song C, Havlin S, Makse HA (2007) Scaling theory of transport in complex biological networks. *Proc Natl Acad Sci USA* 104:7746–7751.
25. Comas I, Moya A, Azad RK, Lawrence JG, Gonzalez-Candelas F (2006) The evolutionary origin of Xanthomonadales genomes and the nature of the horizontal gene transfer process. *Mol Biol Evol* 23:2049–2057.
26. Boetius A, *et al.* (2000) A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* 407:623–626.
27. McInerney JO, Cotton JA, Pisani D (2008) The prokaryotic tree of life: Past, present. . . and future? *Trends Ecol Evol* 276:276–281.
28. Beiko RG, Harlow TJ, Ragan MA (2005) Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA* 102:14332–14337.
29. Huang J, Gogarten JP (2007) Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biol* 8:R99.
30. Ueda K, Beppu T (2007) Lessons from studies of *Symbiobacterium thermophilum*, a unique syntrophic bacterium. *Biosci Biotechnol Biochem* 71:1115–1121.
31. Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. *Nat Genet* 21:108–110.
32. Rivera MC, Lake JA (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431:152–155.
33. Boucher Y, *et al.* (2003) Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet* 37:283–328.
34. Doolittle WF, *et al.* (2003) How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Phil Trans R Soc Lond B* 358:39–58.
35. Dagan T, Martin W (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* 104:870–875.
36. Newman MEJ (2003) The structure and function of complex networks. *SIAM Rev* 45:167–256.
37. Artzy-Randrup Y, Stone L (2005) Generating uniformly distributed random networks: The ADD method. *Phys Rev E* 72:056708.35.
38. Clauset A, Shalizi CR, Newman MEJ (2007) Power-law distributions in empirical data. *Physics* 0706.1062 E-print.
39. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.
40. 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.
41. Bapteste E, Boucher Y, Leigh J, Doolittle WF (2004) Phylogenetic reconstruction and lateral gene transfer. *Trends Microbiol* 12:406–411.
42. Hayashi T, *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* 8:11–22.
43. Sorek R, *et al.* (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318:1449–1452.
44. 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.
45. Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 28:33–36.
46. Thompson JD, Higgins DG, Gibson TJ (1994) ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
47. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575–1584.
48. Brohée S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 7:488.
49. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) (Department of Genome Sciences, Univ of Washington, Seattle), version 3.6.