Review

Cell PRESS

Microbial systems biology

Phylogenomic networks

Tal Dagan

Institute of Molecular Evolution, Heinrich-Heine University of Düsseldorf, Universitätstr. 1, Düsseldorf 40225, Germany

Phylogenomics is aimed at studying functional and evolutionary aspects of genome biology using phylogenetic analysis of whole genomes. Current approaches to genome phylogenies are commonly founded in terms of phylogenetic trees. However, several evolutionary processes are non tree-like in nature, including recombination and lateral gene transfer (LGT). Phylogenomic networks are a special type of phylogenetic network reconstructed from fully sequenced genomes. The network model, comprising genomes connected by pairwise evolutionary relations, enables the reconstruction of both vertical and LGT events. Modeling genome evolution in the form of a network enables the use of an extensive toolbox developed for network research. The structural properties of phylogenomic networks open up fundamentally new insights into genome evolution.

Phylogenomics

The evolutionary history of a species is most commonly depicted as a bifurcating phylogenetic tree (see Glossary) comprising nodes and branches. The nodes in the tree correspond to contemporary species (external nodes) and their ancestors (internal nodes). The branches represent vertical inheritance linking ancestors with their descendants (Figure 1a). The accumulation of fully sequenced genomes since the early 2000s has enabled the practice of phylogenomics, that is, the study of phylogenetic relationships at the whole genome level [1]. The evolutionary reconstruction of gene phylogenies from many genomes at once allows a more accurate reconstruction of evolutionary events such as gene loss, gene gain and gene duplication [2] (Figure 1b).

Prokaryotic species evolve not only through vertical inheritance but also by DNA acquisition via lateral gene transfer (LGT) [3]. During an LGT event, a recipient genome acquires genetic material from a donor genome. The acquired DNA becomes an integral part of the recipient genome and is inherited by its descendants [4]. LGT is a major mechanism for natural variation in prokaryotes where several mechanisms for DNA acquisition have evolved, including transformation [5], transduction [6], conjugation [7] and gene transfer agents [8,9]. The frequency of orthologous protein families affected by LGT during microbial evolution as inferred from gene phylogenies is estimated to range between 60 [10,11] and 90% [12]. Other authors reported much lower frequencies ranging between 2 [13] and 14% [14] of the protein families. An experimental assessment of recent LGT frequency revealed that the barriers to gene acquisition in prokaryotes can be rather low [15]. Out of 246 045 LGTs from 79 different donor species via a plasmid (similar to LGT by transformation or conjugation) into *Escherichia coli*, only 1402 instances failed to integrate into the recipient genome. In the remaining 99.4% of the transfers, the gene was transferred successfully [15]. Genes that were identified as resistant to lateral transfer are common among proteins involved in complex biological mechanisms, such as the ribosome, where both sequence conservation and gene copy number confer major selective constraints on protein function [15–17].

The widespread occurrence of LGT means that a tree model that takes only vertical inheritance into account fits only a very small fraction of the bacterial genomic repertoire. The most natural generalization and alternative to trees are networks [18–21].

Networks

A network (or a graph) is a mathematical model of pairwise relations among entities. The entities (vertices or nodes) in the network are linked by edges representing the connections or interactions between these entities. In a coauthorship network, for example, the vertices signify scientists and the edges represent common publications to the scientists that they connect [22]. In an aviation network,

Glossary

Conjugation: the transfer of DNA via proteinaceous cell-to-cell junctions in bacteria.
Degree (or connectivity): the number of edges that connect the node with other nodes.
Directed network: a network where the entities are connected by asymmetric relationships.
Edge (or link): related vertices are connected by an edge.
Gene copy number: the number of copies of a certain gene within the genome.
Gene transfer agents (GTA): phage-like DNA-carriers that are produced by a donor cell during the growth phase and released to the environment (observed in oceanic Alphaproteobacteria).
Network: an abstract representation of a set of entities connected by links representing symmetric or asymmetric relations between the entities.
Orthologous protein family: a set of homologous proteins from various genomes that are related through speciation.
Phylogenetics: the study of evolutionary relationships among biological entities (e.g. species, genes and genomes).
Phylogenetic network: a network of biological entities connected by links representing evolutionary relations.
Phylogenetic tree: a schematic representation of the evolutionary relations of biological entities. In a bifurcating tree, each entity is permitted to have only two descendants.
Protein family: a set of homologous proteins, i.e. proteins of a common origin, found in diverse species.
Transduction: DNA acquisition during the course of phage infection in bacteria.
Transformation: the uptake of raw DNA from the environment into a microbial cell.
Vertex (or node): an individual entity within the network.
Vertices: the plural form of vertex.

Corresponding author: Dagan, T. (tal.dagan@uni-duesseldorf.de).

0966-842X/$ – see front matter © 2011 Elsevier Ltd. All rights reserved. doi:10.1016/j.tim.2011.07.001

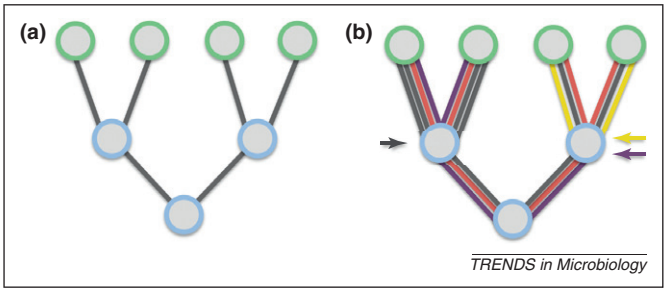

Figure 1. A phylogenetic tree composed of nodes and branches. Contemporary nodes are in green and ancestral nodes in blue. **(a)** A phylogenetic tree of genes or species. The branches represent vertical inheritance. **(b)** A phylogenetic tree of genomes. The multiple lines composing each branch correspond to different genes in the genome. The arrows mark a gene duplication event (gray), a gene loss event (purple) and a gene birth event (yellow).

airports are connected by flights [23]. Network approaches are common in almost all fields of science including social sciences, cell biology, ecology and statistical physics. The network model supplies an abstract representation of whole systems enabling research on the unifying principles behind complex relations among entities. Hence, the most basic issues in networks research are structural [24]. The network properties and connections pattern can teach us about the topology, dynamics and development of the modeled system [24–26].

The information in unweighted networks is limited to whether the vertices are connected or not (Figure 2a). Vertex connectivity (or the degree of a vertex) is the number of vertices connected to the vertex. In a weighted network, the edges can also have a certain weight that signifies the strength of the connection between the vertices. Vertex connectivity in a weighted network is calculated as the total edge weight of edges connected to the vertex [26] (Figure 2a). Edge weight in the coauthorship network is the number of publications coauthored by the two scientists linked by the edge [22]. The connectivity of a scientist in this network is the number of edges connected to it, representing the number of her or his coauthors. A comparison of vertex connectivity in coauthor networks reconstructed for different scientific disciplines reveals stark differences in coauthorship relations depending on the scientific field. For example, the mean number of coauthors per scientist in biomedical studies (18.1 ± 1.3) is much higher than that of physicists (9.7 ± 2) [22].

In directed networks, the edges are oriented from one vertex to another (Figure 2b). Directed networks can be either unweighted or weighted. Vertex connectivity in a directed network is calculated depending on the edge direction. The OUT and IN degrees of any given vertex are defined as the number of edges that are directed from or into the vertex, respectively [26–29] (Figure 2b). For example, in a directed network of phone calls among individuals, the edges signify a phone call between the two individuals that they connect. The edge direction defines the calling individual and the receiving individual [29]. In the phone calls network, the edge weight is the number of phone calls from one individual to another

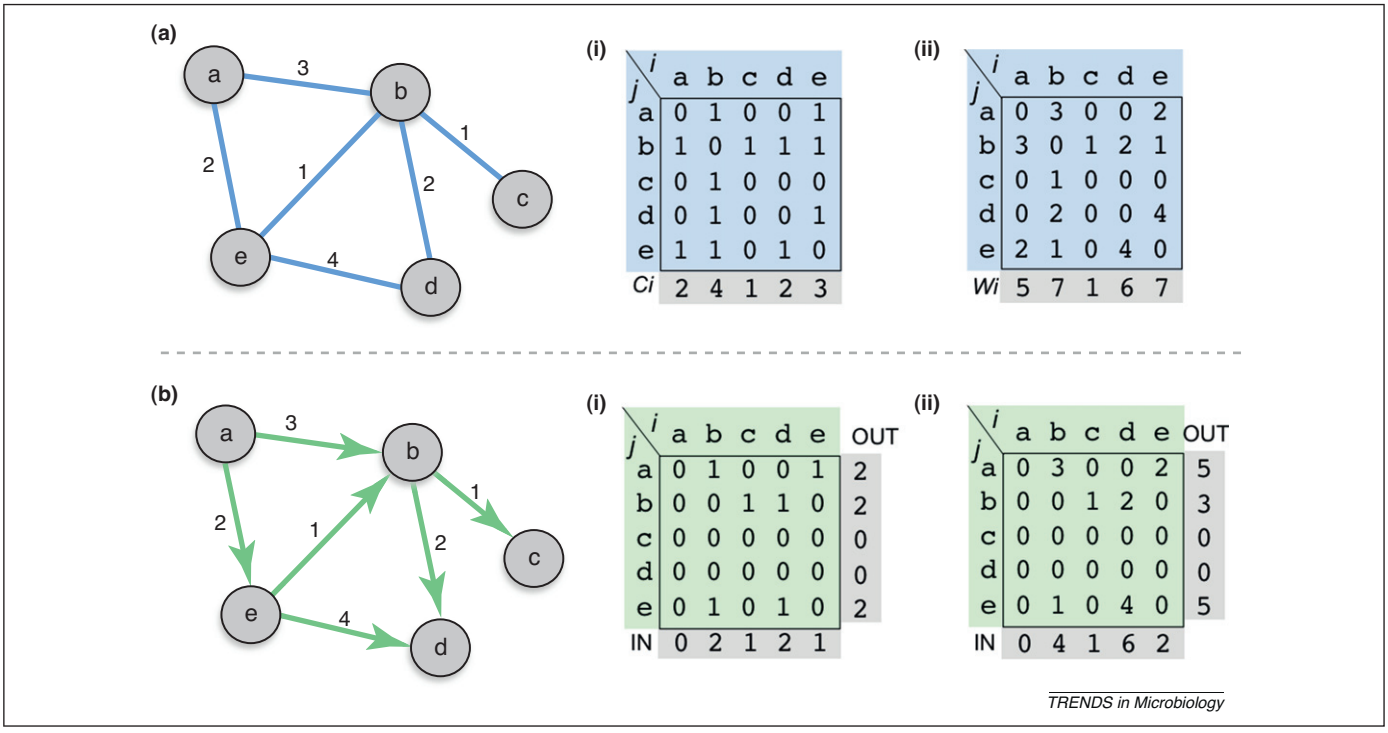

Figure 2. Networks. **(a)** A network composed of vertices (circles) and edges (lines). **(i)** An unweighted network of N vertices can be fully defined by a matrix, $A = [a_{ij}]_{N*N}$, with $a_{ij} = 1$ if an edge is connecting between vertex i and vertex j, and $a_{ji} = 0$ otherwise. Vertex connectivity (C_i) is calculated as the sum of vertices linked to the vertex. **(ii)** A weighted matrix representation of the network. Cells of connected vertices i and j contain the edge weight linking the vertices. Vertex connectivity (W_i) is the sum of edge weights of the edges connected to the vertex. **(b)** A directed network comprising vertices and directed edges. **(i)** In the matrix representation of an unweighted directed network of N vertices, $a_{ij} = 1$ if a directed edge is pointing from vertex i to vertex j, and $a_{ji} = 1$ if a directed edge is pointing from vertex j to vertex i. Vertex IN degree is the sum of vertices connected to the vertex. Vertex OUT degree is the number of vertices to which the vertex is connected. **(ii)** A matrix representation of a weighted directed network. Cells of edges directed from vertex i to vertex j contain the edge weight. Vertex IN degree is the sum of edges connected to the vertex. Vertex OUT degree is the sum of edges connecting the vertex to other vertices.

individual. Vertex OUT and IN degrees correspond to the number of people to whom the individual called and the number of people that called the individual, respectively [29].

Directed networks of biological systems include mainly models of metabolic pathways (e.g. [28,30]) and regulation schemes (e.g. [31–33]). In a directed network of metabolic processes, the vertices represent chemical (metabolite) compounds and the edges represent reactions catalyzed by the corresponding enzyme(s). The edges are directed from the substrate to the product of the enzymatic reaction [28]. Substrate IN and OUT connectivity distribution in metabolic networks is similar among species from the three domains of life, suggesting common principles of metabolic pathway organization within cells [28]. Regulation networks have been used to model different regulatory mechanisms of gene expression. In a transcriptional regulation network, the vertices represent genes and the edges are directed from the regulating gene (i.e. transcription factor) to the regulated gene [31]. The distribution of gene IN and OUT degrees in the transcriptional regulation network of *E. coli* shows that transcription factors regulate the transcription of three genes on average, and that most genes are regulated by one or two transcription factors [31].

Network models are highly efficient as information visualization tools. Modeling complex systems using a networks approach supplies an abstract visual representation of the system [25] enabling our brain (the most powerful of known computers) to look for patterns in the data. Ordering the vertices in the network according to a predefined layout can assist in the search for visual patterns that can then be formulated as hypotheses regarding the modeled system, and be tested statistically. For example, the network of Facebook user connections comprising 500 million people interconnected via the Facebook virtual social network is incomprehensible. However, distributing the vertices in the network according to the geographical coordinates of user address reveals a clear link pattern resembling the globe (http://www.facebook.com/note.php?-note_id=469716398919). The clear geographical structure of human pairwise connections conducted via the World Wide Web suggests that human relations are primarily initiated by a meeting in the real world.

Phylogenetic networks

Networks are commonly used in phylogenetic research for the reconstruction of evolutionary processes that are non-tree-like in nature including hybridization, recombination, genome fusions and LGT [19]. The application of networks to phylogenetic data enables the modeling and visualization of reticulated evolutionary events that cannot be represented using a bifurcating phylogenetic tree [18–21,34–36]. Network applications can also be used for tree-like (vertical inheritance only) gene phylogenies to analyze conflicting phylogenetic signals stemming from either the data or model misspecification [18]. Similar to phylogenetic trees, phylogenetic networks can be reconstructed from various data types including molecular sequences, evolutionary distances, presence/absence data and trees [18,19].

Split networks, for example, are reconstructed from bipartitions of a set of taxa as implied by the underlying data [18,37–39]. The splits are classified as compatible if they correspond to the branching pattern of a phylogenetic tree, and incompatible if they do not [39]. A phylogenetic split network includes both compatible and incompatible splits, hence it can be used to depict and analyze multiple evolutionary scenarios, not only those that are represented by a single phylogenetic tree [18,39]. A phylogenetic reconstruction of a split network from concatenated gene alignments can reveal conflicting phylogenetic signals resulting from hybridization events such as those that occurred during the evolution of the domesticated apple [40] or the origin of the symbiotic hybrid *Euglena gracilis* [41].

The network of shared microbial transposases is an example of a phylogenetic network reconstructed from gene presence/absence data [42]. Transposases, the most abundant enzymes in nature [43], catalyze DNA transposition within and between genomes by a 'cut-and-paste' or 'copy-paste' mechanism [44]. An analysis of sequence divergence patterns among transposases showed that these enzymes are transferred more frequently by LGT than by vertical inheritance [45]. Thus, the distribution of shared transposases among microbial genomes is expected to correlate with LGT. In the microbial transposases network, the vertices are species and the edges correspond to transposase families shared between the genomes that they connect [42]. The shared transposases network reveals that most of the interactions are between closely related species living in the same environment. However, interhabitat connections are also quite common in the network supplying evidence for prokaryotic mobility across habitats, either at present or in the past [42].

Phylogenomic networks are a special type of phylogenetic networks that are reconstructed from the analysis of whole genomes. The vertices in a phylogenomic network correspond to fully sequenced genomes that are linked by edges representing evolutionary relationship reconstructed from whole-genome comparisons. Current applications in the literature include genomes from the three domains of life [46,47] or prokaryotes only [10,14,48–51], as well as genomes of plasmids [47,52,53] and bacteriophages [47,54]. Phylogenomic networks can be divided into two main types: gene-sharing and LGT networks.

Phylogenomic networks of shared genes

Networks of shared genes are reconstructed from the presence/absence pattern of all orthologous protein families distributed across the genomes in the network [11,46–50,54]. The vertices in the network are genomes (species) and the edges correspond to gene sharing between the genomes they connect. The gene sharing network reconstruction procedure includes the following steps: (i) selecting the genomes to be included in the network; (ii) sorting all proteins encoded in the selected genomes into protein families; and (iii) calculating the number of shared genes for each genome pair as the number of protein families in which both genomes are present. Genomes that share at least one protein are connected by an edge. In the simplest form of this network, the edge weight corresponds to the number of shared protein families between the genomes it connects

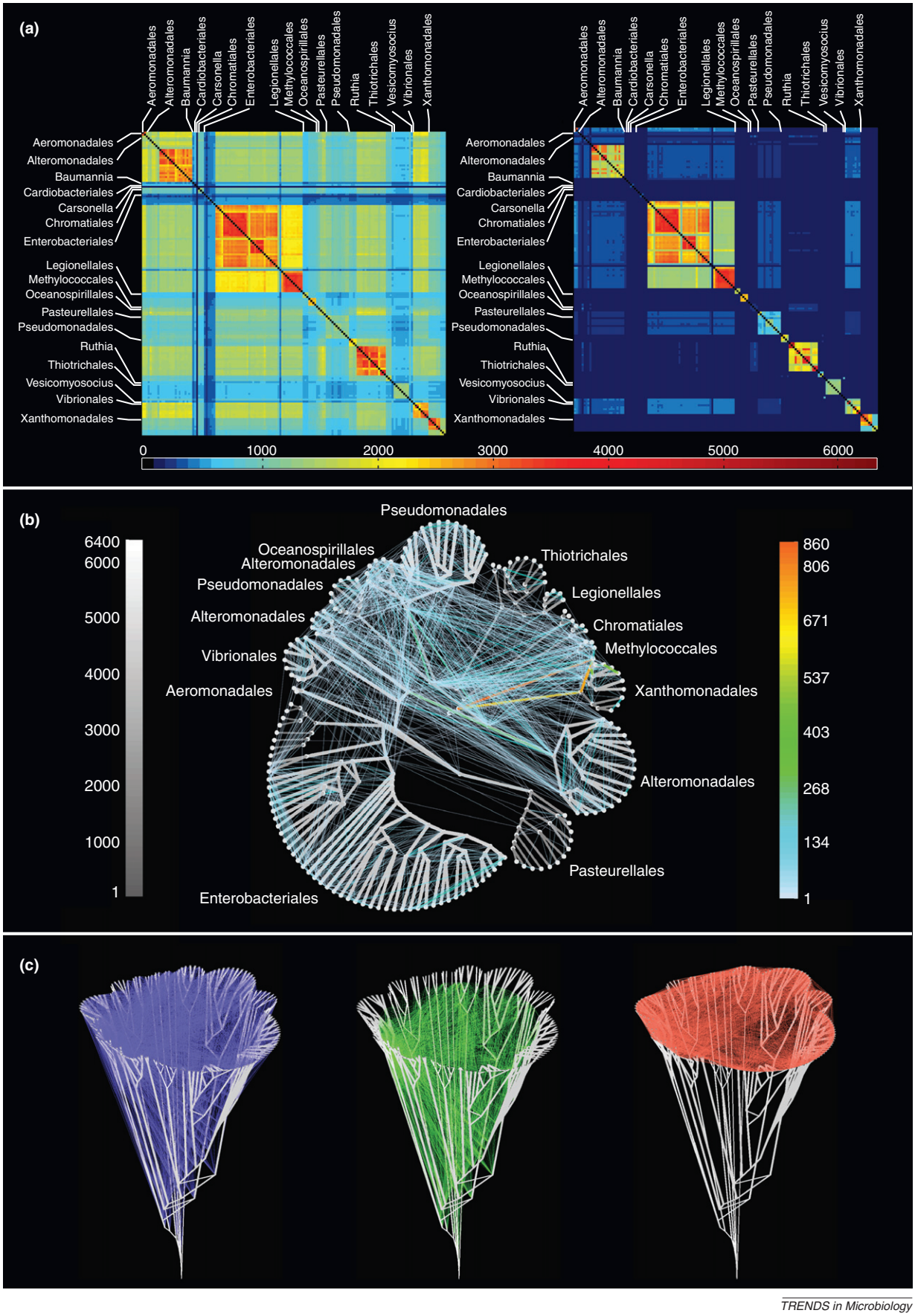

Figure 3. Phylogenomic networks of shared genes reconstructed from 329 gammaproteobacterial genomes. **(a)** A matrix representation of a phylogenomic shared genes network. Protein families were reconstructed under the constraint of 30% (left) and 70% (right) amino acid identities (for details, see [50]). The species are sorted by an alphabetical order of the order and genus. The color scale of cell a_{ij} in the matrix indicates the number of shared protein families between genomes *i* and *j*. The matrix

[47,50] (Figure 3a). Because genome size can vary considerably among species (up to 12-fold in interdomain comparisons), the edge weight in some gene sharing networks is normalized by the genome sizes of the connected vertices [11,46,48,54]. A graphical representation of a gene-sharing network can reveal an internal structure within the network. For example, a network reconstructed from both eukaryotic and prokaryotic genomes reveals a strong phylogenetic structure within the network with a clear distinction between the three domains of life [46]. Phylogenomic shared gene networks of microbial genomes reveal strong connections between closely related species [11,46] (Figure 3a) as well as abundant gene sharing across taxonomic groups that is characteristic of evolution by LGT [11,47,50].

Gene-sharing networks in the literature are typically reconstructed from complete genomes of known taxonomic classification [11,46,50]. Nevertheless, there are also examples for networks comprising genomes of plasmids [52,53] or bacteriophages [54] or even environmental metagenomes [47]. For example, Lima-Mendez *et al*. [54] have looked into the issue of bacteriophage classification using a phylogenomic shared genes network reconstructed from 306 bacteriophage genomes. Similar to prokaryotes, phages also evolve by frequent LGT, making their classification into phylogenetically related groups very difficult [54]. The phylogenomic phage network reveals that clusters of similar genomes in terms of gene sharing comprise phages of various host ranges and nucleic acid types (double- or single-stranded DNA or RNA) [54]. Hence, in this case, the networks approach can contribute to development of a system for phylogenetic classification of phages [54].

Halary *et al*. [47] used a phylogenomic network of shared genes to study the evolution of genetic diversity from a 'DNA centered' point of view. Their network comprises 111 genomes of eukaryotes and prokaryotes, as well as several thousands of phage and plasmid protein sequences; many of the latter were obtained from metagenomic datasets. Using a network of shared genes across the different DNA vehicles (i.e. chromosomes, phages and plasmids) revealed multiple genetic worlds with clear boundaries between the different DNA carriers, with most protein families having a distribution that is limited to a specific type of DNA vehicle. However, the network also contains a large connected component where chromosomes, plasmids and phages are highly interconnected. Frequent links between bacterial chromosomes and plasmids in that component indicate that LGT by conjugation is highly prevalent in natural habitats [47].

Shared gene content among fully sequenced genomes can also be used to reconstruct split networks [49]. Using the extensive set of tools developed for split networks reconstruction [18] enables the analysis and depiction of conflicting phylogenetic signals within gene sharing data. Split networks of shared gene content among prokaryotes can reveal insights into the most ancient splits among microbial genomes [49]. The splits in this type of network are reconstructed from the presence/absence pattern of protein families across fully sequenced microbial genomes. Each protein family defines a partitioning of the genomes into those that encode for that protein and those that do not. Such a split network reconstructed from 22 archaebacterial and 169 eubacterial genomes revealed an ancient divide within microbial life between archaebacteria and eubacteria and an interdomain root position [49].

Phylogenomic LGT networks from shared genes

Phylogenomic LGT networks have been developed to study the lateral component in microbial evolution and are reconstructed from LGT events inferred from genomic data [11,14,48,50,51]. Networks of laterally shared genes (LSG) are a special case of shared genes networks. These are designed specifically to study gene distribution patterns resulting from LGT during prokaryotic evolution. The vertices in the network are the external and internal nodes of a reference species phylogenetic tree. Edges in the network correspond to putative gene transfer events between the nodes they connect [10,48,50] (Figure 3b,c). LGT inference in current applications of LSG networks is based on mapping gene gain and loss events within each protein family onto the reference tree nodes. A gene gain event can be either a gene birth (e.g. by gene duplication, see [55] for review) or a gene acquisition via LGT. The underlying assumption is that gene birth is much rarer than LGT. Hence, in protein families were $N>1$ gain events were inferred, only one of the gains is a gene birth and the remaining $N–1$ gain events are gene acquisitions by LGT. In the LGS network, nodes in the reference tree are connected if there is at least one protein family that is shared between the nodes via a putative LGT event. Edge weight in the LSG network corresponds to the number of laterally shared gene gains between the connected nodes [10,48].

Two different LSG network reconstruction methods are documented in the literature. Gene gain and loss events in the 'net of life' network [10] are inferred by a parsimonious algorithm for ancestral gene content reconstruction. In the minimal lateral network (MLN) approach [48], gene gain and loss events are reconstructed by the ancestral genome size criterion [11]. The application of phylogenomic LSG network including both gene inheritance and gene acquisition by LGT enables an inference of the cumulative impact of LGT during microbial evolution. An MLN

representation of the phylogenomic shared-genes network reconstructed from gammaproteobacterial genomes clearly shows groups of highly connected species having many genes in common. These groups usually comprise closely related species. Examples are 14 *Shewanella* species (Alteromonadales order) at the top left corner, and six *Xanthomonas* species (Xanthomonadales order) at the bottom right corner of the matrix intraconnected species corresponding to (top to bottom) 12 *Escherichia* species, seven *Salmonella* species, six *Shigella* species and 12 *Yersinia* species, which have many genes in common. Applying a higher protein similarity cutoff (right) yields a shared genes network of conserved genes only. The network shows a clear phylogenetic signal with most genes shared among closely related species. **(b)** A phylogenomic network of laterally shared genes reconstructed by the minimal lateral network (MLN) approach [48]. Vertical edges (tree branches) 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 scale on the left). The lateral network is indicated by edges that do not map onto the vertical component, with the number of genes per edge indicated in color (see scale on the right). Edges of weight <10 are excluded [50]. **(c)** A 3D projection of the gammaproteobacterial MLN. Lateral edges are classified into three groups according to the types of vertices they connect within the reference tree. From left to right: (blue) 5083 internal-external edges represent gene sharing between a clade (a group of species) and a contemporary genome; (green) 2191 internal-internal edges correspond to gene sharing between groups of species; (red) 3432 external-external edges correspond to laterally shared genes between contemporary genome.

reconstructed from 181 fully sequenced microbial genomes revealed that, on average, 81 ± 15% of the proteins in each genome are affected by LGT at some time during evolution [48].

Phylogenomic LGT networks from trees

Phylogenomic LGT networks have also been reconstructed from LGT events detected in gene phylogenies [14,51]. As in the LSG network, the phylogenomic LGT network reconstruction requires a species tree that is considered as a reference for distinction between vertical inheritance and LGT. For the network reconstruction, a phylogenetic tree is reconstructed for each protein family. Branches (splits) in the protein family tree that are found in disagreement with the reference species tree are considered as LGT events and are included in the network [14,51].

The LGT network reconstructed by Beiko *et al.* [14] is a summary of all LGT events inferred from 22 432 phylogenies of orthologous protein families encoded in 144 prokaryote genomes. The nodes in the network correspond to 21 higher taxonomic groups of microbes (e.g. Cyanobacteria, Euryarchaeota, Bacilli etc.). Edges in the network correspond to LGT events between members of the groups and are weighted by the number of laterally transferred genes [14]. The network comprises a total edge weight of 1398 LGT events. The heaviest edges in the network connect the vertices of Alphaproteobacteria, Betaproteobacteria and Gammaproteobacteria. The sum of the edge weights linking these three groups corresponds to 56% of the transfers in the network, indicating that LGT is frequent during the evolution of species in these classes [14]. However, the current sampling density of completely sequenced microbial genomes is strongly biased towards Proteobacteria, many of them are human pathogens. Hence, the high frequency of LGTs observed within this phylum could be attributed to their overrepresentation in the data [51 (37%) of the species in the network].

LGT inference methods that include the identification of the donor and recipient in the gene transfer event enable the reconstruction of a directed phylogenomic network. Popa *et al.* [51] described a directed network of LGT (dLGT) comprising 32 027 recent LGT events reconstructed from 657 fully sequenced microbial genomes. The vertices in this network are contemporary and ancestral microbial species (as in the LSG network). Edges in the dLGT network correspond to one or more recent LGT events between the species they connect and are directed from the donor to the recipient. The edge weight is the number of genes that were laterally transferred between the connected genomes [51] (Figure 4). The nodes in the dLGT network are arranged by the density of their connections. Highly connected species, having frequent recent LGTs between them, are placed close together in the graph (Figure 4). Species from the same taxonomic group are colored by the same node color. The resulting network shows that vertices that are close together in the graph often have the same color (e.g. the clusters of Enterobacteriales or Xanthomonadales in Figure 4). Hence, most of the recent LGT events within the dLGT network are among closely related species. Using a dLGT networks approach enables coupling of information regarding LGT events and cellular properties of donors and recipients. The dLGT network reconstructed by Popa *et al.* [51] revealed that DNA repair mechanisms could be involved in DNA integration into the recipient genome during an LGT event, enabling gene acquisition from distantly related donors.

Structural properties of phylogenomic networks

Structural properties of networks can be analyzed and understood using an extensive set of tools developed over the years [24,26]. Node connectivity, for example, is a measure that quantifies the extent to which a node is central within the network [26]. A similar measure, vertex centrality, quantifies the frequency in which the vertex occurs along the shortest path between any vertex pair in the network. The overall distribution of vertex centrality is commonly used to test for internal structure within the network. A distribution that is different from that of a random network indicates that vertices in the network have a preferential attachment resulting from the evolutionary history of the network [27].

Vertex connectivity in phylogenomic LSG networks can serve as a measure for the frequency in which the species donates or acquires genes by LGT. The genomes of the plancomycetes *Rhodopirellula baltica* str. SH1 (*Pirellula* sp.) and the Alphaproteobacteria *Bradyrhizubium japonicum*, for example, are highly connected within the LSG network (hub genomes) [10]. These two species harbor a relatively big proteome, *R. baltica* with 7325 proteins and *B. japonicum* with 8317 proteins. Genome size and the frequency of acquired genes are positively correlated [56], hence species having large genomes are expected to be highly connected in phylogenomic networks of LGT. In the dLGT network, genome size correlates positively with both IN and OUT vertex degree (r_{IN} = 0.38, r_{OUT} = 0.39) indicating that species having large genomes are not only frequent recipients but also frequent donors [51]. In the phylogenomic gene-sharing network among different DNA carriers, plasmids have significantly higher centrality than phages [47]. This result suggests that LGT in nature is more frequently mediated by conjugation than by transduction [47]. Edge weight distribution in weighted networks can also supply information regarding link patterns in the network. The edge weight distribution in the LSG and dLGT networks is linear in a log–log scale indicating that the majority of LGT events are of one or few genes whereas bulk transfers of many genes are rare [10,48,50,51].

Another measure of interest is the diameter of a network, which quantifies the mean shortest path length between any two vertices in the network [26]. In the aviation network, for example, this is the average number of flights that one needs to book in order to travel from any city to any other city in the world [23]. Networks having a small diameter are designated 'small world' networks [24–26,57]. Human society is an example of such a network; the median of distances between any given pair of humans measured by mutual acquaintances is only 5.5 [57]. The diameter of the LSG network measured by the mean shortest path between any genome pair ranges between two and five nodes indicating that they form a small world network [10,48]. This implies that a gene can be

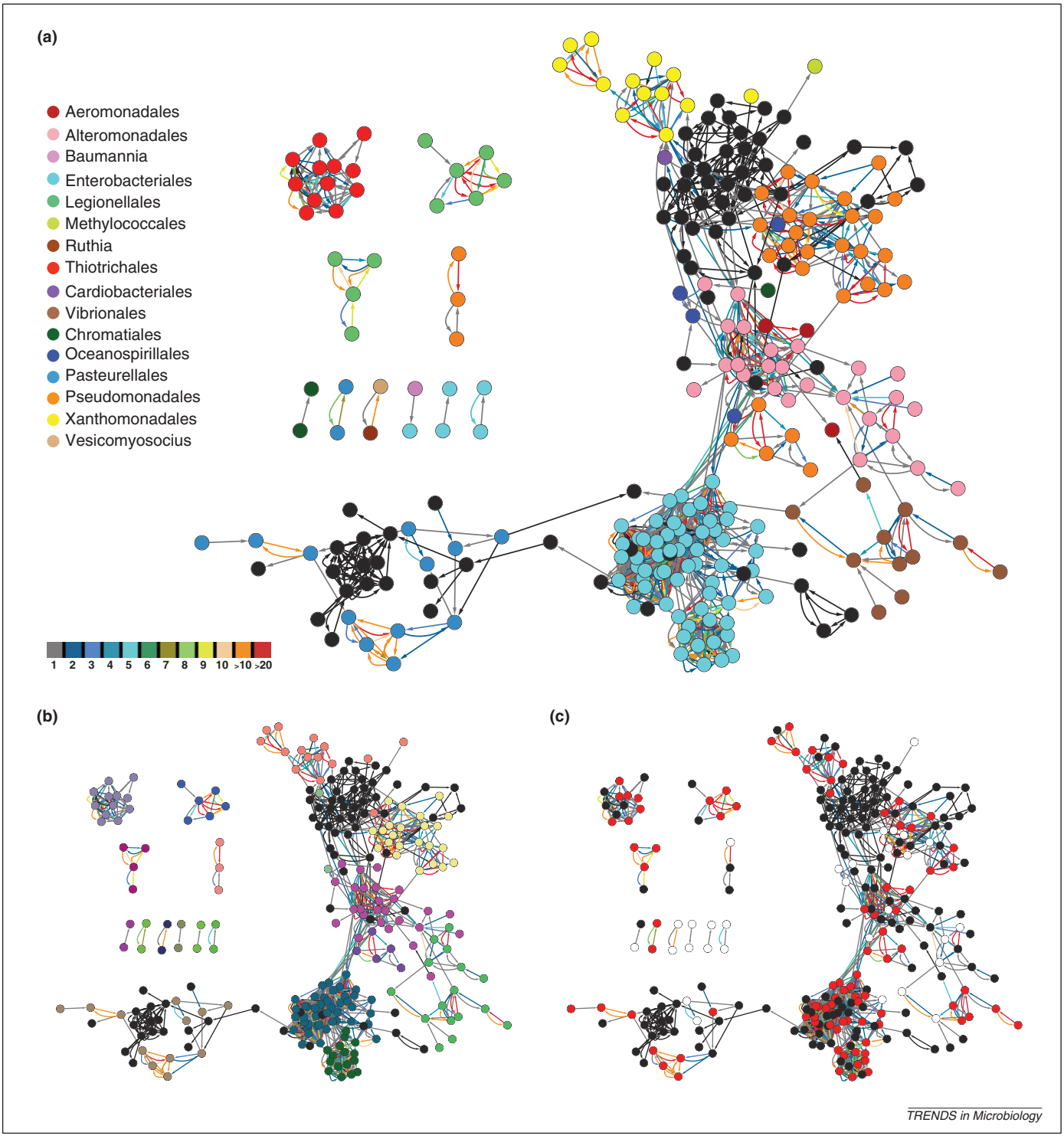

Figure 4. A phylogenomic directed lateral gene transfer (dLGT) network [51]. The nodes represent species and their ancestors. The edges represent LGT events and are directed from the donor to the recipient. Nodes of non-Gammaproteobacteria species are colored in black. Most of these are Betaproteobacteria [51]. **(a)** Node color corresponds to the taxonomic order of donors and recipients listed on the left. The edge color corresponds to the number of transferred genes (see scale at the bottom). Most of the colorful edges connect between nodes having the same color, hence most of the recent LGT in this network occurs between donors and recipients from the same taxonomic group. Genomes of intracellular endosymbionts (e.g. the parasites Legionellales and Thiotrichales) are forming genus-specific clusters that are disconnected from the larger component. The lack of detected recent LGT between those endosymbionts and other species in the network can be due to their interaction with the host, which is a barrier to LGT. **(b)** Community structure within the dLGT network. Node color corresponds to the community to which it belongs. Nodes from the same community are colored in the same shade. Most of the communities comprise closely related species from the same genus. The Enterobacteriales form two communities. The green community includes only *Yersinia* species, the blue community includes *Escherichia, Shigella, Salmonella* and *Citrobacter* species. **(c)** Cellular characteristics in the dLGT network showing the pathogens (red) and non-pathogens (white) in the network. The presence of LGT links between pathogens and non-pathogens suggest that non-pathogens might mediate DNA transfer between pathogenic populations [51].

transferred between any two random species by no more than five LGT events via intermediate recipients/donors. This could be the reason for the rapid percolation of antibiotic resistance genes [58] within pathogenic populations.

Networks can also display community structure [59]. A network that includes groups of vertices that are densely connected within the group but scarcely connected with vertices from other groups is said to have an internal community structure [26,59–61]. Communities are the functional building blocks of the network and could supply information about its evolutionary history [60,61]. An example is the network of protein–protein interactions within the cell. In this network, proteins (vertices) that were found to interact are linked by an edge. The protein–protein interaction network has a significant community structure. Proteins that function in the same cellular process form communities of densely interacting proteins whereas proteins from different cellular processes interact sparsely [59].

The phylogenomic networks of shared genes among prokaryotes have a clear community structure that largely corresponds to the taxonomic classification of the connected species [48]. In Proteobacteria, the community structure within a network comprising 329 genomes reveals a deep split, one that was not detected by common phylogenetic methods, between Alpha-, Delta-, and Epsilon-proteobacteria in one group and Beta- and Gamma-proteobacteria in the other group [50]. Communities in the network of shared genes among DNA carriers are strictly homogeneous with regard to plasmids and phages. This indicates that these two gene vehicles rarely carry the same genes [47]. Community structure within the dLGT network reveals groups of species that are connected by LGT events much more than with species outside the group. Most of the communities in this network comprise species from the same taxonomic group, hence the majority of recent LGT events occur between closely related donors and recipients. The rare communities that group together distantly related species are evidence for frequent LGT within a common habitat or via a common phage [51].

Concluding remarks

Each of the different phylogenomic network types presented here offers a different insight into microbial genome evolution. Networks capture a substantial component of genome evolution, which is not tree-like in nature. Therefore, in biological systems where reticulated evolutionary events are common, phylogenomic networks offer a general computational approach that is more biologically realistic and evolutionarily more accurate. The prevalence of LGT during microbial and viral evolution makes phylogenomic networks an essential tool in the study of these systems.

The networks approach enables the study of several genomic and species characteristics in parallel such as evolutionary relatedness, common habitats, shared gene content and common metabolic pathways. The rapid advance of new sequencing technologies will deliver a genome sample density that was previously unthinkable. It is clear that there is abundant interspecific gene recombination among prokaryotic genomes in nature. Phylogenomic networks will enable the mathematical modeling of evolutionary processes and the investigation of cellular mechanisms that drive microbial genome evolution.

Acknowledgments

I thank Giddy Landan, Liat Shavit-Grievink, Ovidiu Popa, Thorsten Klösges and William Martin for their critical comments on the manuscript. This publication was funded in part by an ERC grant NETWORKORIGINS to William Martin.

References

1 Eisen, J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8, 163–167
2 Eisen, J.A. and Fraser, C.M. (2003) Phylogenomics: intersection of evolution and genomics. *Science* 300, 1706–1707
3 Ochman, H. *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304
4 Babic, A. *et al.* (2008) Direct visualization of horizontal gene transfer. *Science* 319, 1533–1536
5 Chen, I. and Dubnau, D. (2004) DNA uptake during bacterial transformation. *Nat. Rev. Microbiol.* 2, 241–249
6 Thomas, C.M. and Nielsen, K.M. (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* 3, 711–721
7 Chen, I. *et al.* (2005) The ins and outs of DNA transfer in bacteria. *Science* 310, 1456–1460
8 Lang, A.S. and Beatty, J.T. (2007) Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol.* 15, 54–62
9 McDaniel, L.D. *et al.* (2010) High frequency of horizontal gene transfer in the oceans. *Science* 330, 50
10 Kunin, V. *et al.* (2005) The net of life: reconstructing the microbial phylogenetic network. *Genome Res.* 15, 954–959
11 Dagan, T. and Martin, W. (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl. Acad. Sci. U.S.A.* 104, 870–875
12 Mirkin, B.G. *et al.* (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* 3, 2
13 Ge, F. *et al.* (2005) The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol.* 3, e316
14 Beiko, R.G. *et al.* (2005) Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 14332–14337
15 Sorek, R. *et al.* (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318, 1449–1452
16 Jain, R. *et al.* (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 96, 3801–3806
17 Cohen, O. *et al.* (2011) The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* 28, 1481–1489
18 Huson, D.H. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 2, 254–267
19 Huson, D.H. and Scornavacca, C.A. (2011) survey of combinatorial methods for phylogenetic networks. *Genome Biol. Evol.* 3, 23–35
20 Bapteste, E. *et al.* (2009) Prokaryotic evolution and the tree of life are two different things. *Biol. Direct* 4, 34
21 Dagan, T. and Martin, W. (2009) Getting a better picture of microbial evolution en route to a network of genomes. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 364, 2187–2196
22 Newman, M.E. (2001) The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. U.S.A.* 98, 404–409
23 Guimerà, R. *et al.* (2005) The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7794–7799
24 Strogatz, S.H. (2001) Exploring complex networks. *Nature* 410, 268–276
25 Barabási, A.L. (ed.) (2002) *Linked*, Perseus Publishing
26 Newman, M.E.J. (ed.) (2010) *Networks: An Introduction*, Oxford University Press
27 Barabási, A.L. *et al.* (2000) Scale-free characteristics of random networks: the topology of the World-Wide Web. *Phys. A* 281, 69–77
28 Jeong, H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature* 407, 651–654

29 Palla, G. *et al.* (2007) Quantifying social group evolution. *Nature* 446, 664–667
30 Pal, C. *et al.* (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* 37, 1372–1375
31 Thieffry, D. *et al.* (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays* 20, 433–440
32 Shen-Orr, S.S. *et al.* (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31, 64–68
33 Tsang, J.S. *et al.* (2010) Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. *Mol. Cell* 38, 140–153
34 Sneath, P.H.A. (1975) Cladistic representation of reticulate evolution. *Syst. Zool.* 24, 360–368
35 Koonin, E.V. and Wolf, Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* 36, 6688–6719
36 Swithers, K.S. *et al.* (2009) Trees in the web of life. *J. Biol.* 8, 54
37 Bandelt, H-J. and Dress, A.W.M. (1992) A canonical decomposition theory for metrics on a finite set. *Adv. Math.* 92, 47–105
38 Dress, A.W.M. and Huson, D.H. (2004) Constructing splits graphs. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1, 109–115
39 Bryant, D. and Moulton, V. (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* 21, 255–265
40 Velasco, R. *et al.* (2010) The genome of the domesticated apple (*Malus* × *domestica Borkh.*). *Nat. Genet.* 42, 833–839
41 Ahmadinejad, N. *et al.* (2007) Genome history in the symbiotic hybrid *Euglena gracilis*. *Gene* 402, 35–39
42 Hooper, S.D. *et al.* (2009) Microbial co-habitation and lateral gene transfer: what transposases can tell us. *Genome Biol.* 10, R45
43 Aziz, R.K. *et al.* (2010) Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* 38, 4207–4217
44 Curcio, M.J. and Derbyshire, K.M. (2003) The outs and ins of transposition: from mu to kangaroo. *Nat. Rev. Mol. Cell Biol.* 4, 865–877
45 Wagner, A. (2006) Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Mol. Biol. Evol.* 23, 723–733
46 Kunin, V. *et al.* (2005) Measuring genome conservation across taxa: divided strains and united kingdoms. *Nucleic Acids Res.* 33, 616–621
47 Halary, S. *et al.* (2010) Network analyses structure genetic diversity in independent genetic worlds. *Proc. Natl. Acad. Sci. U.S.A.* 107, 127–132
48 Dagan, T. *et al.* (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10039–10044
49 Dagan, T. *et al.* (2010) Genome networks root the tree of life between prokaryotic domains. *Genome Biol. Evol.* 2, 379–392
50 Kloesges, T. *et al.* (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
51 Popa, O. *et al.* (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.* 21, 599–609
52 Fondi, M. and Fani, R. (2010) The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks. *Environ. Microbiol.* 12, 3228–3242
53 Fondi, M. *et al.* (2010) Exploring the evolutionary dynamics of plasmids: the *Acinetobacter* pan-plasmidome. *BMC Evol. Biol.* 10, 59
54 Lima-Mendez, G. *et al.* (2008) Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* 25, 762–777
55 Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20, 1313–1326
56 Nakamura, Y. *et al.* (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet.* 36, 760–766
57 Milgram, S. (1967) The small world problem. *Psychol. Today* 2, 60–67
58 Croucher, N.J. *et al.* (2011) Rapid pneumococcal evolution in response to clinical interventions. *Science* 331, 430–434
59 Girvan, M. and Newman, M.E. (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826
60 Palla, G. *et al.* (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814–818
61 Newman, M.E.J. (2006) The structure and function of complex networks. *SIAM Rev.* 45, 167–256