Phylogenomic Networks

Habilitationsschrift

zur Erlangung der Habilitation im Fach Mikrobiologie an der Mathematisch-Naturwissenschaftlichen Fakultät der Heinrich-Heine-Universität Düsseldorf

von

Tal Dagan

May 2011

Contents

1 Abstract

Phylogenomics is a field of molecular evolutionary research devoted to the study of functional and evolutionary aspects of genomes from the perspective of phylogenetic reconstruction based on whole genomes. Current approaches to genome phylogenies usually operate within the framework of bifurcating phylogenetic trees. However, several evolutionary process are non tree-like in nature, including recombination, hybridization, genome fusions, and lateral/horizontal gene transfer (LGT or HGT). Phylogenetic networks have been therefore developed in order to analyze and depict reticulated evolutionary processes during gene and species evolution. Networks comprise entities (vertices) connected by pairwise relations (edges). Approaching genome evolution with networks, rather than trees, thus enables the reconstruction of both vertical and lateral gene transfer events. Phylogenomic networks are a special type of phylogenetic network reconstructed from fully sequenced genomes. The vertices in phylogenomic networks correspond to genomes that are connected by edges representing evolutionary relations inferred from genomic data. In the literature, phylogenomic networks have mainly been used to study genome evolution in prokaryotes and bacteriophages where lateral gene transfer is a common mechanism of natural variation. Their applications in the literature can be divided into two network types depending on the evolutionary relations to be characterized. In gene-sharing networks, edges represent shared orthologous protein families among the genomes in the network. In LGT networks, genomes are connected by edges representing LGT events reconstructed by other means, for example using phylogenetic trees. Modeling genome evolution using networks offers access to the extensive available toolbox of network research. The structural properties of phylogenomic networks open up fundamentally new insights into genome evolution.

2 Zusamenfassung

Phylogenomik ist ein Zweig der molekularen Evolutionsforschung. Ziel der Phylogenomik ist es, evolutionäre Aspekte in der Genombiologie anhand phylogenetischer Einordnung kompletter Genome zu erfassen. Heutige Ansätze der Genomphylogenie beziehen sich meist auf bifurzierende, phylogenetische Bäume. Viele evolutionäre Prozesse allerdings, wie z. B. die Rekombination, Hybridisierung, Genomfusionen, oder auch lateraler/horizontaler Gentransfer (LGT oder HGT) sind nicht baumartig aufgebaut. Es wurden daher phylogenetische Netzwerke entwickelt, um neben vertikalen auch komplizierter verästelte evolutionäre Prozesse in der Genom- und Speziesevolution zu analysieren und darzustellen. Das Netzwerkmodell – bestehend aus paarweise verbundenen Objekten – ermöglicht die Rekonstruktion sowohl von Vertikalevererbung, als auch von lateralen Gentransferereignissen. Phylogenomische Netzwerke sind eine aus vollständig sequenzierten Genomen berechnete spezielle Klasse phylogenetischer Netzwerke. Die Knoten dieser Graphen entsprechen den Genomen, welche über Kanten, die evolutionären Beziehungen entsprechen, miteinander verbunden sind. In der Literatur werden phylogenetische Netzwerke vorwiegend zur Erforschung der Genomevolution von Prokaryoten und Bakteriophagen verwendet. Lateraler Gentransfer ist bei diesen Organismen ein gängiger Mechanismus zum Austausch genetischen Materials und trägt somit wesentlich zur Artenvielfalt bei. In der Literatur finden zwei Klassen von Netzwerken Verwendung, abhängig von der Art der evolutionären Verbindung. In Netzwerken gemeinsamer Gene repräsentieren die Kanten orthologe Proteinfamilien, die in den verbundenen Genomen gemeinsam vorkommen. Kanten in LGT-Netzwerken repräsentieren durch phylogenetische Bäume abgeleitete LGT Ereignisse. Die Modellierung der Evolution von Genomen in Form eines Netzwerks ermöglicht es, vielseitige Werkzeuge, die für die Erforschung von Netzwerken und Graphen in anderen Wissenschaftsbereichen entwickelt wurden, zu benutzen. Strukturelle Eigenschaften phylogenomischer Netzwerke können neue Erkenntnisse über grundlegende Prozesse und biologische Mechanismen während der Genomevolution liefern, welche vorher verborgen blieben.

3 Introduction

The evolutionary history of species is most commonly depicted as a bifurcating phylogenetic tree 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 1995 has enabled the practice of phylogenomics, that is, the study of phylogenetic relationships at the whole genome level[1,2]. The evolutionary reconstruction of gene phylogenies from many genomes allows a more accurate reconstruction of evolutionary events such as gene loss, gene gain, and gene duplication[1] (Figure 1b).

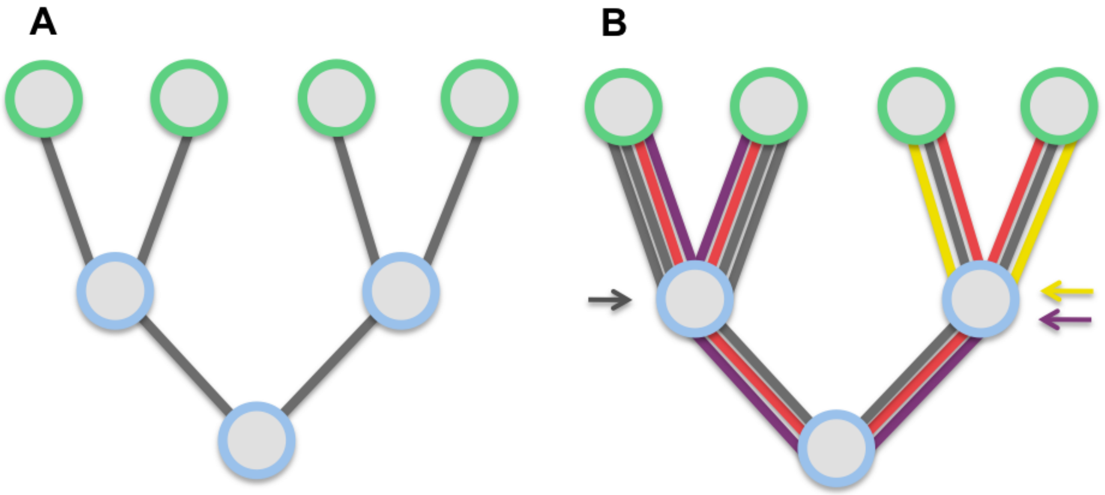


Figure 1 | A phylogenetic tree composed of nodes and branches with contemporary nodes in green and ancestral nodes in blue. (A) A phylogenetic tree of genes or species. The branches represent vertical inheritance. (B) A phylogenetic tree of genomes. The multiple lines composing each branch correspond 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).

But there is more to microbial genome evolution than branching patterns in bifurcating trees. Prokaryotic species evolve not only through vertical inheritance but also by DNA acquisition via lateral gene transfer[3,4]. 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[5]. LGT is a major mechanism for natural variation in prokaryotes where several mechanisms for DNA acquisition have evolved, including transformation[6], transduction[7], conjugation[8], and gene transfer agents[9,10] (Box 1). LGT among eukaryotic species is a rare event except in the case of endosymbiosis where donors and recipients are found in intimate relations making gene transfer feasible and sometimes also beneficial to the host[11,12]. But looking more deeply into the evolutionary past, eukaryotes have a bacterial ancestry; the bioenergetic organelles of eukaryotes – mitochondrion and chloroplast – have evolved through an endosymbiosis event of an eubacterial symbiont[13-17]. The evolution of eukaryotic organelles is characterized by extensive LGT from the engulfed bacterium to the host nuclear genome[18].

Box 1: LGT mechanisms in bacteria

Conjugation: the transfer of DNA via proteinaceous cell-to-cell junction in bacteria.

Gene transfer agents (GTA): phage-like DNA-carriers that are produced by a donor cell under stress conditions and released to the environment (observed in oceanic Alphaproteobacteria).

Transduction: DNA acquisition during the course of phage infection in bacteria.

Transformation: the uptake of raw DNA from the environment into a microbial cell.

While evolution by gene transfer during eukaryotic evolution was abundant in the past but is rare today, LGT during prokaryotic evolution is frequent and abundant today and probably throughout the past as well[3,4]. The frequency of protein families affected by LGT during microbial evolution as inferred from gene phylogenies is estimated to range between 60%[19,20] and 90%[21]. Other authors reported much lower frequencies: Ge et al.[22] estimated that merely 2% of the protein families evolved by LGT, and Beiko et al.[23] detected LGT in only 14% of the protein families. However, these low estimates should be contrasted with the experimental assessment of LGT frequency calculated *in vitro* from gene acquisition rate in *Escherichia coli*[24]. Out of 246,045 LGTs from 79 different donor species via a plasmid (similar to LGT by transformation or conjugation), only 1,402 instances failed to integrate into the *E. coli* genome. In remaining 99.4% of the transfers the gene was transferred successfully[24]. 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[24]. Cellular pathways posing a barrier to LGT of their member proteins are most common in information processing pathways[25] but are found in metabolism and cellular processes pathways as well[26].

The widespread occurrence of LGT means that a tree model that takes only vertical inheritance into account fits only a very small fraction of prokaryotic genome history. The most natural generalization and alternative to trees are networks[27,28]. By graph theory definitions, a phylogenetic tree is a connected, acyclic, directed graph[29]. If we allow the graphs to be cyclic (not only at the root), then we enter the realm of phylogenetic networks[30,31].

3.1 Networks

A network 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 (Box 2). In a co-authorship network, for example, the vertices represent scientists and the edges represent common publications to the scientists that they connect[32]. In an aviation network, airports are connected by flights[33]. 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 a whole system of interacting entities enabling investigation of the unifying principles behind complex relations among them. Hence the most basic issues in network research are structural[34]. Network properties and connection patterns can inform us about the topology, dynamics, and development of the modeled system[34-37].

A network of N vertices can be fully defined by a matrix, $A=[a_{ij}]_{N*N}$, with $a_{ij} \neq 0$ if an edge is connecting between vertex i and vertex j, and $a_{ji} = 0$ otherwise. In a binary network the information is limited to whether the vertices are connected, so that $a_{ij} = 1$ if vertices i, j are connected and $a_{ij} = 0$ if vertices i, j are disconnected. In a weighted matrix the edges can also have a certain weight that signifies the strength of the connection between the vertices. Vertex connectivity is the number of vertices connected to the vertex. In a weighted network the vertex connectivity is calculated as the total edge weight of edges connecting to the vertex[38] (Figure 2a).

Box 2: Network terms

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.

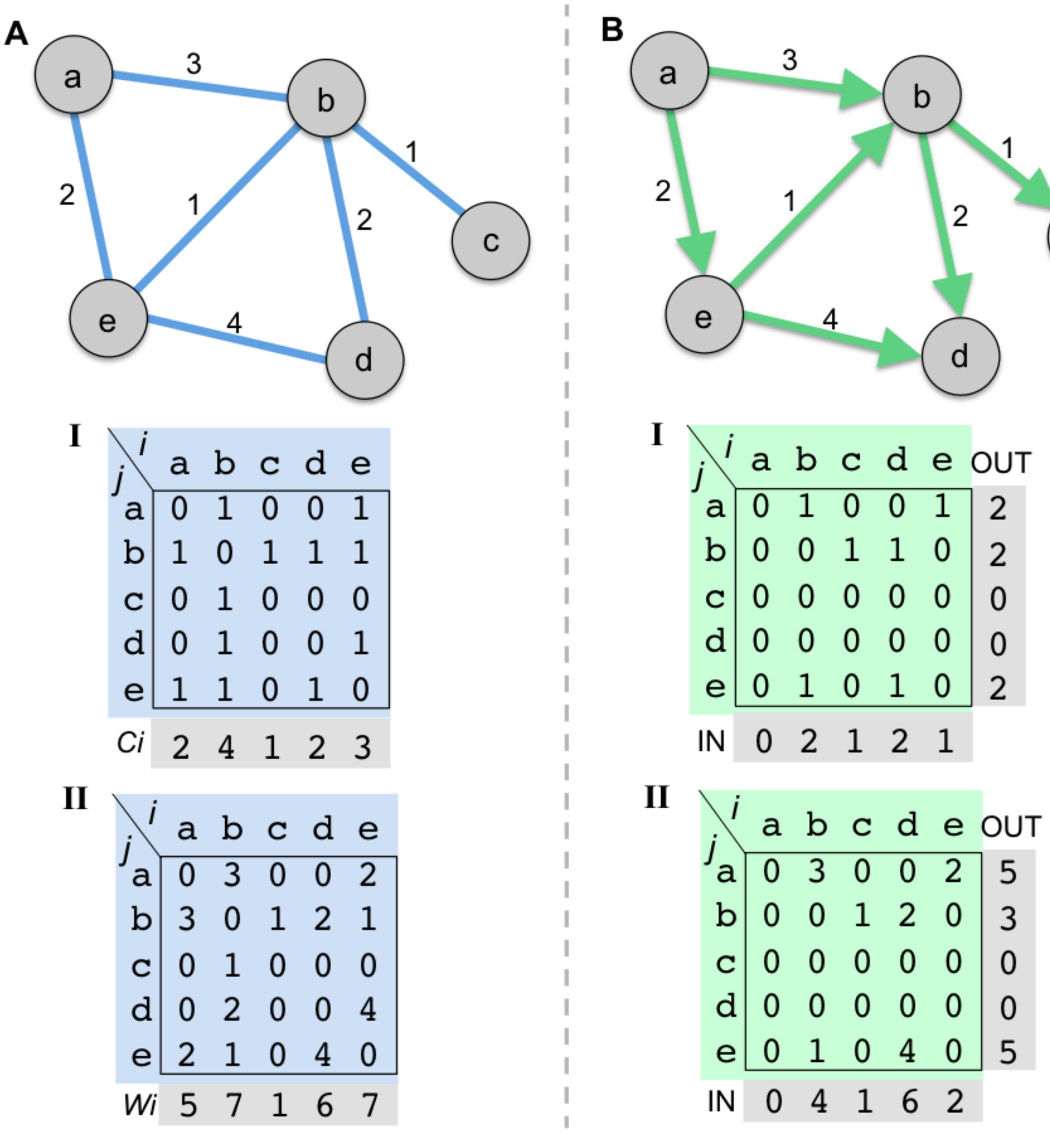
Edge (or Link): related vertices are connected by an edge.

Phylogenetic network: a network of biological entities connected by links representing evolutionary relations.

Network: an abstract representation of a set of entities connected by links representing symmetric or asymmetric relations between the entities.

Vertex (or Node): an individual entity within the network.

For example, in a co-authorship network, scientists are connected by edges signifying common publications[32]. The edge weight is the number of publications coauthored by the two scientists linked by the edge. The connectivity of a scientist in this network is the number of edges connected to it, representing the number of her or his co-authors. The weighted connectivity of a scientist vertex is the total edge weight of all edges linked to it, which is the total number of her or his publications with all co-authors[32]. A comparison of vertex connectivity in co-author networks reconstructed for different scientific disciplines reveals stark differences in co-authorship relations depending on the scientific field.



A I

j \ i	a	b	c	d	e
a	0	1	0	0	1
b	1	0	1	1	1
c	0	1	0	0	0
d	0	1	0	0	1
e	1	1	0	1	0
Ci	2	4	1	2	3

A II

j \ i	a	b	c	d	e
a	0	3	0	0	2
b	3	0	1	2	1
c	0	1	0	0	0
d	0	2	0	0	4
e	2	1	0	4	0
Wi	5	7	1	6	7

B I

j \ i	a	b	c	d	e	OUT
a	0	1	0	0	1	2
b	0	0	1	1	0	2
c	0	0	0	0	0	0
d	0	0	0	0	0	0
e	0	1	0	1	0	2
IN	0	2	1	2	1	

B II

j \ i	a	b	c	d	e	OUT
a	0	3	0	0	2	5
b	0	0	1	2	0	3
c	0	0	0	0	0	0
d	0	0	0	0	0	0
e	0	1	0	4	0	5
IN	0	4	1	6	2	

Figure 2 | Network models. (A) A network composed of vertices (circles) and edges (lines) (top). I) A binary matrix representation of the network. Cells of connected vertices *i* and *j* are set to 1. Vertex connectivity (C_i) is calculated as the sum of edges 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. (B) A directed network comprising vertices and directed edges. I) A binary representation of the directed matrix. Cells of edges directed from vertex *i* to vertex *j* are set to 1. 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 weighted matrix representation of the 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.

For example, the mean co-authors per scientist in the biomedical studies (18.1±1.3) is much higher than that of physicists (9.7±2). Scientists in the field of high-energy physics where large experimental collaborations are common have on average 173±6 co-authors[32].

In directed networks the edges are polarized from one vertex to the other (Figure 2b). In the matrix representation of a directed network of *N* vertices, $a_{ij} \neq 0$ if a directed edge is pointing from vertex *i* to vertex *j*, and $a_{ji} \neq 0$ if a directed edge is pointing from vertex *j* to vertex *i*. Directed networks can be either binary 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[39-43] (Figure 2b). For example, in a directed network of phone calls among individuals the edges signify a phone call between the two individuals they connect, and the edge direction defines the calling individual and the receiving individual[41]. In the phone calls network, the edge weight a_{ij} is the number of phone calls from individual *i* to individual *j*. Vertex OUT and IN degrees correspond to the number people to whom the individual called and the number of people that called the individual respectively[41].

Directed networks of biological systems include mainly models of metabolic pathways[40,44] and regulation schemes[45-47]. In a directed network of metabolic processes the vertices represent chemicals (metabolites) and the edges represent the reactions catalyzed by the corresponding enzyme(s). The edges are directed from the substrate to the product of the enzymatic reaction[40]. 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[40]. 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[45]. 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[45].

Network models are highly efficient as information visualization tools. Modeling complex systems using networks approach supplies an abstract visual representation of the system[37] enabling our brains (the most powerful computers known) 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 graph 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[48]. The clear geographical structure of human pairwise connections conducted via the World Wide Web (WWW) suggests that human relations are primarily initiated by a meeting in the real world[48].

3.2 Phylogenetic networks

Network models are commonly used in phylogenetic research for the reconstruction of evolutionary processes that are non tree-like in nature, including hybridization, recombination, genome fusions, lateral gene transfer, and the like[31]. The application of networks to phylogenetic data permits the modeling and visualization of reticulated evolutionary events that cannot be represented using a bifurcating phylogenetic tree [3,27,28,49-52]. Network applications can be also used for tree-like (vertical inheritance only) gene phylogenies in order to analyze conflicting phylogenetic signals stemming from either the data or model misspecification[30]. Similarly to phylogenetic trees, phylogenetic networks can be reconstructed from various data types including molecular sequences, evolutionary distances, presence/absence data, and trees themselves[30,31].

Split networks, for example, are reconstructed from bipartitions in a set of taxa as implied by the underlying data[31,53-55]. The splits are classified as compatible if they correspond to the branching pattern of a phylogenetic tree, and incompatible if they do not[55]. A phylogenetic splits 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[31,55]. 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[56] or the origin of the symbiotic hybrid *Euglena gracilis*[57].

The network of shared microbial transposases is an example of a phylogenetic network reconstructed from gene presence/absence data[58]. Transposase are the most abundant genes in nature[59]. These enzymes promote DNA transfer between microbial genomes during conjugation[60]. An analysis of transposase sequence divergence patterns showed that these enzymes are transferred more frequently by LGT then by vertical inheritance[61]. Thus the distribution of shared transposases among microbial genomes is expected to correlate with LGT via conjugation. Since this gene transfer mechanism requires a physical contact between the donor and recipient[7], a network of shared transposases is expected to reveal genetic interactions by LGT among species residing in the same habitat. In the microbial transposase network the vertices are species and the edges correspond to transposase families shared between the genomes that they connect[58]. The shared transposase network reveals that most of the interactions are between closely related species living in the same environment. However, inter-habitat connections are also quite common in the network supplying evidence for prokaryotic mobility across habitats, either at present or in the past[58].

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 kingdoms of life[62,63], or prokaryotes only[19,23,64-67] as well as genomes of plasmids[63,68,69] and bacteriophages[63,70]. Phylogenomic networks can be divided into two main types: gene-sharing and lateral gene transfer networks.

3.3 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[20,62-66,70]. 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: (1) select the genomes to be included in the network, (2) sort all proteins encoded in the selected genomes into protein families, and (3) calculate the number of shared genes for each genomes pair by 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[63,66] (Figure 3A). Because genome size can vary considerably among species (up to twelve-fold in inter-kingdom comparisons) the edge weight in some gene sharing networks is normalized by the genome sizes of the connected vertices[20,62,64,70]. 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[62]. Phylogenomic shared-genes networks of microbial genomes reveal strong connections between closely related species[20,62] (e.g. Figure 3A) as well as abundant gene sharing across taxonomic groups that is characteristic of evolution by LGT[20,63,66].

Gene-sharing networks in the literature are typically reconstructed from complete genomes of known taxonomic classification[20,62,66]. Nevertheless, there are also examples for networks comprising genomes of plasmids[68,69] or bacteriophages[70] or even environmental metagenomes[63]. For example, Lima-Mendez et al.[70] have looked into the issue of bacteriophage classification using a phylogenomic shared genes network reconstructed from 306 bacteriophage genomes. Even more so than prokaryotes, phages also evolve by frequent LGT, making their classification into phylogenetically related groups very difficult[70]. 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/RNA)[70]. Hence, in this case the network approach can contribute to a development of a system for phylogenetic classification of phages[70].

Halary et al.[63] 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. The network of genes shared across the different DNA carrier (that is, chromosomes, phages, 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 carrier. 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[63].

Shared gene content among fully sequenced genomes can also be used to reconstruct splits networks[65,71]. Using the extensive set of tools developed for splits network reconstruction[30] enables the analysis and depiction of conflicting phylogenetic signals within gene sharing data. An example is the splits network of protein domain order reconstructed for 167 fully sequenced genomes from the three domains of life[71]. The network reveals clear conflicting phylogenetic signals at the origin of Viridiplantae that are grouped with eukaryotes but share a significant split with cyanobacteria. The plants-cyanobacteria split is a phylogenetic evidence for the cyanoacterial origin of plastids within photosynthetic eukaryotes. Many genes in plant genomes originated by endosymbiotic gene transfer from the genome of the cyanobacterial endosymbiont into the nuclear genome of the host[72]. Consequently, plant genomes are a mosaic of eukaryotic and cyanobacterial genes[18,72,73].

Splits networks of shared gene content among prokaryotes can also reveal insights into the most ancient splits among microbial genomes[65]. 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 reveals a deep split between the two prokaryotic domains[65]. To test for the position of the root in the microbial network of life Dagan et al.[65] used the mid-point rooting approach, according to which the root in a phylogenetic tree is placed at middle of the longest branch[75]. An application of this approach to the shared genes split networks revealed an ancient divide within microbial life between archaebacteria and eubacteria and an inter-domain root position[65].

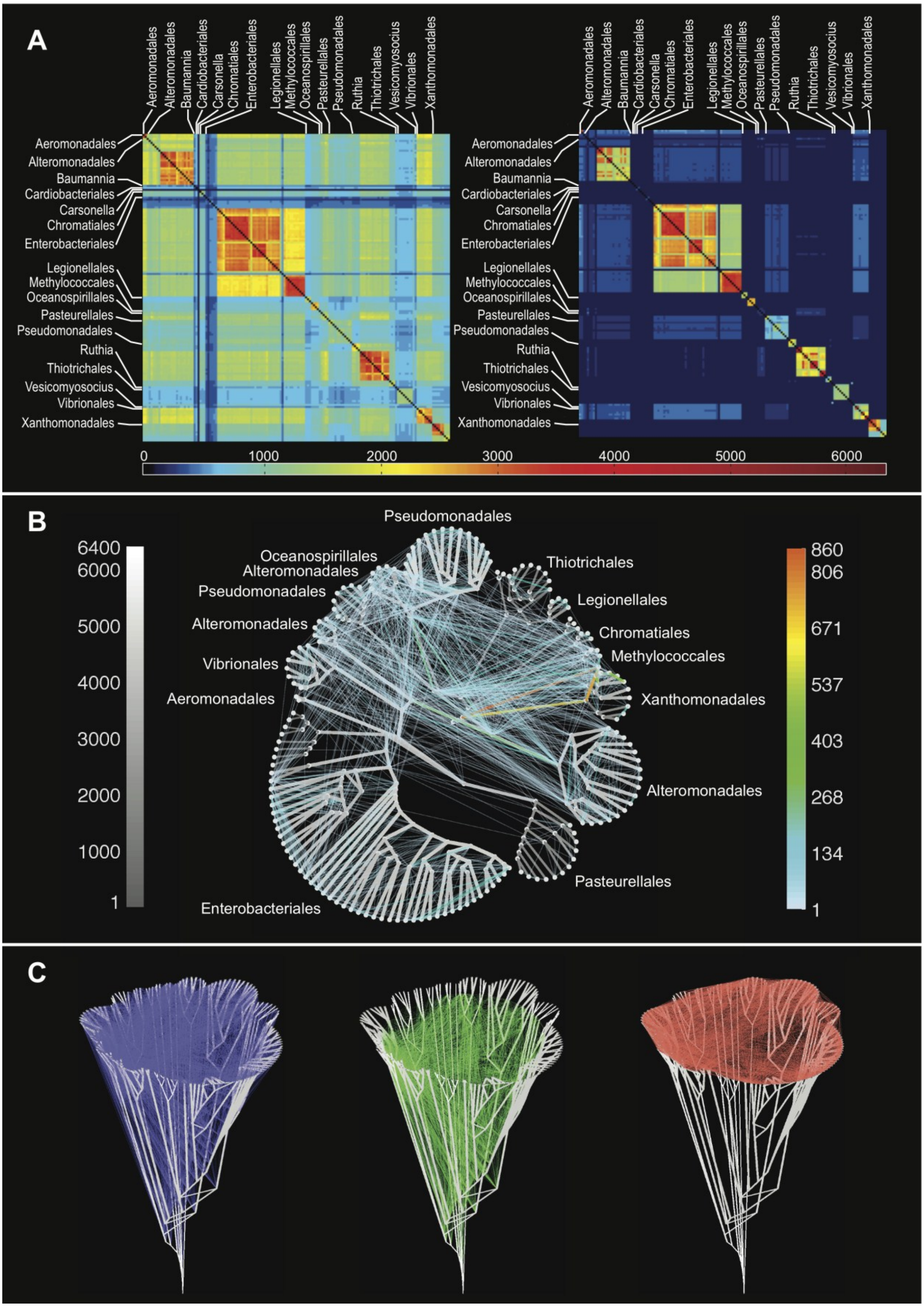

A
Aeromonadales
Alteromonadales
Baumannia
Cardiobacteriales
Carsonella
Chromatiales
Enterobacteriales
Legionellales
Methylococcales
Oceanospirillales
Pasteurellales
Pseudomonadales
Ruthia
Thiotrichales
Vesicomyosocius
Vibrionales
Xanthomonadales
0
1000
2000
3000
4000
5000
6000
B
6400
6000
5000
4000
3000
2000
1000
1
Pseudomonadales
Oceanospirillales
Alteromonadales
Pseudomonadales
Alteromonadales
Vibrionales
Aeromonadales
Thiotrichales
Legionellales
Chromatiales
Methylococcales
Xanthomonadales
Alteromonadales
Pasteurellales
Enterobacteriales
860
806
671
537
403
268
134
1
C

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% (top) and 70% (bottom) amino acid identities (for details see ref. 66). The species are sorted by an alphabetical order of the order and genus. The color scale of cell aij in the matrix indicates the number of shared protein families between genomes *i* and *j*. The matrix 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 intra-connected species corresponding to (top to bottom) 12 *Escherichia* species, 7 *Salmonella* species, 6 *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[64] (left). 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 number of genes per edge indicated in color (see scale on the right). Edges of weight<10 are excluded[66]. (C) A three-dimensional projection of the gammaproteobacterial MLN. Lateral edges are classified into three groups according to the types of vertices they connect within the reference tree. (i) 3,432 external-external edges (red) correspond to laterally shared genes between contemporary genome. (ii) 5,083 internal-external edges (blue) represent gene sharing between a clade (a group of species) and a contemporary genome. (iii) 2,191 internal-internal edges (green) correspond to gene sharing between groups of species.

3.4 Phylogenomic LGT networks from shared genes

Phylogenomic LGT networks are still a young area of endeavor. They were developed to study the lateral component in microbial evolution and are reconstructed from LGT events inferred from genomic data[19,23,64,66,67]. Networks of laterally shared genes (LSG) are a special case of gene sharing networks. They focus on 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 gene transfer events between the nodes they connect[19,64,66] (Figure 2B). 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 or a gene acquisition via LGT. The underlying assumption is that gene birth is much more rare than LGT for genes of related sequence. 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[19,64].

Two different LSG network reconstruction methods are documented in the literature. Gene gain and loss events in the ‘net of life’ network[19] are inferred by a parsimonious algorithm for ancestral gene content reconstruction[76]. In the minimal lateral network (MLN) approach[64] gene gain and loss events are reconstructed by the ancestral genome size criterion[20]. 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 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[64].

3.5 Phylogenomic LGT networks from trees

Phylogenomic LGT networks have been reconstructed from LGT events detected in gene phylogenies as well[23,67]. As in LSG networks, 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[23,67].

The LGT network depicted by Beiko et al.[23] 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[23]. The network comprises a total edge weight of 1,398 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 highly frequent among species in these classes[23].

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.[67] 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[67] (Figure 3C). The nodes in the dLGT network are arranged by the density of their connections.

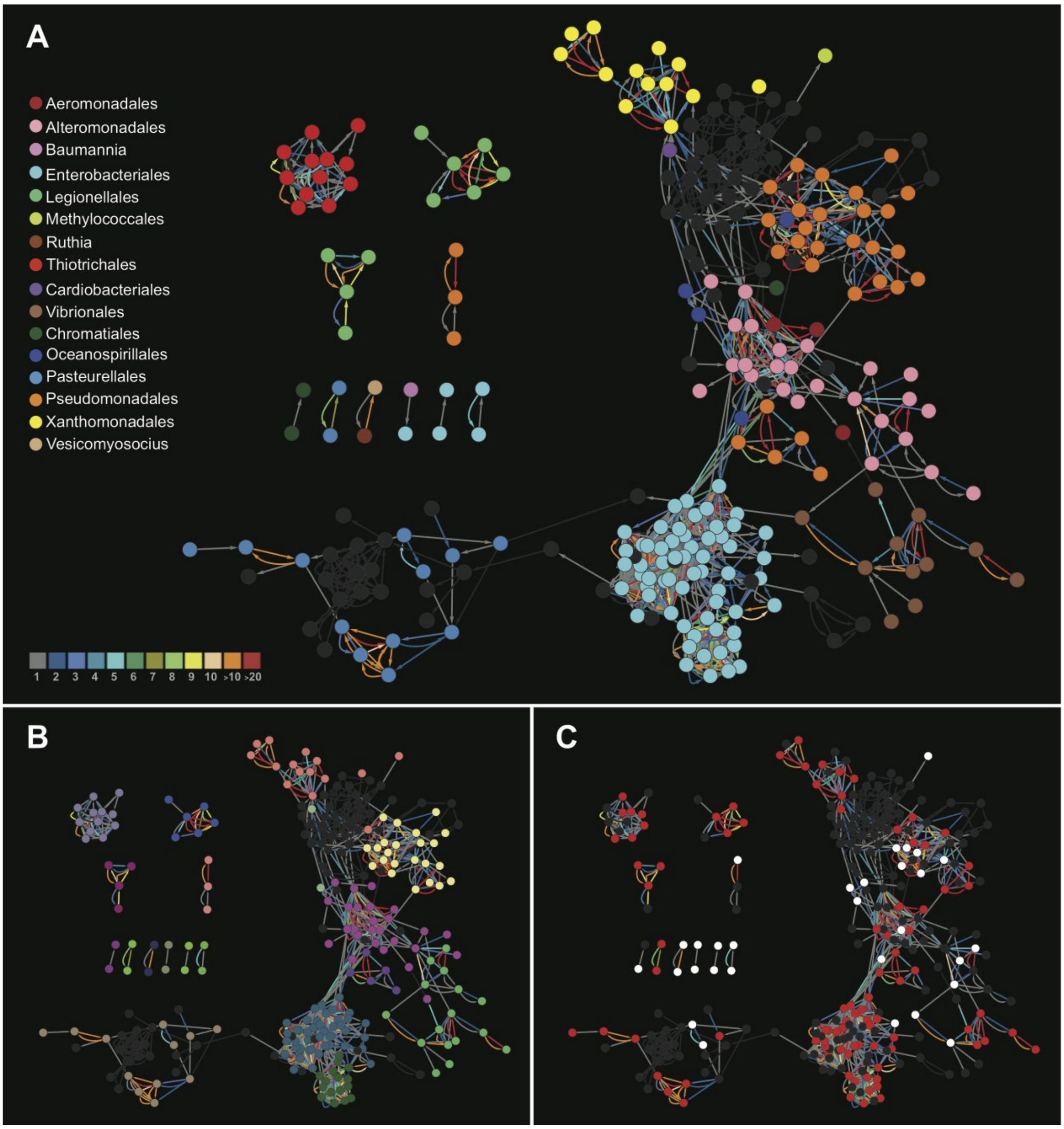


Figure 4 | A phylogenomic directed LGT (dLGT) network. The nodes represent species and their ancestors. The edges represent LGT events and are directed from the donor to the recipient. Nodes of non-Gammaproteobacteria species are colored in gray. Most of these are Betaproteobacteria[67]. (A) Node color 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 may mediate DNA transfer between pathogenic populations[67].

Highly connected species, having frequent recent LGTs between them, are placed close together in the graph (Figure 3C). 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 3C). Hence most of the recent LGT events within the dLGT network are among closely related species. Using an approach based on directed LGT networks enables coupling of information regarding LGT events and cellular properties of donors and recipients. The dLGT network reconstructed by Popa et al.[67] revealed that DNA repair mechanisms such as non-homologous end joining may be involved in DNA integration into the recipient genome during an LGT event, enabling gene acquisition from distantly related donors.

3.6 Structural properties of phylogenomic networks

Structural properties of networks can be analyzed and understood using an extensive set of tools developed over the years[34,37]. Node (or vertex) connectivity, for example, is a measure that quantifies the extent to which a node is central within the network[37,38]. 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[38] .

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 alphaproteobecteria *Bradyrhizubium japonicum*, for example, are highly connected within the LSG network (hub genomes)[19]. These two species harbor a relatively big proteome, *R. baltica* with 7,325 proteins and *B. japonicum* with 8,317 proteins. Genome size and the frequency of acquired genes are positively correlated[77] 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[67]. In the phylogenomic gene-sharing network among different DNA carriers, plasmids have significantly higher centrality than phages[63]. This result suggests that LGT in nature is more frequently mediated by conjugation then by transduction[63]. 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 while bulk transfers of many genes are rare[19,64,66,67].

Another measure of interest is the diameter of a network, which quantifies the mean shortest path length between any two vertices in the network[37]. 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[33]. Networks having a small diameter are designated 'small world' networks[34,35,37,78]. The human society is an example for such a network; the median of distances between any given pair of humans measured by mutual acquaintances is only 5.5 [(ref. 78)]. The diameter of the LSG network measured by the mean shortest path between any genomes pair ranges between 2-5 nodes indicating that they form a small world network[19,64]. This implies that a gene can be 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[79] within pathogenic populations.

Networks may also display community structure[80]. 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[42,80-82]. Communities are the functional building blocks of the network and may supply information about its evolutionary history[81,82]. 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 while proteins from different cellular processes interact sparsely[82].

The phylogenomic networks of shared genes among prokaryotes have a clear community structure that largely corresponds to the taxonomic classification of the connected species[64]. In Proteobacteria, the community structure within a network comprising 329 genomes reveals a deep split, one that was not detected by common phylogentic methods, between alpha-, delta-, and epsilon-proteobacteria in one group and beta- and gamma-proteobacteria in the other group[66]. Communities in the network of shared genes among DNA carriers are strictly homogeneous with regards to plasmids and phages. This indicates that these two gene vehicles rarely carry the same genes[63]. 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 an evidence for frequent LGT within a common habitat or via a common phage[67].

3.7 Summary

Network models 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 which is more realistic biologically and evolutionarily more accurate. The prevalence of LGT during microbial and viral evolution make phylogenomic networks an essential tool in the study these systems.

Each of the different phylogenomic network types presented here offers a different insight into microbial genome evolution. Phylogenomic networks reconstructed from gene sharing are an efficient visualization tool to examine gene distribution patterns across genomes. Community structure within these networks may be helpful for taxonomic classification of bacterial species and bacteriophages[27,70]. Networks of laterally shared genes (LSG) pose an alternative to whole genome phylogenies by supplying a more realistic model of microbial evolution including a distinction between vertical and lateral gene transfer[19,64]. The directed phylogenomic network reconstructed from trees, where both donor and recipient in the recent LGT event are inferred[67], enables a detailed analysis of species characteristics that are related to evolution by LGT.

Much of microbial evolution is better described by networks then by trees, owing to the prevalence of LGT. The network model enables to study several genomic and species characteristics in parallel such as evolutionary relations, 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 enable the mathematical modeling of that evolutionary process and the investigation of cellular mechanisms that drive microbial genome evolution.

3.8 Thematic contents of this thesis

This habilitation, while as a whole dealing with phylogenomic networks, is divided into five complementary sections comprising a total of 13 publications.

The first chapter presents an introduction to non tree-like evolutionary processes. This chapter includes two review articles focusing on the prevalence of reticulated processes during genome evolution (Dagan and Martin 2006) and the need to utilize network approaches in the study of genome evolution (Dagan and Martin 2009).

The second chapter deals with the inference of gene transfer from the organelles to the nuclear genome during eukaryotic evolution. Two research articles are included in this chapter. The first deals with the reconstruction of the mitochondrion ancestor (Esser et al. 2007). The second includes a survey for genes of cyanobacterial origin within plants genomes, the evolutionary history of which suggests that the plastid ancestor was a heterocyst-forming cyanobacterium (Deusch et al. 2008).

The third chapter deals with the inference of lateral gene transfer (LGT) during prokaryote evolution. One article investigates possible biases in LGT inference from phylogenetic trees (Roettger et al. 2009). The other article presents a novel approach to quantify LGT frequency during microbial evolution using the ancestral genome size constraint (Dagan and Martin 2007).

The fourth chapter comprises five applications of phylogenomic networks to different evolutionary questions. Two articles present the utility of a minimal lateral network for estimating the cumulative impact of LGT during microbial evolution (Dagan et al. 2008) and for studying proteobacterial phylogeny (Kloesges et al. 2011). A third article presents a phylogenomic network approach to inferring the root of the tree of life (Dagan et al. 2010). The fourth article presents a novel approach to study microbial genome evolution using directed LGT networks (Popa et al. 2011). A fifth article demonstrates the utility of minimal lateral networks to examine the prevalence of lateral word transfer (borrowing) during language evolution (Nelson-Sathi et al. 2011).

The final chapter deals with the cumulative impact of chaperone-mediated folding on genome evolution. The first article reveals the genomic imprints of chaperone-mediated folding in prokaryotes (Bogumil and Dagan 2010). The second article uncovers common physiochemical properties among proteins that are folded by the same molecular chaperones in yeast (Bogumil et al. 2011).

This introductory chapter is itself currently under consideration for publication as a review article.

4 References

1. Eisen JA: **Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis**. *Genome Res* 1998, **8**:163-167

2. Eisen JA, Fraser CM: **Phylogenomics: intersection of evolution and genomics.** *Science* 2003, **300**:1706-1707.

3. Doolittle WF: **Phylogenetic classification and the universal tree.** *Science* 1999,

4. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405**:299–304.

5. Babic A, Lindner AB, Vulic M, Stewart EJ, Radman M: **Direct visualization of horizontal gene transfer.** *Science* 2008, **319**:1533–1536.

6. Chen I, Dubnau D: **DNA uptake during bacterial transformation.** *Nat Rev Microbiol* 2004, **2**:241-249.

7. Thomas CM, Nielsen KM: **Mechanisms of, and barriers to, horizontal gene transfer between bacteria.** *Nat Rev Microbiol* 2005, **3**:711-721.

8. Chen I, Christie PJ, Dubnau D: **The ins and outs of DNA transfer in bacteria.** *Science* 2005, **310**:1456-1460.

9. Lang AS, Beatty JT: **Importance of widespread gene transfer agent genes in alpha-proteobacteria** *Trends Microbiol* 2007, **15**:54–62.

10. McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH: **High frequency of horizontal gene transfer in the oceans.** *Science* 2010, **330**:50.

11. Nikoh N, Tanaka K, Shibata F, Kondo N, Hizume M, Shimada M, Fukatsu T: ***Wolbachia* genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes.** *Genome Res* 2008, **18**:272-280.

12. Gilbert C, Schaack S, Pace JK 2nd, Brindley PJ, Feschotte C: **A role for host-parasite interactions in the horizontal transfer of transposons across phyla.** *Nature* 2010, **464**:1347-1350.

13. Mereschkowsky C: **Über Natur und Ursprung der Chromatophoren im Pflanzenreiche.** *Biol Centralbl* 1905, **25**:593–604. [English translation by Martin W and Kowallik KV in *Eur J Phycol* 1999, **34**:287–295].

14. Gray MW, Burger G, Lang BF: **Mitochondrial evolution.** *Science* 1999, **283**:1476–1481.

15. Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D, Bryant D, Steel MA, Lockhart PJ, Penny D, Martin W: **A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes.** *Mol Biol Evol* 2004, **21**:1643-1660.

16. Embley TM, Martin W: **Eukaryotic evolution, changes and challenges.** *Nature* 2006, **440**: 623-630.

17. Cotton JA, McInerney JO: **Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function.** *Proc Natl Acad Sci USA* 2010, **107**:17252-17255.

18. Timmis JN, Ayliffe MA, Huang CY, Martin W: **Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes.** *Nat Rev Genet* 2004, **5**:123-135.

19. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA: **The net of life: reconstructing the microbial phylogenetic network.** *Genome Res* 2005, **15**:954-959.

20. Dagan T, Martin W: **Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution.** *Proc Natl Acad Sci USA* 2007, **104**:870-875.

21. Mirkin BG, Fenner TI, Galperin MY, Koonin EV: **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* 2003, **3**:2.

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

23. Beiko RG, Harlow TJ, Ragan MA: **Highways of gene sharing in prokaryotes.** *Proc Natl Acad Sci USA* 2005, **102**:14332-14337.

24. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM: **Genome-wide experimental determination of barriers to horizontal gene transfer.** *Science* 2007, **318**:1449-1452.

25. Jain R, Rivera MC, Lake JA: **Horizontal gene transfer among genomes: the complexity hypothesis.** *Proc Natl Acad Sci USA* 1999, **96**:3801-3806.

26. Cohen O, Gophna U, Pupko T: **The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer.** *Mol Biol Evol* 2011, **28**:1481-1489.

27. Bapteste E, O'Malley MA, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, Lapointe FJ, Dupré J, Dagan T, Boucher Y, Martin W: **Prokaryotic evolution and the tree of life are two different things.** *Biol Direct* 2009, **4**:34.

28. Dagan T, Martin W: **Getting a better picture of microbial evolution en route to a network of genomes.** *Philos Trans R Soc Lond B Biol Sci* 2009, **364**:2187-2196.

29. Harary F: *Graph Theory* Reading, MA: Perseus Books Publishing L. L. C. 1969.

30. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol* 2006, **2**:254-267.

31. Huson DH, Scornavacca C: **A survey of combinatorial methods for phylogenetic networks.** *Genome Biol Evol* 2011, **3**:23-35.

32. Newman ME: **The structure of scientific collaboration networks.** *Proc Natl Acad Sci USA* 2001, 98:404-409.

33. Guimerà R, Mossa S, Turtschi A, Amaral LA: **The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles.** *Proc Natl Acad Sci USA* 2005, 102:7794-7799.

34. Strogatz SH: **Exploring complex networks.** *Nature* 2001, **410**:268-276.

35. Barabási AL: ***Linked***, Cambridge, MA: Perseus Publishing. 2002.

36. Alon U: **Network motifs: theory and experimental approaches.** *Nat Rev Genet* 2007, **8**:450-461.

37. Newman MEJ: ***Networks: An introduction***, Oxford University Press. 2010.

38. Proulx SR, Promislow DE, Phillips PC: **Network thinking in ecology and evolution.** *Trends Ecol Evol* 2005, **20**:345–353.

39. Barabási AL, Albert R, Jeong H: **Scale-free characteristics of random networks: the topology of the World-Wide Web.** *Physica A* 2000, **281**:69-77.

40. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407**:651-654.

41. Palla G, Barabási AL, Vicsek T: **Quantifying social group evolution.** *Nature* 2007, **446**:664-667.

42. Leicht EA, Newman ME: **Community structure in directed networks.** *Phys Rev Lett* 2008, **100**:118703.

43. Foster JG, Foster DV, Grassberger P, Paczuski M. **Edge direction and the structure of networks.** *Proc Natl Acad Sci USA* 2010, **107**:10815-10820.

44. Pal C, Papp B, Lercher MJ: **Adaptive evolution of bacterial metabolic networks by horizontal gene transfer.** *Nat Genet* 2005, **37**:1372–1375.

45. Thieffry D, Huerta AM, Pérez-Rueda E, Collado-Vides J: **From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*.** *Bioessays* 1998, **20**:433-440.

46. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of Escherichia coli.** *Nat Genet* 2002, **31**:64-68.

47. Tsang JS, Ebert MS, van Oudenaarden A: **Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures.** *Mol Cell* 2010, 38:140-153.

48. Butler P: **Visualizing Friendships**. [http://www.facebook.com/note.php?note_id=469716398919]

49. Sneath PHA: **Cladistic Representation of Reticulate Evolution.** *Systematic Zoology* 1975, **24**:360-368.

50. Gogarten JP, Townsend JP: **Horizontal gene transfer, genome innovation and evolution.** *Nat Rev Microbiol* 2005, **3**:679-687.

51. Koonin EV, Wolf YI: **Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world.** *Nucleic Acids Res* 2008, **36**:6688-6719.

52. Swithers KS, Gogarten JP, Fournier GP: **Trees in the web of life.** *J Biol* 2009, **8**:54.

53. Bandelt H-J, Dress AWM: **A canonical decomposition theory for metrics on a finite set.** *Adv Math* 1992, **92**:47–105.

54. Dress AWM, Huson DH: **Constructing splits graphs.** *IEEE/ACM Trans Comput Biol Bioinform* 2004, **1**:109–115.

55. Bryant D, Moulton V: **Neighbor-net: an agglomerative method for the construction of phylogenetic networks.** *Mol Biol Evol* 2004, **21**:255–265.

56. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, *et al.*: **The genome of the domesticated apple (*Malus* × *domestica* Borkh.).** *Nat Genet* 2010, **42**:833-839.

57. Ahmadinejad N, Dagan T, Martin W: **Genome history in the symbiotic hybrid *Euglena gracilis*.** *Gene* 2007, **402**:35-39.

58. Hooper SD, Mavromatis K, Kyrpides NC: **Microbial co-habitation and lateral gene transfer: what transposases can tell us.** *Genome Biol* 2009, **10**:R45.

59. Aziz RK, Breitbart M, Edwards RA: **Transposases are the most abundant, most ubiquitous genes in nature.** *Nucleic Acids Res* 2010, **38**:4207-4217.

60. Curcio MJ, Derbyshire KM: **The outs and ins of transposition: from mu to kangaroo.** *Nat Rev Mol Cell Biol* 2003, **4**:865-877.

61. Wagner A: **Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes.** *Mol Biol Evol* 2006, **23**:723-733

62. Kunin V, Ahren D, Goldovsky L, Janssen P, Ouzounis CA: **Measuring genome conservation across taxa: divided strains and united kingdoms.** *Nucleic Acids Res* 2005, **33**:616-621.

63. Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E: **Network analyses structure genetic diversity in independent genetic worlds.** *Proc Natl Acad Sci USA* 2010, **107**:127-132.

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

65. Dagan T, Roettger M, Bryant D, Martin W: **Genome networks root the tree of life between prokaryotic domains.** *Genome Biol Evol* 2010, **2**:379-392.

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

67. Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T: **Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes.** *Genome Res* 2011, **21**:599-609.

68. Fondi M, Fani R: **The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks.** *Environ Microbiol* 2010, **12**:3228-3242.

69. Fondi M, Bacci G, Brilli M, Papaleo MC, Mengoni A, Vaneechoutte M, Dijkshoorn L, Fani R: **Exploring the evolutionary dynamics of plasmids: the *Acinetobacter* pan-plasmidome.** *BMC Evol Biol* 2010, **10**:59.

70. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R: **Reticulate representation of evolutionary and functional relationships between phage genomes.** *Mol Biol Evol* 2008, **25**:762-777.

71. Fukami-Kobayashi K, Minezaki Y, Tateno Y, Nishikawa K: **A tree of life based on protein domain organizations.** *Mol Biol Evol* 2007, **24**:1181-1189.

72. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D: **Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus.** *Proc Natl Acad Sci USA* 2002, **99**:12246–12251.

73. Archibald JM: **Algal genomes: exploring the imprint of endosymbiosis.** *Curr Biol* 2002, **16**:R1033–R1035.

74. Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D: **Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions.** *Curr Biol* 2006, **16**:2320–2325.

75. Farris JS: **Estimating phylogenetic trees from distance matrices.** *Am Nat* 1972, **106**:645–668.

76. Kunin V, Ouzounis CA: **GeneTRACE-reconstruction of gene content of ancestral species.** *Bioinformatics* 2003, **19**:1412-1416.

77. Nakamura Y, Itoh T, Matsuda H, Gojobori T: **Biased biological functions of horizontally transferred genes in prokaryotic genomes.** *Nat Genet* 2004, **36**:760-766.

78. Milgram S: **The small world problem.** *Psychol Today* 1967, **2**:60–67.

79. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, *et al.*: **Rapid pneumococcal evolution in response to clinical interventions.** *Science* 2011, **331**:430-434.

80. Girvan M, Newman ME: **Community structure in social and biological networks.** *Proc Natl Acad Sci USA* 2002, **99**:7821-7826.

81. Newman MEJ: **The structure and function of complex networks.** *SIAM Rev* 2003, **45**:167–256.

82. Palla G, Derenyi I, Farkas I, Vicsek T: **Uncovering the overlapping community structure of complex networks in nature and society.** *Nature* 2005, **435**:814-818.

5 An introduction to non tree-like evolutionary processes.

5.1 The tree of one percent

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

(Own contribution: 50%).

Opinion

The tree of one percent

Tal Dagan and William Martin

Address: Institute of Botany, University of Düsseldorf, D-40225 Düsseldorf, Germany.

Correspondence: Tal Dagan. Email: tal.dagan@uni-duesseldorf.de

Published: 1 November 2006

Genome **Biology** 2006, **7:**118 (doi:10.1186/gb-2006-7-10-118)

The electronic version of this article is the complete one and can be
found online at http://genomebiology.com/2005/7/10/118

comment
reviews
reports
deposited research
refereed research
interactions
information

Abstract

Two significant evolutionary processes are fundamentally not tree-like in nature - lateral gene transfer among prokaryotes and endosymbiotic gene transfer (from organelles) among eukaryotes. To incorporate such processes into the bigger picture of early evolution, biologists need to depart from the preconceived notion that all genomes are related by a single bifurcating tree.

Evolutionary biologists like to think in terms of trees. Since Darwin, biologists have envisaged phylogeny as a tree-like process of lineage splittings. But Darwin was not concerned with the evolution of microbes, where lateral gene transfer (LGT; a distinctly non-treelike process) is an important mechanism of natural variation, as prokaryotic genome sequences attest [1-4]. Evolutionary biologists are not debating whether LGT exists. But they are debating - and heatedly so - how much LGT actually goes on in evolution. Recent estimates of the proportion of prokaryotic genes that have been affected by LGT differ 30-fold, ranging from 2% [5] to 60% [6]. Biologists are also hotly debating how LGT should influence our approach to understanding genome evolution on the one hand, and our approach to the natural classification of all living things on the other. These debates erupt most acutely over the concept of a tree of life. Here we consider how LGT and endosymbiosis bear on contemporary views of microbial evolution, most of which stem from the days before genome sequences were available.

A tree of life?

When it comes to the concept of a tree of life, there are currently two main camps. One camp, which we shall call the positivists, says that there is a tree of life, that microbial genomes are, in the main, related by a series of bifurcations, and that when we have sifted out a presumably small amount of annoying chaff (LGT), the wheat (the tree) will be there and will still our hunger for a grand and natural system [7-10]. The other camp, which we will call the microbialists, says that LGT is just as natural among prokaryotes as is point mutation, and that furthermore, it has occurred throughout microbial history. This means that even were we to agree on a grand natural classification, the process of microbial evolution underlying it would be fundamentally undepictable as a single bifurcating tree, because a substantial component of the evolutionary process - LGT - is not tree-like to begin with [1,11,12].

A recent paper by Ciccarelli *et al.* [9] brings these two views head-to-head. It purports to weigh in heavily for the positivists, but in doing so it inadvertently provides some of the strongest support for the microbialist camp that has been published so far. A closer look reveals why. Ciccarelli *et al.* [9] report an automated procedure for identifying protein families that are universally distributed among all genomes, with pipeline alignment and tree building. Their routine looked for possible cases of LGT (detected as unusual tree topologies), excluded such proteins, and reiterated the procedure until the universe of proteins had been examined. This left them with 31 presumably orthologous protein sequences present in 191 genomes each, the alignments of which were concatenated to produce a data matrix with 8,089 sites (of which only 1,212 would have remained had gapped sites been excluded). A maximum likelihood tree was inferred from this matrix, motivating a brief discussion of some important events in life's history as inferred from that tree.

Fair enough, one might say, what is there to debate? Lots. Bearing in mind that an average prokaryotic proteome represents about 3,000 protein-coding genes, the 31-protein tree of life represents only about 1% of an average prokaryotic proteome and only 0.1% of a large eukaryotic proteome. Thus, the positivists can say that there is a tree of life after all: a bit skimpier than expected, but a tree nonetheless. But the microbialists, glaring at the same data, can say that the glass is only 1% full at best, and more than 99% empty! There might be a tree there, but it is not the tree of life, it is the 'tree of one percent of life'.

Looking at the issue openly, the finding that, on average, only 0.1% to 1% of each genome fits the metaphor of a tree of life overwhelmingly supports the central pillar of the microbialist argument that a single bifurcating tree is an insufficient model to describe the microbial evolutionary process. If throwing out all non-universally distributed genes and all suspected cases of LGT in our search for the tree of life leaves us with a tree of one percent, then we should probably abandon the tree as a working hypothesis. When chemists or physicists find that a given null hypothesis can account for only 1% of their data, they immediately start searching for a better hypothesis. Not so with microbial evolution, it seems, which is rather worrying. Could it be that many biologists have their heart set on finding a tree of life, regardless of what the data actually say?

Which hypotheses (if any) are we testing?

By themselves, genomes cannot tell us anything about evolution, microbial or otherwise. Evolutionary biology is about hypothesis testing: one checks to see if data from genomes provide support or not for one or the other hypothesis that was generated independently of the genome data used to test it. What ideas about early evolution that could be tested with genome data are currently discussed by specialists in the field? We consider five distinctly different views, each of which enjoys some popularity.

The rRNA tree

The first is the classical ribosomal RNA (rRNA) tree of life as constructed by Carl Woese and colleagues [13-16] from the late 1970s onwards (Figure 1a). It suggests, in its current interpretations, that the universal ancestor of all life (the progenote) was a communal collection of information-storing and information-processing entities that were not yet organized as cells. LGT is seen as the main mode of genetic novelty at the early stages of evolution, and the process of vertical inheritance arises only with the process of 'genetic annealing' from within this mixture. At this point, the emerging cellular lineages of prokaryotes and eukaryotes become refractory to LGT, and are considered to traverse a kind of 'Darwinian threshold' from the organizational state of supramolecular aggregates to the organizational state of cells. Traversing that threshold is seen as equivalent to the primary emergence, from the broth in which life arose, of the three kinds of cells that we recognize today - archaebacteria, eubacteria and eukaryotes. The classical tree [13] assumed its current shape when anciently diverged protein-coding genes suggested that the root of the universal tree lies on the bacterial branch [17,18]. This view admits that chloroplasts and mitochondria did arise via endosymbiosis, but it sees no role for mitochondria or any other kind of symbiosis in the emergence of the eukaryotic lineage, and the genetic contribution of mitochondria to eukaryotes is seen as detectable, but negligible in evolutionary or mechanistic terms [19]. The classical tree is taken by some to indicate that eukaryotes are in fact sisters of archaea at the level of the whole genome [9,16], a view that is, however, mainly founded on extrapolation from the rRNA tree to the rest of the genome without actually looking at all of the data.

The introns-early tree

The introns-early (or eukaryotes-first) tree emerged when Ford Doolittle [20] suggested that the ancestral state of genes might be 'split', and that some introns in eukaryotic genes might thus be carryovers from the assembly of primordial protein-coding regions. In that case, the organizational state of eukaryotic genes (having introns) would represent the organizational state of the very first genomes [21] and the intronless prokaryotic state would be a derived condition (Figure 1b), a view that was christened 'introns-early' [22]. Doolittle has since abandoned this view [23], but it has found other proponents [24,25]. They draw upon different lines of evidence in support, and call their position 'introns-first' rather than introns-early [25]. They agree that the eubacterial root assumed for the rRNA tree is questionable and that a eukaryote root is more likely [26,27].

Some of the proponents of the introns-first hypothesis interpret various aspects of RNA processing in eukaryotes (in addition to introns), such as rRNA modification through small nucleolar RNAs (snoRNAs), as direct carryovers from the RNA world and hence as evidence for eukaryote antiquity [26,28,29]. There is no prokaryote-to-eukaryote transition in the introns-early tree, because prokaryotic genome organization is seen as a very early derivative of eukaryotic gene organization. Accordingly, the relationship of eukaryotes and prokaryotes is depicted largely as a more-or-less unresolved trichotomy [19], and the contribution of organelles or symbiosis to eukaryote evolution is admitted as existing, but negligible in terms of evolutionary significance.

The neomuran tree

The neomuran tree (Figure 1c) stems from the work of Tom Cavalier-Smith [30-32]. No theory on the relationship of prokaryotes to eukaryotes, current or otherwise, is more explicit in terms of details of mechanism [32]. In the main, it suggests that the common ancestor of all cells was a free-living eubacterium (in the most recent version of the theory, a *Chlorobium*-like anoxygenic photosynthesizer) and that

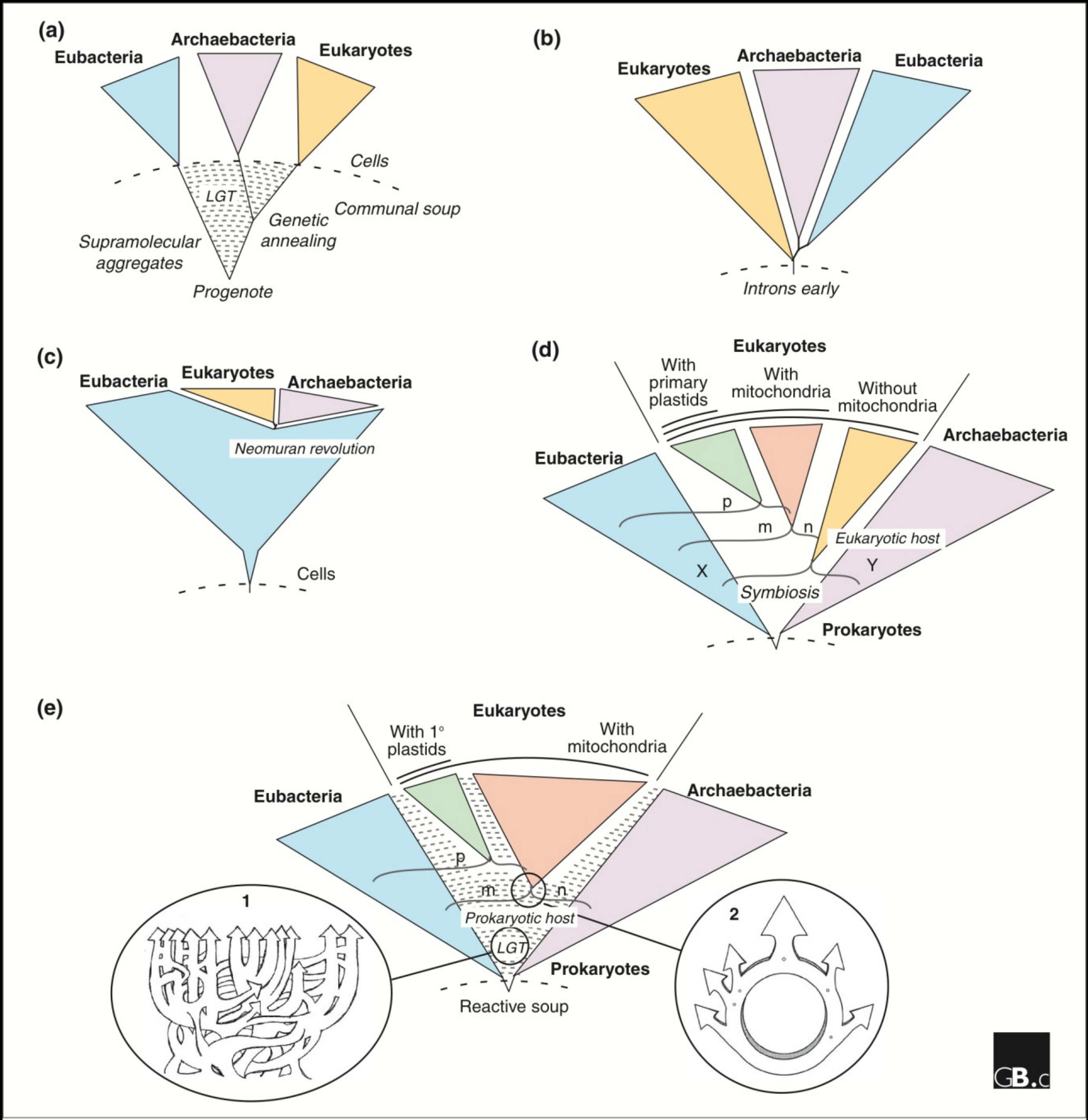
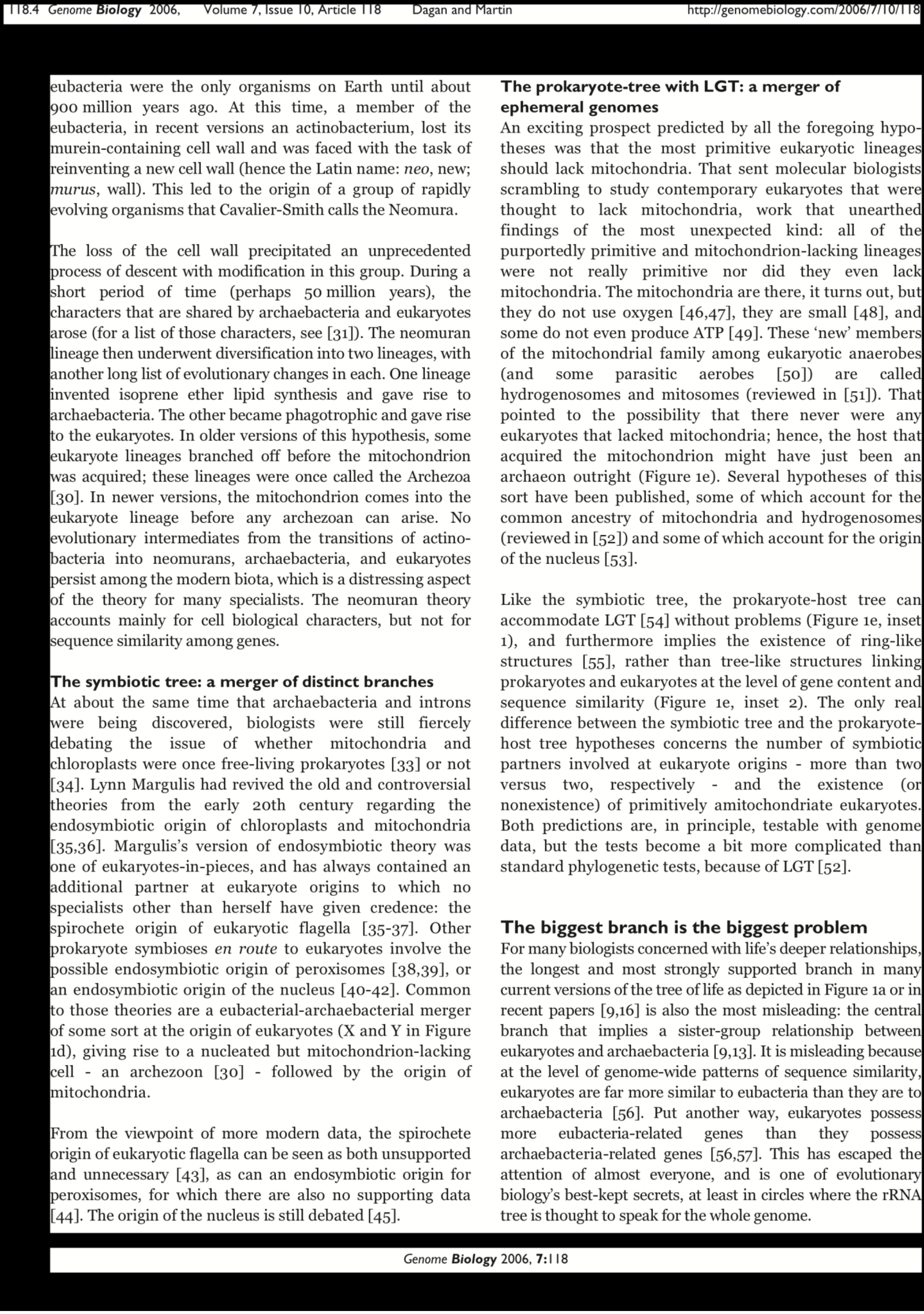


Figure 1
Five different current views of the general shape of microbial evolution. **(a)** The 'classical' tree derived from comparison of rRNA sequence and rooted with ancient paralogs. It is thought to arise from a collection of non-cellular supramolecular aggregates in the primordial soup, between which there is lateral gene transfer (LGT). A process dubbed genetic annealing gives rise to cells. In this scenario, the three domains of life - Eubacteria, Archaebacteria and Eukaryotes - branch off in that order. **(b)** The introns-early tree. This proposes that the ancestor of all three domains contained introns, which were lost in the Archaebacteria and Euacteria. **(c)** The neomuran tree. This introduces an ancestral group of organisms from which Archaeabacteria and Eukaryotes arose after the loss of the eubacterial-type cell wall in one lineage (the neomuran revolution). **(d)** The symbiotic tree. This proposes that the ancestor of eukaryotes originated by the endosymbiosis of one prokaryote (X) in another prokaryote host (Y), giving rise to nucleated (n) eukaryotic cells. The different groups of eukaryotes arose by subsequent separate endosymbiotic events involving various prokaryotes - the ancestors of plastids (p) and mitochondria (m) - in host cells of this lineage. **(e)** The prokaryote-host tree. This also incorporates endosymbiosis as the origin of mitochondria and plastids, but proposes that the endosymbiotic event that gave rise to a cell containing nucleus and mitochondria occurred in a prokaryotic host. This leads to a ring-like relationship between the ancestral organisms rather than a tree (see inset 2). This model also invokes extensive LGT throughout microbial evolution (see inset 1). See text for further details.



eubacteria were the only organisms on Earth until about 900 million years ago. At this time, a member of the eubacteria, in recent versions an actinobacterium, lost its murein-containing cell wall and was faced with the task of reinventing a new cell wall (hence the Latin name: *neo*, new; *murus*, wall). This led to the origin of a group of rapidly evolving organisms that Cavalier-Smith calls the Neomura.

The loss of the cell wall precipitated an unprecedented process of descent with modification in this group. During a short period of time (perhaps 50 million years), the characters that are shared by archaebacteria and eukaryotes arose (for a list of those characters, see [31]). The neomuran lineage then underwent diversification into two lineages, with another long list of evolutionary changes in each. One lineage invented isoprene ether lipid synthesis and gave rise to archaebacteria. The other became phagotrophic and gave rise to the eukaryotes. In older versions of this hypothesis, some eukaryote lineages branched off before the mitochondrion was acquired; these lineages were once called the Archezoa [30]. In newer versions, the mitochondrion comes into the eukaryote lineage before any archezoan can arise. No evolutionary intermediates from the transitions of actinobacteria into neomurans, archaebacteria, and eukaryotes persist among the modern biota, which is a distressing aspect of the theory for many specialists. The neomuran theory accounts mainly for cell biological characters, but not for sequence similarity among genes.

The symbiotic tree: a merger of distinct branches

At about the same time that archaebacteria and introns were being discovered, biologists were still fiercely debating the issue of whether mitochondria and chloroplasts were once free-living prokaryotes [33] or not [34]. Lynn Margulis had revived the old and controversial theories from the early 20th century regarding the endosymbiotic origin of chloroplasts and mitochondria [35,36]. Margulis's version of endosymbiotic theory was one of eukaryotes-in-pieces, and has always contained an additional partner at eukaryote origins to which no specialists other than herself have given credence: the spirochete origin of eukaryotic flagella [35-37]. Other prokaryote symbioses *en route* to eukaryotes involve the possible endosymbiotic origin of peroxisomes [38,39], or an endosymbiotic origin of the nucleus [40-42]. Common to those theories are a eubacterial-archaebacterial merger of some sort at the origin of eukaryotes (X and Y in Figure 1d), giving rise to a nucleated but mitochondrion-lacking cell - an archezoon [30] - followed by the origin of mitochondria.

From the viewpoint of more modern data, the spirochete origin of eukaryotic flagella can be seen as both unsupported and unnecessary [43], as can an endosymbiotic origin for peroxisomes, for which there are also no supporting data [44]. The origin of the nucleus is still debated [45].

The prokaryote-tree with LGT: a merger of ephemeral genomes

An exciting prospect predicted by all the foregoing hypotheses was that the most primitive eukaryotic lineages should lack mitochondria. That sent molecular biologists scrambling to study contemporary eukaryotes that were thought to lack mitochondria, work that unearthed findings of the most unexpected kind: all of the purportedly primitive and mitochondrion-lacking lineages were not really primitive nor did they even lack mitochondria. The mitochondria are there, it turns out, but they do not use oxygen [46,47], they are small [48], and some do not even produce ATP [49]. These 'new' members of the mitochondrial family among eukaryotic anaerobes (and some parasitic aerobes [50]) are called hydrogenosomes and mitosomes (reviewed in [51]). That pointed to the possibility that there never were any eukaryotes that lacked mitochondria; hence, the host that acquired the mitochondrion might have just been an archaeon outright (Figure 1e). Several hypotheses of this sort have been published, some of which account for the common ancestry of mitochondria and hydrogenosomes (reviewed in [52]) and some of which account for the origin of the nucleus [53].

Like the symbiotic tree, the prokaryote-host tree can accommodate LGT [54] without problems (Figure 1e, inset 1), and furthermore implies the existence of ring-like structures [55], rather than tree-like structures linking prokaryotes and eukaryotes at the level of gene content and sequence similarity (Figure 1e, inset 2). The only real difference between the symbiotic tree and the prokaryote-host tree hypotheses concerns the number of symbiotic partners involved at eukaryote origins - more than two versus two, respectively - and the existence (or nonexistence) of primitively amitochondriate eukaryotes. Both predictions are, in principle, testable with genome data, but the tests become a bit more complicated than standard phylogenetic tests, because of LGT [52].

The biggest branch is the biggest problem

For many biologists concerned with life's deeper relationships, the longest and most strongly supported branch in many current versions of the tree of life as depicted in Figure 1a or in recent papers [9,16] is also the most misleading: the central branch that implies a sister-group relationship between eukaryotes and archaebacteria [9,13]. It is misleading because at the level of genome-wide patterns of sequence similarity, eukaryotes are far more similar to eubacteria than they are to archaebacteria [56]. Put another way, eukaryotes possess more eubacteria-related genes than they possess archaebacteria-related genes [56,57]. This has escaped the attention of almost everyone, and is one of evolutionary biology's best-kept secrets, at least in circles where the rRNA tree is thought to speak for the whole genome.

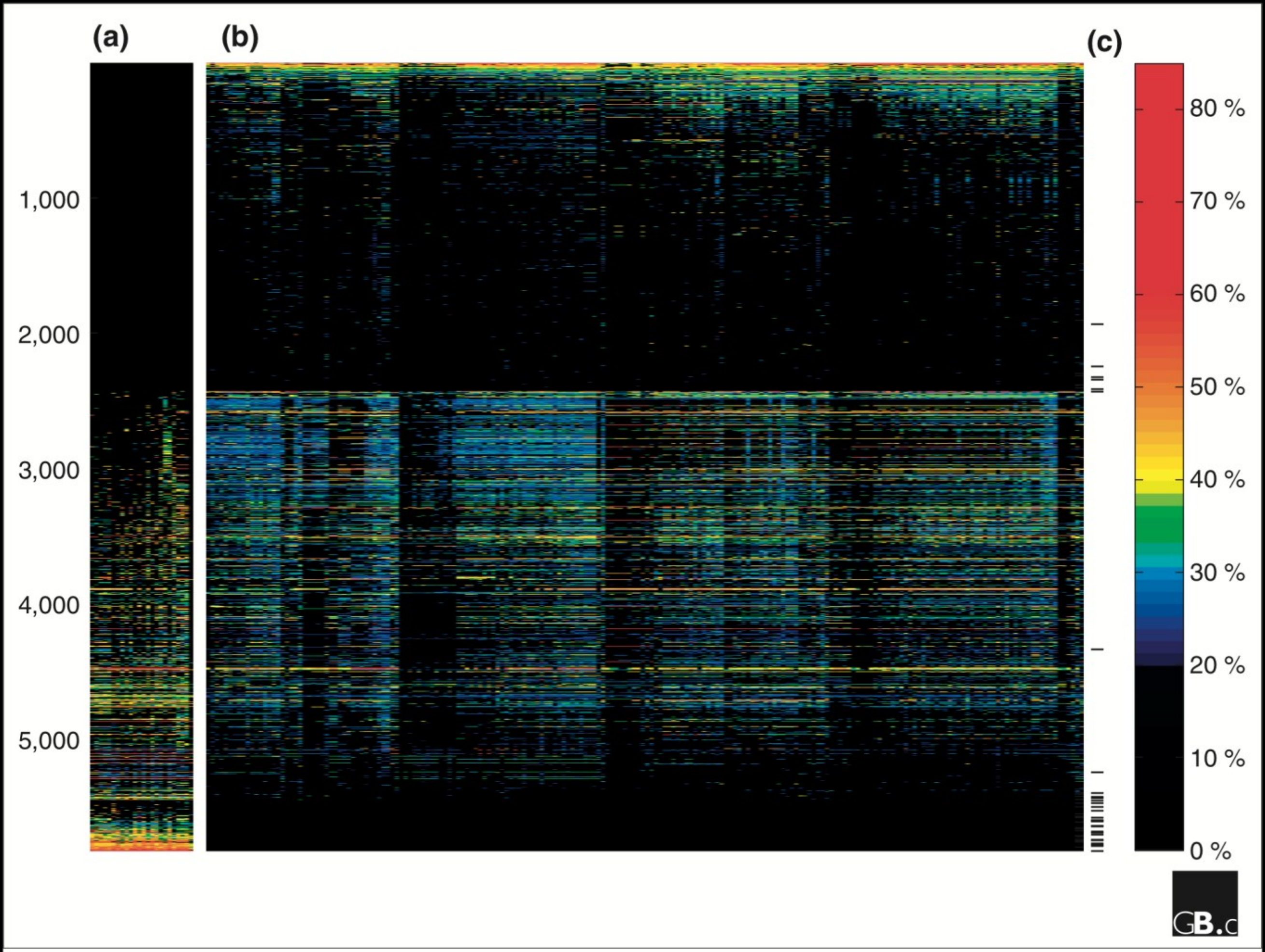


Figure 2
As a representative eukaryote example, the non-redundant set of human proteins (NCBI's Refseq database [70]) was compared using BLAST to a data set containing all proteins from 224 prokaryotic genomes: **(a)** 24 archaebacteria and **(b)** 200 eubacteria. In each panel, individual genomes are represented by columns and individual proteins by rows; numbers of proteins are indicated on the left and percentage amino-acid identity by the color scale shown on the right. BLAST hits with an e-value ≤ 10^{-20} and ≥ 20% amino-acid identity were recorded. The percent identity of the best blast hit for each human protein in each prokaryote was color coded as shown on the right and plotted with MATLAB©. The 31 proteins that were used in the recent tree of life [9] are marked with ticks in column **(c)**. A table containing the numbers, genes, and species underlying the figure is available as additional data file 1.

An example emphasizing this point is shown in Figure 2, where the percentage amino-acid identity between eukaryotic proteins (human in this example; yeast in [56]) and their homologs in prokaryotes (when present) is depicted. Of the 5,833 human proteins that have homologs in these prokaryotes at the specified thresholds, 2,811 (48%) have homologs in eubacteria only, while 828 (14%) have homologs in archaebacteria only, and 4,788 (80%) have greater sequence identity with eubacterial homologs, whereas 877 (15%) are more similar to archaebacterial homologs (196 are ties). The proteins comprising the recent tree of life - or the tree of one percent [9] - belong almost exclusively to the informational class [57]; that is, they are involved in information storage and processing. It is well known that eukaryotic informational genes are archaea-like [55-57]. They indicate a close relationship of eukaryotes and archaebacteria, but as is clearly visible in Figure 2, they speak for only a very small minority of eukaryotic genes [56].

Eukaryotes possess genes that they have inherited from archaebacteria and from eubacterial organelles [58]. But in plants, the acquisition of genes from cyanobacteria (plastids) has been estimated as 18% of the genome; the acquisition from mitochondria could be even greater [52]. Because such substantial gene influxes cannot be represented with bifurcating trees, they are usually just ignored.

A refreshing exception to the assumption that the tree of life is a tree to begin with is the recent paper by Rivera and Lake [55], who reported a procedure that takes LGT into account;

it shows eukaryotes as the sisters of archaebacteria and eubacteria simultaneously (Figure 1e, inset 2). But Rivera and Lake [55] did not force the data onto a tree; rather, they looked to see whether the data were actually tree-like in structure, and found that a directed acyclic graph (a ring) represents the underlying evolutionary process linking prokaryotes to eukaryotes better than a tree does. They offered crisp arguments that endosymbiosis is the most likely cause for the ring-like nature of the data.

But not everyone agrees that symbiosis was important in eukaryote evolution. Some biologists, mainly from the positivist camp, categorically reject the idea that eukaryotes acquired many, or any, genes from endosymbionts, and they scorn the notion that endosymbiosis had anything to do with eukaryote origins [15,19,39]. An argument salient to that view is the sweeping claim that endosymbiosis and gene transfer from endosymbionts fails to account for the evolution of any outstanding eukaryote characters [19], such as the nucleus. A more optimistic view from the microbialist camp is that the endosymbiotic origin of mitochondria could have made a major contribution to the genetic makeup of eukaryotes [58,59]. This could account for the finding that operational genes of bacterial origin are in the majority in eukaryote genomes [52]. The origin of mitochondria could have even precipitated the origin of the nucleus via the introduction of introns into eukaryotic lineages [53]. The roles of LGT and endosymbiosis in evolution have always been controversial. Genomes attest that both processes are important [23], but neither can be handled by strictly bifurcating trees as a means to represent genome evolution.

Seeing the wood for the trees

The need to incorporate non-treelike processes into ideas about microbial evolution has long been evident [57,60-63]. But mathematicians and bioinformaticians are just now beginning to explore the biological utility of graphs that can recover and represent non-treelike process that sometimes underlie patterns of sequence similarity in molecular data and patterns of shared genes. These approaches can involve networks [64-67], rings [55], or simply tack inferred gene exchanges onto trees [4,68,69]. These newer approaches aim to recover and depict both the tree-like (vertical inheritance through common descent) and the non-treelike (LGT and endosymbiosis) mechanisms of microbial evolution. As such, they represent important advances, because both mechanisms are germane to the processes through which microbes evolve in nature.

So, are we close to having a microbial tree of life [9]? Or are we closer to rejecting a single tree as the null hypothesis for the process of microbial genome evolution [1,54]? All in all, the latter seems more likely, for if our search for the tree of life delivers the tree of one percent, then we should be searching for graphs and theories that fit the data better than a single bifurcating tree.

Additional data file

Additional data file 1 is a table containing the numbers, genes, and species on which Figure 2 is based.

References

1. Doolittle WF: **If the tree of life fell, would it make a sound?** In *Microbial Phylogeny and Evolution: Concepts and Controversies.* Edited by Sapp J. New York: Oxford University Press; 2004:119-133.
2. Mirkin BG, Fenner TI, Galperin MY, Koonin EV: **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* 2003, **3:**2.
3. Snel B, Bork P, Huynen MA: **Genomes in flux: the evolution of archaeal and proteobacterial gene content.** *Genome Res* 2002, **12:**17-25.
4. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA: **The net of life: reconstructing the microbial phylogenetic network.** *Genome Res* 2005, **15:**954-959.
5. Ge F, Wang LS, Kim J: **The cobweb of life revealed by genome-scale estimates of horizontal gene transfer.** *PLoS Biol* 2005, **3:**e316.
6. Lerat E, Daubin V, Ochman H, Moran NA: **Evolutionary origins of genomic repertoires in bacteria.** *PLoS Biol* 2005, **3:**e130.
7. Woese CR: **Interpreting the universal phylogenetic tree.** *Proc Natl Acad Sci USA* 2000, **97:**8392-8396.
8. Kurland CG, Canback B, Berg OG: **Horizontal gene transfer: a critical view.** *Proc Natl Acad Sci USA* 2003, **100:**9658-9662.
9. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life.** *Science* 2006, **311:**1283-1287.
10. Daubin V, Moran NA, Ochman H: **Phylogenetics and the cohesion of bacterial genomes.** *Science* 2003, **301:**829-832.
11. Gogarten JP, Townsend JP: **Horizontal gene transfer, genome innovation and evolution.** *Nat Rev Microbiol* 2005, **3:**679-687.
12. Pal C, Papp B, Lercher MJ: **Adaptive evolution of bacterial metabolic networks by horizontal gene transfer.** *Nat Genet* 2005, **37:**1372-1375.
13. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria and Eukarya.** *Proc Natl Acad Sci USA* 1990, **87:**4576-4579.
14. Woese CR: **The universal ancestor.** *Proc Natl Acad Sci USA* 1998, **95:**6854-6859.
15. Woese CR: **On the evolution of cells.** *Proc Natl Acad Sci USA* 2002, **99:**8742-8747.
16. Pace NR: **Time for a change.** *Nature* 2006, **441:**289.
17. Iwabe N, Kuma K-I, Hasegawa M, Osawa S, Miyata T: **Evolutionary relationship of archaebacteria, eubacteria and eukaryotes inferred from phylogenetic trees of duplicated genes.** *Proc Natl Acad Sci USA* 1989 **86:**9355-9359.
18. Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, *et al.*: **Evolution of the vacuolar H$^+$-ATPase: Implications for the origin of eukaryotes.** *Proc Natl Acad Sci USA* 1989, **86:**6661-6665.
19. Kurland CG, Collins LJ, Penny D: **Genomics and the irreducible nature of eukaryote cells.** *Science* 2006, **312:**1011-1014.
20. Doolittle WF: **Genes in pieces: Were they ever together?** *Nature* 1978, **272:**581-582.
21. Doolittle WF: **Revolutionary concepts in evolutionary cell biology.** *Trends Biochem Sci* 1980, **5:**147-149.
22. Doolittle WF: **The origin and function of intervening sequences in DNA: a review.** *Am Nat* 1987, **130:**915-928.
23. Stoltzfus A, Spencer DF, Zuker M, Logsdon JM Jr, Doolittle WF: **Testing the exon theory of genes: the evidence from protein structure.** *Science* 1994, **265:**202-207.
24. Forterre P: **Thermoreduction, a hypothesis for the origin of prokaryotes.** *C R Acad Sci III* 1995, **318:**415-422.

25. Jeffares DC, Poole AM, Penny D: **Relics from the RNA world.** *J Mol Evol* 1998, **46:**18-36.
26. Poole A, Jeffares D, Penny D: **Prokaryotes, the new kids on the block.** *BioEssays* 1999, **21:**880-889.
27. Forterre P, Philippe H: **Where is the root of the universal tree of life?** *BioEssays* 1999, **21:**871-879.
28. Penny D, Poole A: **The nature of the last universal common ancestor.** *Curr Opin Genet Dev* 1999, **9:**672-677.
29. Jeffares DC, Mourier T, Penny D: **The biology of intron gain and loss.** *Trends Genet* 2006, **22:**16-22.
30. Cavalier-Smith T: **The origin of eukaryote and archaebacterial cells.** *Ann NY Acad Sci* 1987, **503:**17-54.
31. Cavalier-Smith T: **The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa.** *Int J Syst Evol Microbiol* 2002, **52:**297-354.
32. Cavalier-Smith T: **Cell evolution and Earth history: stasis and revolution.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361:**969-1006.
33. Bonen L, Doolittle WF: **On the prokaryotic nature of red algal chloroplasts.** *Proc Natl Acad Sci USA* 1975, **72:**2310-2314.
34. Cavalier-Smith T: **The origin of nuclei and of eukaryotic cells.** *Nature* 1975, **256:**463-468.
35. Sagan L: **On the origin of mitosing cells.** *J Theor Biol* 1967, **14:** 225-274.
36. Margulis L: *Origin of Eukaryotic Cells.* New Haven, CT: Yale University Press; 1970.
37. Margulis L, Dolan MF, Guerrero R: **The chimeric eukaryote: Origin of the nucleus from the karyomastigont in amitochondriate protists.** *Proc Natl Acad Sci USA* 2000, **97:**6954-6959.
38. Cavalier-Smith T: **The simultaneous symbiotic origin of mitochondria, chloroplasts and microbodies.** *Ann NY Acad Sci* 1987, **503:**55-71.
39. de Duve C: *Singularities. Landmarks on the Pathways of Life.* Cambridge, UK: Cambridge University Press; 2005.
40. Lake JA, Rivera MC: **Was the nucleus the first endosymbiont?** *Proc Natl Acad Sci USA* 1994, **91:**2880-2881.
41. Gupta RS: **Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes.** *Microbiol Mol Biol Rev* 1998, **62:**1435-1491.
42. Horiike T, Hamada K, Miyata D, Shinozawa T: **The origin of eukaryotes is suggested as the symbiosis of *Pyrococcus* into γ-proteobacteria by phylogenetic tree based on gene content.** *J Mol Evol* 2004, **59:**606-619.
43. Jekely G, Arendt D: **Evolution of intraflagellar transport from coated vesicles and autogenous origin of the eukaryotic cilium.** *BioEssays* 2006, **28:**191-198.
44. Gabaldon T, Snel B, van Zimmeren F, Hemrika W, Tabak H, Huynen MA: **Origin and evolution of the peroxisomal proteome.** *Biol Direct* 2006, **1:**8.
45. Martin W: **Archaebacteria (Archaea) and the origin of the eukaryotic nucleus.** *Curr Opin Microbiol* 2005, **8:**630-637.
46. Müller M: **The hydrogenosome.** *J Gen Microbiol* 1993, **139:**2879-2889.
47. Müller M: **Energy metabolism. Part I: Anaerobic protozoa.** In *Molecular Medical Parasitology.* Edited by Marr J. London: Academic Press; 2003: 125-139.
48. Tovar J, Fischer A, Clark CG: **The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite *Entamoeba histolytica*.** *Mol Microbiol* 1999, **32:**1013-1021.
49. Tovar J, León-Avila G, Sánchez LB, Sutak R, Tachezy J, van der Giezen M, Hernández M, Müller M, Lucocq JM: **Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation.** *Nature* 2003, **426:**172-176.
50. Williams BA, Hirt RP, Lucocq JM, Embley TM: **A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*.** *Nature* 2002, **418:**865-869.
51. van der Giezen M, Tovar J, Clark CG: **Mitochondrion-derived organelles in protists and fungi.** *Int Rev Cytol* 2005, **244:**175-225.
52. Embley TM, Martin W: **Eukaryotic evolution, changes and challenges.** *Nature* 2006, **440:**623-630.
53. Martin W, Koonin EV: **Introns and the origin of nucleus-cytosol compartmentation.** *Nature* 2006, **440:**41-45.
54. Doolittle WF: **Phylogenetic classification and the universal tree.** *Science* 1999, **284:**2124-2128.
55. Rivera MC, Lake JA: **The ring of life provides evidence for a genome fusion origin of eukaryotes.** *Nature* 2004, **431:**152-155.
56. Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D, *et al.*: **A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes.** *Mol Biol Evol* 2004, **21:**1643-1660.
57. Rivera MC, Jain R, Moore JE, Lake JA: **Genomic evidence for two functionally distinct gene classes.** *Proc Natl Acad Sci USA* 1998, **95:**6239-6244.
58. Timmis JN, Ayliffe MA, Huang CY, Martin W: **Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes.** *Nat Rev Genet* 2004, **5:**123-135.
59. Doolittle WF, Boucher Y, Nesbo CL, Douady CJ, Andersson JO, Roger AJ: **How big is the iceberg of which organellar genes in nuclear genomes are but the tip?** *Philos Trans R Soc Lond B Biol Sci* 2003, **358:**39-58.
60. Brown JR: **Ancient horizontal gene transfer.** *Nat Rev Genet* 2003, **4:**121-132.
61. Martin W: **Is something wrong with the tree of life?** *BioEssays* 1996, **18:**523-527.
62. Doolittle WF: **Fun with genealogy.** *Proc Natl Acad Sci USA* 1997, **94:**12751-12753.
63. Feng D-F, Cho G, Doolittle RF: **Determining divergence times with a protein clock: update and reevaluation.** *Proc Natl Acad Sci USA* 1997, **94:**13028-13033.
64. Holland BR, Jermiin LS, Moulton V: **Improved consensus network techniques for genome-scale phylogeny.** *Mol Biol Evol* 2006, **23:**848-855.
65. Holland BR, Huber KT, Moulton V, Lockhart PJ: **Using consensus networks to visualize contradictory evidence for species phylogeny.** *Mol Biol Evol* 2004, **21:**1459-1461.
66. Huson DH, Dezulian T, Klöpper T, Steel MA: **Phylogenetic supernetworks from partial trees.** *IEEE/ACM Trans Comput Biol Bioinform* 2004, **1:**151-158.
67. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol* 2006, **23:**254-267.
68. Hao W, Golding GB: **Patterns of bacterial gene movement.** *Mol Biol Evol* 2004, **21:**1294-1307.
69. Susko E, Leigh J, Doolittle WF, Bapteste E: **Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the γ-proteobacteria.** *Mol Biol Evol* 2006, **23:**1019-1030.
70. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33:** D501-D504.

5.2 En route to a network of genomes

Dagan T, Martin W: **Getting a better picture of microbial evolution en route to a network of genomes**. *Phil Trans Roy Soc Lond* B 2009, **364**:2187-2196.

(Own contribution: 50%)

PHILOSOPHICAL
TRANSACTIONS
OF
THE ROYAL
SOCIETY
B

Phil. Trans. R. Soc. B (2009) **364**, 2187–2196
doi:10.1098/rstb.2009.0040

Getting a better picture of microbial evolution en route to a network of genomes

Tal Dagan* and William Martin

Institute of Botany, University of Düsseldorf, 40225 Düsseldorf, Germany

Most current thinking about evolution is couched in the concept of trees. The notion of a tree with recursively bifurcating branches representing recurrent divergence events is a plausible metaphor to describe the evolution of multicellular organisms like vertebrates or land plants. But if we try to force the tree metaphor onto the whole of the evolutionary process, things go badly awry, because the more closely we inspect microbial genomes through the looking glass of gene and genome sequence comparisons, the smaller the amount of the data that fits the concept of a bifurcating tree becomes. That is mainly because among microbes, endosymbiosis and lateral gene transfer are important, two mechanisms of natural variation that differ from the kind of natural variation that Darwin had in mind. For such reasons, when it comes to discussing the relationships among all living things, that is, including the microbes and all of their genes rather than just one or a select few, many biologists are now beginning to talk about networks rather than trees in the context of evolutionary relationships among microbial chromosomes. But talk is not enough. If we were to actually construct networks instead of trees to describe the evolutionary process, what would they look like? Here we consider endosymbiosis and an example of a network of genomes involving 181 sequenced prokaryotes and how that squares off with some ideas about early cell evolution.

Keywords: phylogeny; networks; genomics

1. INTRODUCTION

Some evolutionary relationships are well described by a series of recursive bifurcations—a tree. The phylogeny of birds, fish or mammals are examples. As the lineages split, so do the gene histories, leading to the expectation that different genes for such groups should tend to give roughly the same phylogeny, provided that molecular phylogeny generally works (Landan & Graur 2008), and provided that recurrent genome duplications, as are common among eukaryotes (Scannell *et al.* 2006), have not led to rampant (hidden) paralogy. Xenology is quite rare in eukaryotes (e.g. Rumpho *et al.* 2008) but, among the microbes, evolutionary relationships can entail cell mergers (endosymbiosis), donation and acquisition of genes, such that within a single organism or genome different genes can have fundamentally different histories.

For example, when we retrace the evolutionary process of lineage splittings down into the origin of different algal groups possessing plastids surrounded by three or four membranes, we are confronted with the process of secondary endosymbiosis, where cellular individuals of highly disparate eukaryotic lineages have merged on at least three different occasions to bring forth novel algal lineages at very high taxonomic levels (Stoebe & Maier 2002; Lane & Archibald 2008). This is sketched in figure 1, where the chromists and alveolates are drawn as separate groups as recent findings suggest (Frommolt *et al.* 2008; Sanchez-Puerta & Delwiche 2008). Going back further, the origin of the plant lineage is attributable to the symbiosis of a cyanobacterium with a eukaryotic host, another cellular merger in the phylogeny of life (Gould *et al.* 2008). Going back further still, the origin of mitochondria at the origin of known eukaryotes is yet another decisive cellular merger (Dyall *et al.* 2004; van der Giezen *et al.* 2005; Embley & Martin 2006). At each such symbiotic merger, genes are transferred from symbionts to the chromosomes of their host, a process called endosymbiotic gene transfer (Martin *et al.* 1993, 1998, 2002; Timmis *et al.* 2004). Further back still, among free-living prokaryotes, from which the ancestors of plastids and mitochondria stem, lateral gene transfer has resulted in distributions of genes across prokaryotic chromosomes that do not strictly correspond to a hierarchical classification or any single bifurcating tree (Doolittle 1999; Doolittle & Bapteste 2007; McInerney *et al.* 2008).

Ho hum, one might say, we knew that, so what's new? Maybe the more important question is not what's new, but what isn't new? What isn't new is that despite knowing that many processes in microbial evolution are not tree-like in nature, biologists still tend to use the metaphor of trees to conceive, discuss and represent the process of the overall relatedness of things (Ciccarelli *et al.* 2006). That is not terribly surprising because most, but not all, approaches to describing the evolutionary process involve phylogeny at the computer in some form, and phylogenetic work at the computer generally produces trees, because trees are the most simple way to model protein divergence (a node), evolution (a branch) and homology (a clade) but they contain no information as for the type of homology (e.g. orthology, xenology, etc.).

* Author for correspondence (tal.dagan@uni-duesseldorf.de).

One contribution of 11 to a Theme Issue 'The network of life: genome beginnings and evolution'.

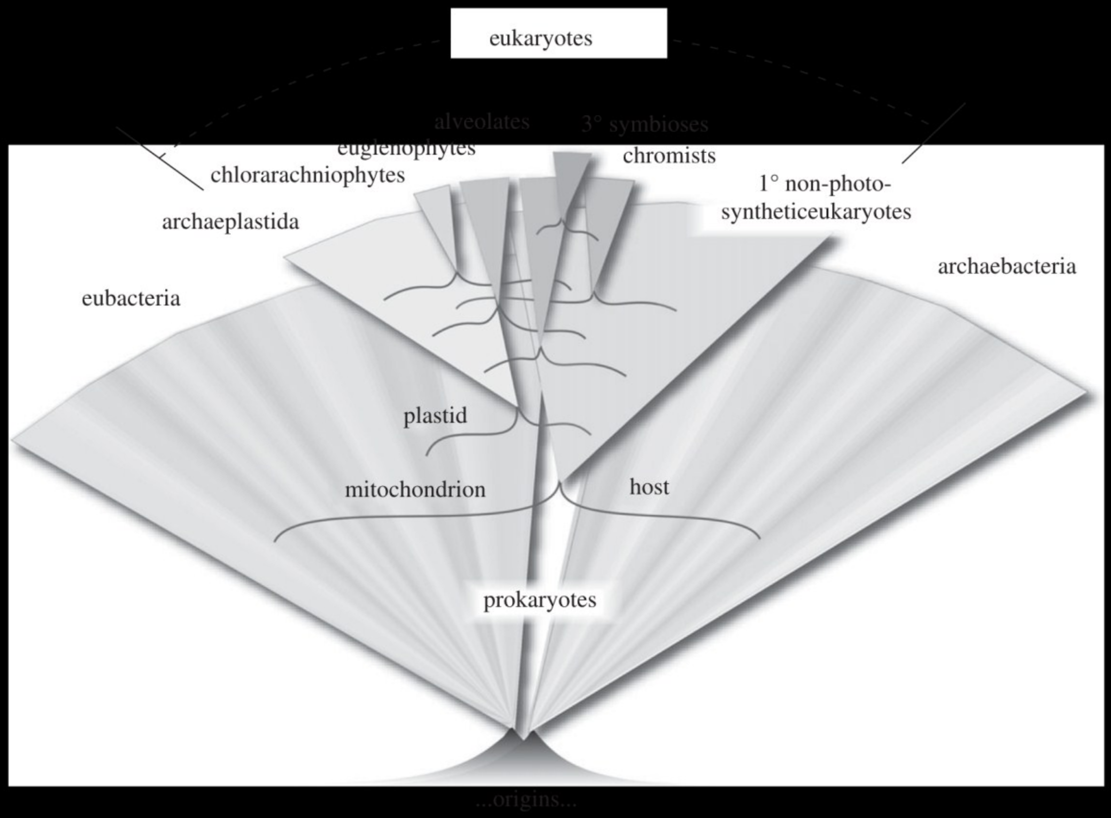


Figure 1. A schematic depiction of cell evolution connecting prokaryotes and eukaryotes and eukaryotes with complex plastids. Diversification of groups is symbolized by triangles, but branching patterns for groups is not. Thin lines subtending each triangle indicate that only two cellular partners (one cell per thin line each) participate in any given endosymbiosis. Origins symbolizes the origins of prokaryotes from the elements on early Earth, eubacteria and archaebacteria are prokaryotes, the rest of the lineages are eukaryotes. See text for more details.

There are some exceptions, the directed cyclic graphs used by Rivera & Lake (2004) to generate a ring instead of a tree being one, and there are others (Beiko *et al.* 2005; Kunin *et al.* 2005; Dagan *et al.* 2008; Lima-Mendez *et al.* 2008). But by and large, biologists tend to have a tree in mind when they approach the issue of evolutionary relatedness, also among microbes, and they (we) tend to use programs that generate trees, and therefore they (we) tend to see trees as the result of investigation on the topic. Many graphical representations of microbial evolution that incorporate LGT and/or that depict endosymbiosis have been published (Doolittle 1999; Martin 1999; Brown 2003; Huang & Gogarten 2006; McInerney *et al.* 2008) but, like figure 1, which schematically depicts how endosymbiosis runs contrary to the notion of a strictly bifurcating tree, they tend to involve something like an artist's impression of the evolutionary process that takes into account many different kinds of observations. It would probably help matters, that is, it would probably help us as evolutionary biologists studying microbial evolution, to convey in a more objective and scientific manner to ourselves and to non-specialists alike, if we could produce as a computer-generated printed product of our current understanding of microbial evolution. That would require computer-based methods that would allow us to depict non tree-like processes for the simple purpose of having a better level of congruence between what we think is actually going on in microbial evolution in nature (based on observations) and how we model it. If microbial evolution is not tree-like in salient aspects, then we need tools to investigate it that do not force the data into the straightjacket of a tree.

2. NETWORKS

Phylogenetic relationship can be modelled as graphs, in which the species (or genes) are represented by vertices and their evolutionary relationship are represented by edges. By graph theory definitions, a phylogenetic tree is a connected, acyclic, directed (and sometime also rooted) graph (Harary 1969). But there are some alternatives to trees. Networks are one such kind of alternative. Using the same mathematical model, if we allow the graph to be cyclic, then we get a phylogenetic network (Huson & Bryant 2006). Hence, we could construct an evolutionary graph of shared genes among prokaryotic genomes in which the nodes (or vertices) of the graph represent sequenced genomes with the edges between nodes representing shared genes. If all of the gene inheritance were vertical, then we should obtain a tree. If there are lateral components of inheritance, then the network should recover and depict them, too.

A problem arises though, in that it is not as simple as it might seem at first sight to discern between vertical inheritance and lateral transfer. Any gene tree that is assumed to be an accurately inferred gene tree but is also discordant with the *a priori* expected relationships for the taxon labels (the species containing the respective gene), disregarding for the moment the issue of whence those *a priori* expectations stem, can readily be explained by assuming some number of ancient gene duplications and differential loss. But each time we assume a duplication and a loss to explain discordant branches, we are assuming the presence of an additional gene in the genome ancestral to the species under study. That is fine for one or two genes, or maybe a dozen or maybe a hundred. If, however, we have to add that kind of corollary assumption to *every* prokaryotic gene and its tree, then the size of the ancestral genome that results from those corollary assumptions begins to burgeon and quickly reaches an untenable size, that is, it becomes the genome of Eden, as Doolittle *et al.* (2003) put it. That logical constraint turns out to be a very useful tool, it turns out, in our efforts to understand gene transfer and chromosome evolution, as we briefly explain in the following.

We recently undertook an endeavour to describe prokaryote genome evolution in terms of networks (Dagan *et al.* 2008). In essence, we assorted 539 723 protein coding genes among 181 sequenced prokaryote genomes into 54 349 families using the standard MCL algorithm. Many of those families have a very patchy distribution, that is, members of many families are found in a few genomes from different taxonomic groups. If we make the extreme and testable assumption that there has been no LGT in the evolution of those genes among the 160 eubacteria and 21 archaebacteria sampled, then the distributions of those genes shared across more than one genome would be governed by lineage specific gene origin and gene loss only. That assumption can be tested by comparing the distribution of inferred ancestral genome sizes under such assumptions with the modern distribution of contemporary genome sizes, measured in gene families to see if they are significantly different, which they are (Dagan & Martin 2007*a*). A premise underlying that test is that there is no *a priori* reason to expect that prokaryotic genome sizes in the past were fundamentally different from those observed today. If we assume that there is no LGT then we are also assuming that all gene trees are compatible, and each gene is present in the genome ancestral to its first appearance in the evolution of the genomes we are considering. Thus gene distributions alone demand a certain amount of LGT among prokaryotic genomes, at least approximately 1 LGT per gene family per gene family lifespan, because too much vertical inheritance leads us into the genome of Eden problem (Dagan & Martin 2007*a*). Allowing LGT reduces the inferred size of ancestral genomes, but allowing too much LGT reduces their size to distributions that are once again significantly different from modern genome sizes, but too small (the genome of Lilliput) rather than too large.

The constraint of ancestral genome size opens two inroads to studying genome evolution. First, it permits estimates for how much LGT has gone on in prokaryote evolution (Dagan & Martin 2007*a*). Those estimates are attained without comparing gene trees and furthermore by assuming all gene trees to be compatible, hence they constitute minimum lower bound estimates. Second, it permits us to address genome evolution in terms of evolutionary networks consisting of vertically and laterally inherited genes. How? Given an assumed (or inferred) phylogeny for any given component of the genomes in question, then each of the ancestral nodes in that phylogeny corresponds to a genome-sized collection of genes. The constraint of genome size provides a criterion to decide whether a gene is present at a given ancestral node, that is, present in an inferred ancestral genome, or not. That is important because if we have a criterion for deciding which genes are present at which nodes, then shared genes across nodes correspond to edges in a network, and we can construct an evolutionary network that captures both vertical and horizontal components of gene inheritance, as illustrated in figure 2.

Figure 2*a* shows an assumed phylogeny for 181 genomes and corresponds to the topology and species designations shown in fig. 3a and the supplementary material of Dagan *et al.* (2008). The tree that we use as a vertical backbone was not just assumed from thin air, rather it was constructed from analyses of the rRNA operon assuming monophyly for the prokaryotic taxa shown, but its specific branching order might as well just have been assumed, for two reasons. First, there is currently little evidence to suggest that any genes in prokaryote genomes have strictly co-evolved with the rRNA operon over the whole of evolutionary time (Bapteste *et al.* 2008), hence even if we had the right rRNA tree, there remains the more pressing question of ‘for what would it be a proxy?’ (Doolittle & Bapteste 2007). Second, for 181 genomes there are 3.6×10^{379} possible trees, and the chances of getting the right tree are comfortingly negligible (by comparison there are about 10^{80} protons in the universe, very close to the number of trees for 60 genomes). Nonetheless, we can work with that assumed tree and specify as its root the branch between eubacteria and archaebacteria, because that is where genome similarity as measured in shared proportions of shared genes would place the root (Dagan & Martin 2007*a*), notwithstanding other suggestions as to where the root might be (see the contribution by Lake *et al.* 2009). Then, given the genome of Eden constraint, we can draw an edge between all nodes that are connected by a shared gene, that is, nodes that are connected by the presence of a member of one of our 54 349 protein families, which would give us a network of genomes.

Before drawing such a network, there is a matter to consider concerning the congruence between the edges to be drawn in the figure (shared genes) and the process they are intended to represent (LGT). If we infer that there was only one LGT in the history of a given gene family, then there is only one lateral edge connecting the nodes bearing that gene. In that case, there is a 1 : 1 correspondence between the number of lateral edges and the number of LGTs. But if three nodes need to be connected by lateral edges, then there are three edges that connect them but there are only two LGT events at the minimum need to be assumed, which applies to 27 per cent of the genes in the present example. Similarly, if four

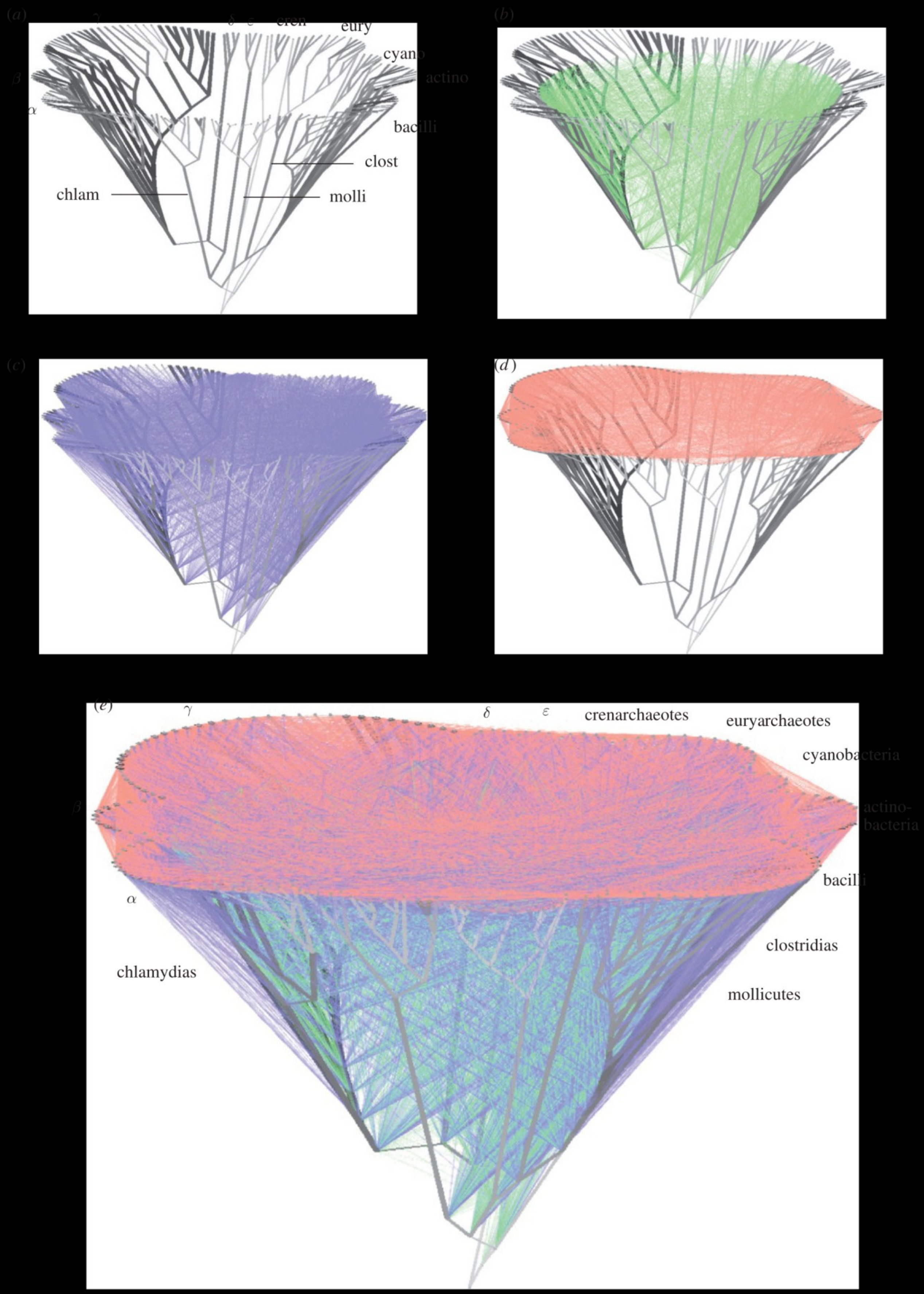


Figure 2. A network representation of vertical inheritance and lateral exchange among prokaryotes. (*a*–*d*) The individual components from which (*e*), modified from Dagan *et al.* (2008), was constructed. Prokaryotic groups sampled are indicated, greek letters designate proteobacterial subdivisions. Note that in (*c*), where lateral edges connecting internal nodes to external nodes are shown, some genes seem to be transferred from 'ancient' nodes to more modern nodes, whereas the lateral edge specifies the group of taxa from within which the donor or recipient is inferred. Similar applies to (*b*) for edges connecting internal nodes. See text for more details.

nodes need to be connected by lateral edges, then there are six edges that can connect them, but only three LGT events are needed to explain the gene distribution. Kunin *et al.* (2005) dealt with this problem by assigning weights to lateral edges corresponding to their probabilities of 2/3, 3/6, etc. We dealt with it by taking 1000 replicate samples from the matrix representation of the lateral network in which superfluous edges are randomly deleted, such that the number of lateral edges and the number of LGT events exactly correspond (Dagan *et al.* 2008). Using that procedure, we can plot the lateral edges onto figure 2*a*, which represents vertical inheritance among genomes, and furthermore depict lateral transfer among genomes as well, which was one aim of our undertaking to use networks for describing genome evolution.

Among the 1000 replicates, there are 2330 ± 16 lateral edges that connect internal nodes to internal nodes (figure 2*b*), 5886 ± 20 lateral edges that connect internal nodes to external nodes (figure 2*c*) and 4046 ± 16 lateral edges that connect external nodes to external nodes (figure 2*d*). Each of these edges corresponds to a lateral gene transfer, and if we plot all edges in one figure, the result is that shown in figure 2*e*. We designate the network as a minimal lateral network because the procedures that were used to determine gene presence or absence at nodes entail two simplifying assumptions that severely underestimate the amount of LGT that has actually gone on among genomes: (i) we assume that all genes are orthologous, that is, that all multiple occurrences of a gene family in a genome are assumed to be the result of recent gene duplications within that genome, and (ii) we assume that all gene trees for all families are compatible. Those are rather severe assumptions, but they do deliver estimates for the minimum LGT rate and the minimum number of LGT events to be plotted in the network.

3. INCLUDING EUKARYOTES IN THE NETWORK

Ideally, one would like to see eukaryotes and prokaryotes in the same network of shared genes and it can be expected that such graphs will eventually emerge. But current gene sharing networks encompass only prokaryotes (Kunin *et al.* 2005; Dagan *et al.* 2008) or phage (Lima-Mendez *et al.* 2008). They depict genome evolution among prokaryotes as a process of donor–recipient relationships that has been more or less continuous over evolutionary time, with genes acquired by conjugation (plasmids), transduction (phages), transformation (natural competence) (Thomas & Nielsen 2005) or gene transfer agents (Lang & Beatty 2007) but also being transmitted vertically by the process of chromosome replication as well, in agreement with some current views the process of microbial genome evolution (Doolittle & Bapteste 2007). While these four mechanisms of gene spread among prokaryotes just mentioned are very well characterized at the molecular and genetic level, similar genetically and molecular defined mechanisms have not been characterized among eukaryotes. Thus it would seem that there is a big difference between prokaryotes and eukaryotes concerning the prevalence, mechanisms and biological significance of lateral gene transfer. Indeed, it is not unusual to find that three strains of the same prokaryotic species such as *E. coli* might share less than 40 per cent of their genes in common (Welch *et al.* 2002), while sequencing a representative for a eukaryotic lineage, such as *Entamoeba*, might reveal only 1–2% of the genome consisting of genes that might have been specifically acquired in that lineage (Loftus *et al.* 2005). Clearly, the frequency and impact of LGT in prokaryote and eukaryote genomes is different.

But at the same time, a particular kind of gene transfer among eukaryotes, namely gene transfer from organelles, or endosymbiotic gene transfer (Martin *et al.* 1993), represents a very important source of genetic novelty among the eukaryotes (Timmis *et al.* 2004; Lane and Archibald 2008). Gene transfer from organelles sets eukaryotes apart from prokaryotes, which in contrast to eukaryotes lack organelles descended from free-living prokaryotes. That is not to say that no prokaryotes harbour prokaryotic endosymbionts, for there are two such examples known (Wujek 1979; van Dohlen *et al.* 2001), but there are no prokaryotes known to harbour double-membrane bounded organelles, raising the question of what is an endosymbiont and what is an organelle (Cavalier-Smith & Lee 1985). A practical distinction between the two is whether the endosymbiont has evolved a protein import apparatus, as in the case of chloroplasts (Kanalon & McFadden 2008), mitochondria (Dolezal *et al.* 2006) and secondary plastids (Hempel *et al.* 2007), in which case it would qualify as an organelle, or not, in which case it is best called an endosymbiont (Theissen & Martin 2006).

Endosymbionts living in the cytosol are very common among eukaryotes today and probably have been throughout evolution (Dagan & Martin 2007*b*), but endosymbiotic associations that give rise to organelles are not common at all. Available evidence indicates that there was only one origin of plastids from cyanobacteria (Gould *et al.* 2008), and only one origin of mitochondria from proteobacteria (see contribution by Embley in this volume), as sketched in figure 1. Once every 4 Gyr is rare. Both symbioses entailed the origin of a specific protein-import machinery. Both entailed the origin of a novel taxon at the highest levels (known plants and known eukaryotes). Both entailed a symbiosis of one cell within another, each possessing a genome's worth of genes. If an endosymbiont lyses, its chromosome is free to recombine with that of its host, if the host lyses, the symbiosis is over, hence the transfer of genes is generally unidirectional from endosymbiont to host, which can be seen as a ratchet mechanism (Doolittle 1998). We can see the workings of endosymbiotic gene transfer in eukaryote genomes today. We can see that bulk recombination is involved, as the 367 kb insertion of the complete mitochondrial genome in Arabidopsis and the 121 kb insertion of the complete chloroplast genome attest (Huang *et al.* 2005). The mechanism of insertional recombination involves non-homologous end joining (Hazcani-Covo & Covo 2008). Gene transfer from transformed mitochondria and from transformed plastids can be demonstrated in the lab

(Thorsness & Fox 1990; Huang *et al.* 2003) and there is increasing interest in the role of stress factors, such as oxidative stress, that might promote the rate incorporation of organelle sequences in nuclear genomes over recent evolutionary time (Cullis *et al.* 2008). Given the ease and frequency with which genes are transferred from organelles to the nucleus, the question arises as to why there are any genes left in organelles at all, and despite many different proposals to account for this observation, only one really fills the bill, namely that of Allen (1993, 2003), who suggested that organelles have retained genomes in order to allow redox-dependent regulation of gene expression within individual organelles that possess bioenergetic membranes. This proposal is strongly supported by recent characterization of proteins involved in redox-regulated plastid gene regulation (Puthiyaveetil *et al.* 2008) and would furthermore directly account for the lack of DNA in hydrogenosomes, anaerobic forms of mitochondria that generate energy via substrate level phosphorylation, and hence lack membrane-associated electron transport (Müller 2007).

There is also evidence for the workings of endosymbiotic gene transfer early in evolution as well. In plants, estimates for the fraction of genes acquired from the ancestor of plastids range from approximately 15 to 20 per cent of nuclear protein coding genes, with systematic underestimations owing to the difficulties of phylogenetic inference with poorly conserved sequences figuring prominently in the issue (Deusch *et al.* 2008). In eukaryotes that never possessed plastids, such as yeast, the majority of genes having homologues among prokaryotes are more similar to eubacterial homologues than they are to archaebacterial homologues and the former are generally involved in metabolic functions (operational genes) while the latter are generally involved in information storage and expression (informational genes) (Rivera *et al.* 1998; Esser *et al.* 2004; Rivera & Lake 2004).

The generally surprising observation that eukaryotes possess a majority of eubacterial genes (Martin *et al.* 2007) is distinctly at odds with the view that eukaryotes are sisters of archaebacteria, but it is readily accounted for under endosymbiotic models for the origin of eukaryotes (Pisani *et al.* 2007), if we allow for the very real possibility that there was a substantial quantity of endosymbiotic gene transfer subsequent to the origin of mitochondria. That brings us to the question of which genes, exactly, the ancestor of mitochondria, or the ancestor of plastids for that matter, possessed? We can phrase that question another way, and in the specific context of this paper, namely, what is the relationship of figure 1 to figure 2? Both figures purport to represent something that most people who will ever read this paper generally accept, namely that plastids and mitochondria really are descended from free living endosymbionts (figure 1) and that prokaryotes really do redistribute their genes across chromosomes over time (figure 2). If we add to that the recognition that many genes in eukaryote genomes really do stem from those two endosymbionts via endosymbiotic gene transfer, then which genes did those endosymbionts harbour in their chromosomes at the time when they became endosymbionts?

If we take the evidence seriously that prokaryotes really do pass their genes around over time, as we should, then it would appear that the collection of genes possessed by the ancestor of mitochondria is probably best preserved in its most contiguous form among eukaryote genomes, rather than among prokaryote genomes. This issue has been around for about 10 years (Martin 1999; Esser *et al.* 2007) but for the most part it has been disregarded, with some exceptions (Gross *et al.* 2008). For example, Huang *et al.* (2005) recently reported that there are some genes the plants and chlamydias share more or less specifically, and they suggested that this constitutes evidence for the participation of an additional endosymbiont, a chlamydial one, at the origin of plastids. But if we let go of the notion that the chromosomes of prokaryotic 'lineages' are static collections of genes that have co-evolved in a linked manner within the same chromosome over billions of years (Doolittle 1999), as data from genomes suggests that we should (Doolittle & Bapteste 2007), then we can contrast two ways of looking at the chlamydia data as an example of many similar sorts of observations emerging from genomes: (i) is it more reasonable to assume that a gene or group of genes can be used as a proxy for the existence of an additional endosymbiont in the plant lineage? Put another way, does every gene, in the extreme, serve as a proxy for the expected patterns of sequence similarity for the rest of the genes present in a given chromosome at a given point in time? Or (ii) are prokaryotic chromosomes, including those related to the ancestors of organelles, really 'fluid' structures, with genes coming in and going out over time? In our view, the latter question is much closer to being a formulation to which we could respond with a straightforward 'yes' and feel comfortable saying so.

It will probably take some time before LGT among prokaryotes (figure 2) and the endosymbiotic origins of chloroplasts and mitochondria (figure 1) can be reconstructed at the computer in a unified framework that starts with genome sequences and ends up with a network that is both readily printable and readily interpretable. It will take longer still before the secondary endosymbioses can be included in such an endeavour, because the data coming from those genomes are painting an increasingly complex picture (Frommolt *et al.* 2008; Sanchez-Puerta & Delwiche 2008). Apropos complexity, as the tsunami of data from eukaryote genome projects rolls in, it is being churned through various alignment and phylogeny pipelines and many of the trees so produced are showing unusual branching patterns or unusual sequence similarities. This has led to a situation where many reports for LGT among eukaryotes are emerging, the most spectacular being the initial claim for several hundred laterally acquired bacterial genes in the human genome, which turned out not to be true (Salzberg *et al.* 2001; Stanhope *et al.* 2001). However, because eukaryotes, in contrast to prokaryotes (Thomas & Nielsen 2005; Lang & Beatty 2007), lack genetically and molecularly well-defined mechanisms of gene transfer across species boundaries, the search for mechanisms to explain the presence of odd branching or otherwise unexpected sequences has been expanded to include mere physical contact between organisms (Keeling & Palmer 2008) or even LGT via meteorites (Bergthorsson *et al.* 2003), to highlight one prominent example. Such suggestions leave us less than comfortable.

In addition to the lack of molecularly characterized mechanisms, another contrast of LGT among prokaryotes to reports of eukaryote-to-eukaryote LGTs is that the latter all too often entail oddly branching copies of highly similar genes (Keeling & Palmer 2008) but without any corresponding effects for organismic ecology, whereas LGT among prokaryotes can, and often does, transform the overall physiology of an organism (Kennedy *et al.* 2001; Boucher *et al.* 2003; Mongodin *et al.* 2005) with dramatic and obvious consequences for its ecology and evolution. In that vein, chloroplasts and mitochondria also transformed the physiology of their hosts through endosymbiosis and donated some fundamentally new genes to their hosts (for example, for photosynthesis and mitochondrial ATP synthesis), not just divergent copies of the same ones.

Thus, LGT among prokaryotes and gene transfer in the context of endosymbiosis can be correlated to changes in ecology and physiology, but most of the reports for 'odd-branch' LGT among eukaryotes cannot (Keeling & Palmer 2008). This is not to say that eukaryotes never acquire genes from other eukaryotes. But the 'odd branch' approach to LGT has some hefty caveats because there are lots of genes out there in the databases and there are thousands of alignments and trees that can be made from them. Some of those trees will have high support values for artefactual branches for reasons intrinsic to the computational process of phylogenetic reconstruction (Delsuc *et al.* 2003; Bapteste *et al.* 2008; Shavit *et al.* 2007), and even the random choice of whether we align amino acids in a protein sequence from N-terminus to C-terminus or in the reverse order can exert a dramatic influence on phylogenetic and phylogenomic results (Landan & Graur 2007; Deusch *et al.* 2008). Such issues still loom somewhat over investigations of LGT that are based in tree comparisons alone and where the inference of LGT can account for differences in observed branching patterns, but little else.

4. WARNING: MANY PEOPLE DISAGREE WITH SOME VIEWS EXPRESSED HERE

We have presented two figures here to illustrate our current views on early evolution from the standpoint of endosymbiosis (figure 1) and LGT (figure 2). Figure 1 might be more controversial than figure 2 in various aspects and we feel obliged to point out that many scientists would staunchly disagree with aspects of the sketch presented in figure 1, hence a few words seem in order to justify why we drew it the way we did. In figure 1, we have sketched the origin of the host lineage for the origin of mitochondria as an archaebacterium outright, because it precludes the notion that nucleated but mitochondrion-lacking cells (archezoa) ever existed (Embley & Martin 2006) in agreement with some recent analyses based on supertrees (Pisani *et al.* 2007) and based on careful phylogenetic studies of informational genes (Cox *et al.* 2008). Some would staunchly disagree, maintaining that there are indeed eukaryotes around that never possessed mitochondria (Margulis *et al.* 2007), that the host that acquired the mitochondrion was a eubacterium (de Duve 2007), or that the common ancestry of mitochondria and hydrogenosomes is somehow tenuous (de Duve 2007; Margulis *et al.* 2007). We politely disagree, and will not argue their case here. We have indicated a later origin of eukaryotes than of prokaryotes, consistent with microfossil evidence suggesting their later emergence (Knoll *et al.* 2006; Rasmussen *et al.* 2008), and this runs contrary to views, with which we disagree, that eukaryotes represent a lineage that is as old as or older than prokaryotes (Kurland *et al.* 2007). We have drawn the root in figure 1 between archaebacteria and eubacteria, with which many scientists would also disagree, maintaining that archaebacteria arose via mutations from a bona fide eubacterium (Cavalier-Smith 2002) or that prokaryotes are derived from eukaryotes (Glansdorff *et al.* 2008) or that other placements of the root are preferable (see contribution by Lake *et al.* 2009). Again, we disagree and do not argue the opposing views.

Our placement of the root is consistent with geochemical evidence for the antiquity of both prokaryotic groups (Nisbet & Sleep 2001; Ueno *et al.* 2006) and with the observation that the two main groups of prokaryotes are deeply divergent, not only at the level of their cell wall and membrane constituents (Martin & Russell 2003), but also at the level of processes so basic as DNA maintenance (Koonin & Martin 2005). Also, we have drawn the base of figure 1 to suggest that the first prokaryotes might have arisen from something that looks like a hydrothermal vent, which need not be true, but there are enough similarities between energy-releasing geochemical reactions involving H_2 and CO_2 at some modern hydrothermal vents and energy releasing biological reactions involving H_2 and CO_2 among some modern microbes to pursue the idea further (Martin *et al.* 2008). Many scientists would disagree with the view that hydrothermal vents had anything to do with the origin of life (Orgel 2008).

Finally, there is the matter that we have not suggested any branching orders for either prokaryotic groups or eukaryotic groups in figure 1, other than implying that the organelle-generating symbioses among eukaryotes correspond to a relative temporal sequence. Among the prokaryotes, we have schematically indicated some kind of metabolic diversification (colours), but without suggesting what the order of appearance for different metabolic types might be. There is quite a lot of phylogenomic and phylogenetic work devoted to the relative branching orders of prokaryotic groups, and serious efforts have been undertaken to link that branching order to geochemical evidence and dates, for example in Battistuzzi *et al.* (2004) and Gribaldo & Brochier-Armanet (2006). Other efforts have focused inferring geological history from phylogenetic trees (Ciccarelli *et al.* 2006). But a general problem arises in such studies. In order to construct a tree for all groups, one has to have genes that are present in all groups, and this usually boils down to the ribosomal proteins or their superoperon (Hansmann and Martin 2000) or what has been called 'the core' (Charlebois & Doolittle 2004). The problem is that it is difficult to demonstrate that sequences differences or branching patterns in 'the core', should it evolve as a coherent unit in the first place (Bapteste *et al.* 2008), serve as a good predictor for which, what kind of, and how many genes we are likely to find in the remainder of the chromosome surrounding that core. For example, methanogens and archaeal halophiles have related and similar cores (Gribaldo & Brochier-Armanet 2006), but methanogens are strictly anaerobic chemolithoautotrophs while halophiles are (usually) aerobic heterotrophs with light-harnessing abilities (Kennedy *et al.* 2001; Boucher *et al.* 2003), while *Salinibacter* has a core similar to the eubacterial Bacteroides/Chlorobi group, but a physiology and gene collection reminiscent of archaeal halophiles (Mongodin *et al.* 2005). That example is certainly not new to anyone, but it perhaps illustrates the point that sequence similarities within the core are not a good proxy for what is likely to be found in the rest of the genome. Eukaryotes are another such example, the archaebacterial nature of their genetic apparatus does not predict the eubacterial nature of their energy metabolism, but some endosymbiotic models for the origin of mitochondria that entail gene transfers from symbiont to host do (Pisani *et al.* 2007).

5. CONCLUSION

The problems relating to the notion that the evolution of all living things can be represented by a tree have been well put by others (Doolittle 1999; Brown 2003; Doolittle & Bapteste 2007; McInerney *et al.* 2008), and we broadly agree with that view. The main non tree-like processes to deal with seem to be LGT among prokaryotes and gene transfer from organelles (endosymbiotic gene transfer) among eukaryotes. The onus of offering alternatives would appear to be upon those of us who are saying that the tree metaphor is inadequate. Networks are an alternative that can be used in the case of prokaryotes (Dagan *et al.* 2008). It is obvious that there exists some amount of vertical inheritance via chromosome replication and segregation as well as some amount of lateral inheritance via other means among prokaryotes; hence the network approach to genome evolution should depict both. If we approach the problem of describing the overall course of prokaryote genome evolution from the standpoint of shared genes among genomes rather than shared phylogeny of some core, as recent studies of phage evolution have (Lima-Mendez *et al.* 2008), then we are taking steps away from the familiar conceptual environment of trees and into the less well-charted territory of evolutionary processes that cannot be modelled by a tree, but might better fit the process of prokaryote genome evolution as it occurs in nature.

We thank the German Israeli Foundation (T.D.), the German Research Foundation, the European Research Council and the BMBF for financial support.

REFERENCES

Allen, J. F. 1993 Control of gene expression by redox potential and the requirement for chloroplast and mitochondrial genomes. *J. Theor. Biol.* **165**, 609–631. (doi:10.1006/jtbi.1993.1210)

Allen, J. F. 2003 The function of genomes in bioenergetic organelles. *Phil. Trans. R. Soc. B* **358**, 19–37. (doi:10.1098/rstb.2002.1191)

Bapteste, E., Susko, E., Leigh, J., Ruiz-Trillo, I., Bucknam, J. & Doolittle, W. F. 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. (doi:10.1093/molbev/msm229)

Battistuzzi, F. U., Feijao, A. & Hedges, S. B. 2004 A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol. Biol.* **4**, 44. (doi:10.1186/1471-2148-4-44)

Beiko, R. G., Harlow, T. J. & Ragan, M. A. 2005 Highways of gene sharing in prokaryotes. *Proc. Natl Acad. Sci. USA* **102**, 14 332–14 337. (doi:10.1073/pnas.0504068102)

Bergthorsson, U., Adams, K. L., Thomasson, B. & Palmer, J. D. 2003 Widespread horizontal gene transfer of mitochondrial genes in flowering plants. *Nature* **424**, 197–201. (doi:10.1038/nature01743)

Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau, M. E., Nesbo, C. L., Case, R. J. & Doolittle, W. F. 2003 Lateral gene transfer and the origins of prokaryotic groups. *Annu. Rev. Genet.* **37**, 283–328. (doi:10.1146/annurev.genet.37.050503.084247)

Brown, J. R. 2003 Ancient horizontal gene transfer. *Nat. Rev. Genet.* **4**, 121–132. (doi:10.1038/nrg1000)

Cavalier-Smith, T. 2002 The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int. J. Syst. Evol. Microbiol.* **52**, 7–76.

Cavalier-Smith, T. & Lee, J. J. 1985 Protozoa as hosts for endosymbioses and the conversion of symbionts into organelles. *J. Protozool.* **32**, 376–379.

Charlebois, R. L. & Doolittle, W. F. 2004 Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res.* **14**, 2469–2477. (doi:10.1101/gr.3024704)

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B. & Bork, P. 2006 Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287. (doi:10.1126/science.1123061)

Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R. & Embley, T. M. 2008 The archaebacterial origin of eukaryotes. *Proc. Natl Acad. Sci. USA* **105**, 20 356–20 361. (doi:10.1073/pnas.0810647105)

Cullis, C. A., Vorster, B. J., Van Der Vyver, C. & Kunert, K. J. 2008 Transfer of genetic material between the chloroplast and nucleus: how is it related to stress in plants? *Ann. Bot.* **103**, 625–633. (doi:10.1093/aob/mcn173)

Dagan, T. & Martin, W. 2007*a* Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl Acad. Sci. USA* **104**, 870–875. (doi:10.1073/pnas.0606318104)

Dagan, T. & Martin, W. 2007*b* Testing hypotheses without considering predictions. *BioEssays* **29**, 500–503. (doi:10.1002/bies.20566)

Dagan, T., Artz-Randrup, Y. & Martin, W. 2008 Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl Acad. Sci. USA* **105**, 10 039–10 044. (doi:10.1073/pnas.0800679105)

de Duve, C. 2007 The origin of eukaryotes: a reappraisal. *Nat. Rev. Genet.* **8**, 395–403. (doi:10.1038/nrg2071)

Delsuc, F., Phillips, M. J. & Penny, D. 2003 Hexapod origins: Monophyletic or paraphyletic? *Science* **301**, 1482. (doi:10.1126/science.1086558)

Deusch, O., Landan, G., Roettger, M., Gruenheit, N., Kowallik, K. V., Allen, J. F., Martin, W. & Dagan, T. 2008 Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol. Biol. Evol.* **25**, 748–761. (doi:10.1093/molbev/msn022)

Dolezal, P., Likic, V., Tachezy, J. & Lithgow, T. 2006 Evolution of the molecular machines for protein import into mitochondria. *Science* **313**, 314–318. (doi:10.1126/science.1127895)

Doolittle, W. F. 1998 You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trend. Genet.* **14**, 307–311.

Doolittle, W. F. 1999 Phylogenetic classification and the universal tree. *Science* **284**, 2124–2128. (doi:10.1126/science.284.5423.2124)

Doolittle, W. F. & Bapteste, E. 2007 Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl Acad. Sci. USA* **104**, 2043–2049. (doi:10.1073/pnas.0610699104)

Doolittle, W. F., Boucher, Y., Nesbo, C. L., Douady, C. J., Andersson, J. O. & Roger, A. J. 2003 How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Phil. Trans. R. Soc. B* **358**, 39–58. (doi:10.1098/rstb.2002.1185)

Dyall, S. D., Brown, M. T. & Johnson, P. J. 2004 Ancient invasions: from endosymbionts to organelles. *Science* **304**, 253–257. (doi:10.1126/science.1094884)

Embley, T. M. & Martin, W. 2006 Eukaryotic evolution, changes and challenges. *Nature* **440**, 623–630. (doi:10.1038/nature04546)

Esser, C. *et al.* 2004 A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* **21**, 1643–1660. (doi:10.1093/molbev/msh160)

Esser, C., Martin, W. & Dagan, T. 2007 The origin of mitochondria in light of a fluid prokaryotic chromosome model. *Biol. Lett.* **3**, 180–184. (doi:10.1098/rsbl.2006.0582)

Frommolt, R., Werner, S., Paulsen, H., Goss, R., Wilhelm, C., Zauner, C., Maier, U. G., Grossman, A. R., Bhattacharya, D. & Lohr, M. 2008 Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Mol. Biol. Evol.* **25**, 2653–2667. (doi:10.1093/molbev/msn206)

Glansdorff, N., Xu, Y. & Labedan, B. 2008 The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol. Direct.* **3**, 29. (doi:10.1186/1745-6150-3-29)

Gould, S. B., Waller, R. R. & McFadden, G. I. 2008 Plastid evolution. *Annu. Rev. Plant. Biol.* **59**, 491–517. (doi:10.1146/annurev.arplant.59.032607.092915)

Gribaldo, S. & Brochier-Armanet, C. 2006 The origin and evolution of Archaea: a state of the art. *Phil. Trans. R. Soc. B* **361**, 1007–1022. (doi:10.1098/rstb.2006.1841)

Gross, J., Meurer, J. & Bhattacharya, D. 2008 Evidence of a chimeric genome in the cyanobacterial ancestor of plastids. *BMC Evol. Biol.* **8**, 117. (doi:10.1186/1471-2148-8-117)

Hansmann, S. & Martin, W. 2000 Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes. *Int. J. Syst. Evol. Microbiol.* **50**, 1655–1663.

Harary, F. 1969 *Graph Theory.* Reading, MA: Perseus Books Publishing L. L. C.

Hazkani-Covo, E. & Covo, S. 2008 *Numt*-mediated double strand break repair mitigates deletions during primate genome evolution. *PLoS Genet.* **4**, e1000237. (doi:10.1371/journal.pgen.1000237)

Hempel, F., Bozarth, A., Sommer, M. S., Zauner, S., Przyborski, J. M. & Maier, U. G. 2007 Transport of nuclear-encoded proteins into secondarily evolved plastids. *Biol. Chem.* **388**, 899–906. (doi:10.1515/BC.2007.119)

Huang, J. L. & Gogarten, J. P. 2006 Ancient horizontal gene transfer can benefit phylog phylogenetic enetic reconstruction. *Trends Genet.* **22**, 361–366. (doi:10.1016/j.tig.2006.05.004)

Huang, C. Y., Ayliffe, M. A. & Timmis, J. N. 2003 Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature* **422**, 72–76. (doi:10.1038/nature01435)

Huang, C. Y., Gruenheit, N., Ahmadinejad, N., Timmis, J. N. & Martin, W. 2005 Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiol.* **138**, 1723–1733. (doi:10.1104/pp.105.060327)

Huson, D. H. & Bryant, D. 2006 Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267. (doi:10.1093/molbev/msj030)

Kalanon, M. & McFadden, G. I. 2008 The chloroplast protein translocation complexes of *Chlamydomonas reinhardtii*: a bioinformatic comparison of Toc and Tic components in plants, green algae and red algae. *Genetics* **179**, 95–112. (doi:10.1534/genetics.107.085704)

Keeling, P. J. & Palmer, J. D. 2008 Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* **9**, 605–618. (doi:10.1038/nrg2386)

Kennedy, S. P., Ng, W. V., Salzberg, S. L., Hood, L. & DasSarma, S. 2001 Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Res.* **11**, 1641–1650. (doi:10.1101/gr.190201)

Knoll, A. H., Javaux, E. J., Hewitt, D. & Cohen, P. 2006 Eukaryotic organisms in Proterozoic oceans. *Phil. Trans. R. Soc. B* **361**, 1023–1038. (doi:10.1098/rstb.2006.1843)

Koonin, E. V. & Martin, W. 2005 On the origin of genomes and cells within inorganic compartments. *Trends Genet.* **21**, 647–654. (doi:10.1016/j.tig.2005.09.006)

Kunin, V., Goldovsky, L., Darzentas, N. & Ouzounis, C. A. 2005 The net of life: reconstructing the microbial phylogenetic network. *Genome Res.* **15**, 954–959. (doi:10.1101/gr.3666505)

Kurland, C. G., Canback, B. & Berg, O. G. 2007 The origins of modem proteomes. *Biochimie* **89**, 1454–1463. (doi:10.1016/j.biochi.2007.09.004)

Lake, J. A., Skophammer, R. G., Herbold, C. W. & Servin, J. A. 2009 Genome beginnings: rooting the tree of life. *Phil. Trans. R. Soc. B* **364**, 2177–2185. (doi:10.1098/rstb.2009.0035)

Landan, G. & Graur, D. 2007 Heads or tails: A simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.* **24**, 1380–1383. (doi:10.1093/molbev/msm060)

Landan, G. & Graur, D. 2008 Characterization of pairwise and multiple sequence alignment errors. *Gene*, PMID: 18614299. Epub ahead of print.

Lane, C. E. & Archibald, J. M. 2008 The eukaryotic tree of life: endosymbiosis takes its TOL. *Trends. Ecol. Evol.* **23**, 268–275. (doi:10.1016/j.tree.2008.02.004)

Lang, A. S. & Beatty, J. T. 2007 Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol.* **15**, 54–62. (doi:10.1016/j.tim.2006.12.001)

Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. 2008 Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* **25**, 762–777. (doi:10.1093/molbev/msn023)

Loftus, B. *et al.* 2005 The genome of the protist parasite *Entamoeba histolytica*. *Nature* **433**, 865–868. (doi:10.1038/nature03291)

Margulis, L., Chapman, M. & Dolan, M. F. 2007 Semes for analysis of evolution: de Duve's peroxisomes and Meyer's hydrogenases in the sulphurous Proterozoiceon. *Nat. Rev. Genet.* **8**, 1. (doi:10.1038/nrg2071-c1)

Martin, W. 1999 Mosaic bacterial chromosomes—a challenge en route to a tree of genomes. *BioEssays* **21**, 99–104. (doi:10.1002/(SICI)1521-1878(199902)21:2<99::AID-BIES3>3.0.CO;2-B)

Martin, W. & Russell, M. 2003 On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Phil. Trans. R. Soc. Lond. B* **358**, 59–85. (doi:10.1098/rstb.2002.1183)

Martin, W., Brinkmann, H., Savona, C. & Cerff, R. 1993 Evidence for a chimaeric nature of nuclear genomes: Eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. *Proc. Natl Acad. Sci. USA* **90**, 8692–8696. (doi:10.1073/pnas.90.18.8692)

Martin, W., Stoebe, B., Goremykin, V., Hansmann, S., Hasegawa, M. & Kowallik, K. V. 1998 Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* **393**, 162–165. (doi:10.1038/30234)

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M. & Penny, D. 2002 Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl Acad. Sci. USA* **99**, 12 246–12 251. (doi:10.1073/pnas.182432999)

Martin, W., Dagan, T., Koonin, E. V., Dipippo, J. L., Gogarten, J. P. & Lake, J. A. 2007 The evolution of eukaryotes. *Science* **316**, 542–543.

Martin, W., Baross, J., Kelley, D. & Russell, M. J. 2008 Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* **6**, 805–814. (doi:10.1038/nrmicro1991)

McInerney, J. O., Cotton, J. A. & Pisani, D. 2008 The prokaryotic tree of life: past, present ... and future? *Trends Ecol. Evol.* **23**, 276–281. (doi:10.1016/j.tree.2008.01.008)

Mongodin, E. F. *et al.* 2005 The genome of *Salinibacter ruber*: Convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proc. Natl Acad. Sci. USA* **102**, 18 147–18 152. (doi:10.1073/pnas.0509073102)

Müller, M. 2007 The road to hydrogenosomes. In *Origin of Mitochondria and Hydrogenosomes.* (eds W. F. Martin & M. Müller), pp. 1–12. Heidelberg, Germany: Springer-Verlag.

Nisbet, E. G. & Sleep, N. H. 2001 The habitat and nature of early life. *Nature* **409**, 1083–1091. (doi:10.1038/35059210)

Orgel, L. E. 2008 The implausibility of metabolic cycles on the prebiotic earth. *PLoS Biol.* **6**, e18. (doi:10.1371/journal.pbio.0060018)

Pisani, D., Cotton, J. A. & McInerney, J. O. 2007 Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol. Biol. Evol.* **24**, 1752–1760. (doi:10.1093/molbev/msm095)

Puthiyaveetil, S. *et al.* 2008 The ancestral symbiont sensor kinase CSK links photosynthesis with gene expression in chloroplasts. *Proc. Natl Acad. Sci. USA* **105**, 10 061–10 066. (doi:10.1073/pnas.0803928105)

Rasmussen, B., Fletcher, I. R., Brocks, J. J. & Kilburn, M. R. 2008 Reassessing the first occurrence of eukaryotes and cyanobacteria. *Nature* **455**, 1101–1105. (doi:10.1038/nature07381)

Rivera, M. C. & Lake, J. A. 2004 The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**, 152–155. (doi:10.1038/nature02848)

Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. 1998 Genomic evidence for two functionally distinct gene classes. *Proc. Natl Acad. Sci. USA* **95**, 6239–6244. (doi:10.1073/pnas.95.11.6239)

Rumpho, M. E., Worful, J. M., Lee, J., Kannan, K., Tyler, M. S., Bhattacharya, D., Moustafa, A. & Manhart, J. R. 2008 Horizontal gene transfer of the algal nuclear gene psbO to the photosynthetic sea slug *Elysia chlorotica*. *Proc. Natl Acad. Sci. USA* **105**, 17 867–17 871. (doi:10.1073/pnas.0804968105)

Salzberg, S. L., White, O., Peterson, J. & Eisen, J. A. 2001 Microbial genes in the human genome: lateral transfer or gene loss? *Science* **292**, 1903–1906. (doi:10.1126/science.1061036)

Sanchez-Puerta, M. V. & Delwiche, C. F. 2008 A hypothesis for plastid evolution in chromalveolates. *J. Phycol.* **44**, 1097–1107. (doi:10.1111/j.1529-8817.2008.00559.x)

Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. 2006 Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**, 341–345. (doi:10.1038/nature04562)

Shavit, L., Penny, D., Hendy, M. D. & Holland, B. R. 2007 The problem of rooting rapid radiations. *Mol. Biol. Evol.* **24**, 2400–2411. (doi:10.1093/molbev/msm178)

Stanhope, M. J., Lupas, A., Italia, M. J., Koretke, K. K., Volker, C. & Brown, J. R. 2001 Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* **411**, 940–944. (doi:10.1038/35082058)

Stoebe, B. & Maier, G. 2002 One, two, three: nature's toolbox for building plastids. *Protoplasma* **219**, 123–130. (doi:10.1007/s007090200013)

Theissen, U. & Martin, W. 2006 The difference between organelles and endosymbionts. *Curr. Biol.* **16**, R1016–R1017. (doi:10.1016/j.cub.2006.11.020)

Thorsness, P. E. & Fox, T. D. 1990 Escape of DNA from the mitochondria to the nucleus in the yeast *Saccharomyces cerevisiae*. *Nature* **346**, 376–379. (doi:10.1038/346376a0)

Thomas, C. M. & Nielsen, K. M. 2005 Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721. (doi:10.1038/nrmicro1234)

Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. 2004 Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **5**, 123–135. (doi:10.1038/nrg1271)

Ueno, Y., Yamada, K., Yoshida, N., Maruyama, S. & Isozaki, Y. 2006 Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. *Nature* **440**, 516–519. (doi:10.1038/nature04584)

van der Giezen, M., Tovar, J. & Clark, C. G. 2005 Mitochondrion-derived organelles in protists and fungi. *Int. Rev. Cytol.* **244**, 175–225.

von Dohlen, C. D., Kohler, S., Alsop, S. T. & McManus, W. R. 2001 Mealybug β-proteobacterial endosymbionts contain γ-proteobacterial symbionts. *Nature* **412**, 433–436. (doi:10.1038/35086563)

Welch, R. A. *et al.* 2002 Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **99**, 17 020–17 024. (doi:10.1073/pnas.252529799)

Wujek, D. E. 1979 Intracellular bacteria in the blue-green-alga *Pleurocapsa minor*. *Trans. Am. Micros. Soc.* **98**, 143–145. (doi:10.2307/3225953)

6 Endosymbiotic gene transfer at the origin of eukaryotes.

6.1 A fluid prokaryotic chromosome model

Esser C, Martin W, Dagan T: **The origin of mitochondria in light of a fluid prokaryotic chromosome model**. *Biol Lett* **3**:180-184.

(Own contribution: analyzed the data, and wrote the paper).

biology letters
Genome biology

Gray et al. 1999 Dolezal et al. 2006

John & Whatley 1975 Stackebrandt et al. 1988

Yang et al. 1985

Lang et al. 1999 Emelyanov 2003

Andersson et al. 1998

Wu et al. 2004

http://dx.doi.org/10.1098/rsbl.2006.0582 http://www.journals.royalsoc.ac.uk

Esser et al. 2004 Gabaldon & Huynen (2003)

Müller 2003 van der Giezen & Tovar 2005 Embley & Martin 2006

Lang et al. 1999 Esser et al. 2004

Lawrence & Ochman 1998 Martin 1999 Doolittle 2004 Kunin et al. 2005 Lerat et al. 2005

http://www.ncbi.nlm.nih.gov/

Altschul et al. 1990

Saitou & Nei 1987 Koski & Golding 2001 Penny et al. 2001

http://www.ncbi.nlm.nih.gov/Taxonomy/

Thompson et al. 1994 Felsenstein 2005

Felsenstein 2005 Pupko et al. 2000

figure 1

FirstCite®
e-publishing

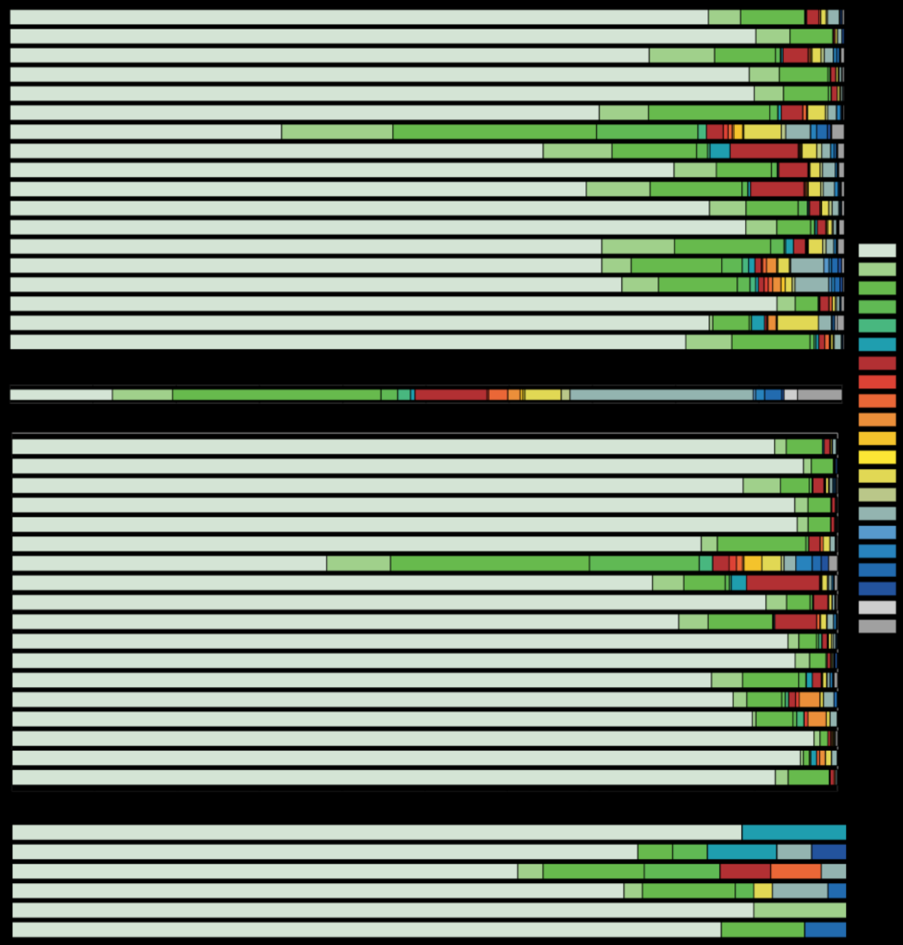


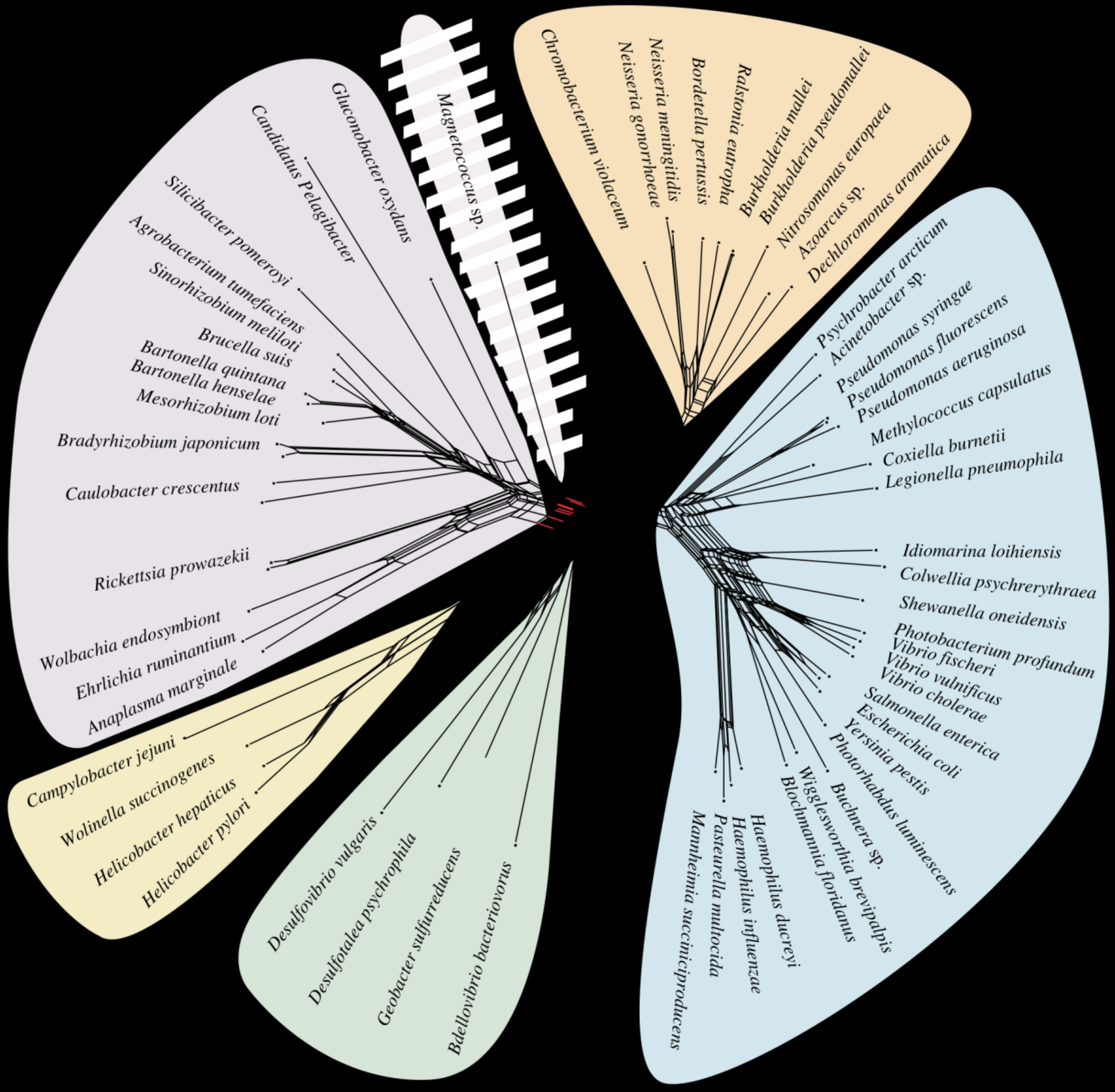
figure 2 figure 1

figure 1

figure 1

figure 1

figure 1



Huson & Bryant 2006

Gabaldon & Huynen
2003

Esser *et al.* 2004 Rivera & Lake 2004
Embley & Martin 2006 Boussau *et al.* 2004

Baughn & Malamy 2002

Susko *et al.* 2006

figure 1

Doolittle 2004 Kunin *et al.* 2005 Lerat *et al.* 2005

Esser *et al.* 2004

Kurland & Andersson 2000

Embley & Martin 2006

doi:10.1006/jmbi.1990.9999

doi:10.1038/24094

doi:10.1073/pnas.052710199

doi:10.1073/pnas.0400975101

doi:10.1126/science.1127895

doi:10.1038/nature04546

doi:10.1016/j.abb.2003.09.031

doi:10.1093/molbev/msh160

doi:10.1126/science.1085463

doi:10.1126/science.283.5407.1476

doi:10.1093/molbev/msj030

doi:10.1038/254495a0

doi:10.1101/gr.3666505

doi:10.1128/MMBR.64.4.786-820.2000

doi:10.1146/annurev.genet.33.1.351

doi:10.1073/pnas.95.16.9413

doi:10.1371/journal.pbio.0030130

doi:10.1002/(SICI)1521-1878(199902)21:2<99::AID-BIES3>3.0.CO;2-B

doi:10.1007/s002390010258

doi:10.1038/nature02848

doi:10.1093/molbev/msj113

doi:10.1038/sj.embor.7400440

doi:10.1371/journal.pbio.0020069

doi:10.1073/pnas.82.13.4443

6.2 A heterocyst-forming plastid ancestor

Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T: **Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor**. *Mol Biol Evol* 2008, **25**:748-761.

(Own contribution: designed the experiment, performed part of the analysis, analyzed the data, and wrote the paper).

Genes of Cyanobacterial Origin in Plant Nuclear Genomes Point to a Heterocyst-Forming Plastid Ancestor

Oliver Deusch, Giddy Landan,† Mayo Roettger,* Nicole Gruenheit,* Klaus V. Kowallik,* John F. Allen,‡ William Martin,* and Tal Dagan**

*Institut für Botanik III, Heinrich-Heine Universität Düsseldorf, Universitätsstrasse 1, Düsseldorf, Germany; †Department of Biology and Biochemistry, University of Houston; and ‡School of Biological and Chemical Sciences, Queen Mary, University of London, London, United Kingdom

Plastids are descended from a cyanobacterial symbiosis which occurred over 1.2 billion years ago. During the course of endosymbiosis, most genes were lost from the cyanobacterium's genome and many were relocated to the host nucleus through endosymbiotic gene transfer (EGT). The issue of how many genes were acquired through EGT in different plant lineages is unresolved. Here, we report the genome-wide frequency of gene acquisitions from cyanobacteria in 4 photosynthetic eukaryotes—*Arabidopsis*, rice, *Chlamydomonas*, and the red alga *Cyanidioschyzon*—by comparision of the 83,138 proteins encoded in their genomes with 851,607 proteins encoded in 9 sequenced cyanobacterial genomes, 215 other reference prokaryotic genomes, and 13 reference eukaryotic genomes. The analyses entail 11,569 phylogenies inferred with both maximum likelihood and Neighbor-Joining approaches. Because each phylogenetic result is dependent not only upon the reconstruction method but also upon the site patterns in the underlying alignment, we investigated how the reliability of site pattern generation via alignment affects our results: if the site patterns in an alignment differ depending upon the order in which amino acids are introduced into multiple sequence alignment—N- to C-terminal versus C- to N-terminal—then the phylogenetic result is likely to be artifactual. Excluding unreliable alignments by this means, we obtain a conservative estimate, wherein about 14% of the proteins examined in each plant genome indicate a cyanobacterial origin for the corresponding nuclear gene, with higher proportions (17–25%) observed among the more reliable alignments. The identification of cyanobacterial genes in plant genomes affords access to an important question: From which type of cyanobacterium did the ancestor of plastids arise? Among the 9 cyanobacterial genomes sampled, *Nostoc* sp. PCC7120 and *Anabaena variabilis* ATCC29143 were found to harbor collections of genes which are—in terms of presence/absence and sequence similarity—more like those possessed by the plastid ancestor than those of the other 7 cyanobacterial genomes sampled here. This suggests that the ancestor of plastids might have been an organism more similar to filamentous, heterocyst-forming (nitrogen-fixing) representatives of section IV recognized in Stanier's cyanobacterial classification. Members of section IV are very common partners in contemporary symbiotic associations involving endosymbiotic cyanobacteria, which generally provide nitrogen to their host, consistent with suggestions that fixed nitrogen supplied by the endosymbiont might have played an important role during the origin of plastids.

Introduction

The idea that plastids arose from cyanobacteria through endosymbiosis is old and was scorned for many decades (Mereschkowsky 1905) but is no longer debated (Douglas 1998; Delwiche 1999; Matsuzaki et al. 2004; McFadden and van Dooren 2004; Archibald 2006). However, the issue of how, exactly, cyanobacteria contributed to plant genome evolution is still unresloved, as is the issue of what kind of cyanobacterium participated in plastid origin (Sato 2006). As an evolutionary mechanism, endosymbiosis differs substantially from point mutation because a genome's worth of new genetic material is the currency unit of genetic change, rather than a succession of fixed nucleotide polymorphisms or duplicated genes with mutated promoters. Accordingly, the evolutionary transition that transformed an oxygen-producing, prokaryotic endosymbiont into the diverse spectrum of photosynthetic organelles found among modern photosynthetic eukaryotes involved both inheritance from the prokaryote and invention by the eukaryote (Sato 2001).

From the standpoint of cell function and physiology at plastid origin, the most important inheritance was photosynthesis itself, which conferred the photolithoautotrophic lifestyle upon a heterotrophic host, whereas the most important invention was the protein import machinery (McFadden and van Dooren 2004; Soll and Schleiff 2004), which permitted the endosymbiont to import nuclear-encoded proteins. The invention of a protein import apparatus allowed the ancestral plastid to relinquish genes to the nucleus over evolutionary time without relinquishing those biochemical functions which are germane to the cyanobacterial lifestyle (Allen 2003). Protein import thus marked a crucial turning point in the evolutionary process that gave rise to plastids. Prior to the invention of protein import, the cyanobacterial endosymbiont was able to donate genes to its host, but unable to import the encoded products, such that, from the host's standpoint, the symbiont served as a virtually inexhaustible source of new and divergent genes for functions in various cell compartments (Martin and Schnarrenberger 1997; Martin and Herrmann 1998; Allen 2003; Bogorad 2008). Once the protein import apparatus had evolved, products of nuclear genes of cyanobacterial origin could be targeted to the plastid compartment, allowing the endosymbiont-encoded copies of such genes to escape purifying selection, and thus undergo pseudogenization and loss. The process just described, endosymbiotic gene transfer (EGT; Martin et al. 1993), resulted in the genetic integration of the endosymbiont with its host (Timmis et al. 2004; Sato 2006; Reyes-Prieto et al. 2007), accompanied by the transition of the former into a double membrane–bound organelle of the latter and by the origin of the eukaryotic lineage that

Key words: endosymbiosis, phylogenomics, multiple sequence alignment, plastid origin, nitrogen.

E-mail: tal.dagan@uni-duesseldorf.de.

Mol. Biol. Evol. 25(4):748–761. 2008
doi:10.1093/molbev/msn022
Advance Access publication January 24, 2008

possesses primary plastids, the archaeplastida (Adl et al. 2005). In the present work, we aim to address 2 questions.

First, we wish to address the quantitative scope of EGT from plastids using whole genome data. Modern plastids are estimated to contain anywhere from about 2,100 to 4,800 different proteins (Richly and Leister 2004), whereas plastid genomes encode only 60–200 proteins in various photosynthetic lineages (Timmis et al. 2004). Plastids thus contain roughly as many proteins as their free-living cyanobacterial cousins but have retained only a handful of the corresponding genes (Allen and Raven 1996). Various and differing estimates for the total number of genes that were acquired by plants from cyanobacteria have been reported. A phylogenomic investigation of *Arabidopsis* including 3 cyanobacterial genomes but only 1 reference eukaryotic genome suggested that about 18% of *Arabidopsis* proteins are cyanobacterial acquisitions (Martin et al. 2002), whereas an EST-based analysis of *Cyanophora paradoxa* indicated that about 11% of the proteins in that genome are cyanobacterial acquisitions (Archibald 2006; Reyes-Prieto et al. 2006).

Those differences are likely to have methodological causes, and it is well recognized that the reliability of phylogenetic inference with highly divergent proteins can affect such estimates (Martin et al. 2002; Reyes-Prieto et al. 2006). Comparisons of plant and cyanobacterial proteins are replete with alignments of highly divergent sequences because the origin of plastids dates back at least 1.2 billion years (Butterfield 2000; Yoon et al. 2004). In order to take the influence of highly divergent alignments into account in this phylogenomic study, we have made use of recent findings showing that phylogenetic results using real data differ markedly in comparisons of alignments generated by reading sequences from N- to C-terminus (the default in all current alignment programs, the "heads" orientation) to those generated using the same program but reading sequences in from C-terminus to N-terminus (the "tails" orientation) (Landan and Graur 2007). The underlying reasoning is straightforward: If a phylogenetic result is contingent upon the order in which amino acids are introduced into multiple sequence alignment—N- to C-terminal versus C- to N-terminal or heads versus tails—then it is likely to be an artifact of phylogeny inference and one that is rooted at the alignment step, where the site patterns for phylogenetic inference are generated. The reproducibility of site pattern generation as a function of heads versus tails alignment provides a criterion for separating phylogenetic wheat from chaff. Its utility to refine estimates for the number of genes that plants acquired from cyanobacteria is explored in a phylogenomic analysis of proteins encoded by *Arabidopsis*, rice, *Chlamydomonas*, and the red alga *Cyanidioschyzon* in comparison with 9 cyanobacterial and 228 other reference genomes.

Second, we wish to investigate the nature of the plastid ancestor, specifically which genes it contained and to which lineages of modern cyanobacteria it was most closely related as inferred from cyanobacterial genes present in plant nuclear genomes. Previous studies using alignments of single loci (Morden and Golden 1989; Turner et al. 1999; Marin et al. 2005) or concatenated genes encoded in plastid DNA in comparison to homologues from cyanobacteria (Rodriguez-Ezpeleta et al. 2005) have been inconclusive, although Sato (2006) suggested that the *Anabaena–Synechocystis* lineage might be closer to the ancestor of plastids than other cyanobacteria sampled in that study. However, the individual phylogenies of plastid-encoded genes can differ significantly even though they are related by the same evolutionary process (Martin et al. 1998; Lockhart et al. 1999), a prime example of the limitations involved in phylogenetic inference as one goes farther back in time (White et al. 2007). But among cyanobacteria the situation is worse because in addition to the problem of phylogenetic error, they can and do exchange genes both among themselves and with other bacterial groups via lateral gene transfer (LGT) (Raymond et al. 2002; Zhaxybayeva et al. 2006), such that concatenation will tend to mix phylogenetic signals, rather than amplify them (Bapteste et al. 2007), with bootstrap or Bayesian support values offering no guide to confidence, because large alignments tend to yield support values close to unity, regardless of whether or not the topology is correct (Phillips et al. 2004).

As it relates to the origin of genes that eukaryotes acquired from the ancestor of plastids, LGT among prokaryotes means that—even though all available evidence points to a single origin of plastids (McFadden and van Dooren 2004; Archibald 2006) and notwithstanding recent debate as to how to define the term plastid (Bodyl et al. 2007; Larkum et al. 2007)—plant genomes harbor a discrete sample of the collection of genes present in the genome of the plastid ancestor. However, neither that gene sample nor the collection of genes present in the free-living plastid ancestor is likely to have persisted in its original, contiguous state in any modern cyanobacterial chromosome (Martin 1999) because of LGT among cyanobacteria (Zhaxybayeva et al. 2006) and between cyanobacteria and other prokaryotes (Raymond et al. 2002) since the origin of plastids. The same problem exists with regard to finding the ancestor of mitochondria among α-proteobacteria (Esser et al. 2007). Despite this complication that LGT introduces with regard to identifying the "lineage" of cyanobacteria from which plastids arose (for a discussion of modern concepts regarding prokaryotic lineages, see Doolittle and Bapteste 2007), one can still address the question of which modern cyanobacterial genomes possess the highest frequencies of genes with strongest sequence similarity to their nuclear encoded homologues in photosynthetic eukaryotes, which we do, using 9 sequenced cyanobacterial genomes.

Materials and Methods

Data

The nuclear proteomes of *Arabidopsis thaliana* version January 2006 and *Oryza sativa* version May 2006 were downloaded from RefSeq database (Pruitt et al. 2005). The nuclear proteomes of *Cyanidioschyzon merolae* (Matsuzaki et al. 2004) version February 2005 and *Chlamydomonas reinhardtii* (Merchant et al. 2007) version 2.0 were downloaded from their genome projects. Multiple copy proteins were condensed into a single entry. Homologous proteins within prokaryotes and nonphotosynthetic eukaryotes were searched as follows. The proteins of the photosynthetic eukaryotes were Blasted (Altschul et al. 1997) to a data set including 200 eubacteria, 24 archaebacteria, and 13 nonphotosynthetic eukaryotes (supplementary table S1, Supplementary Material online). The Blast hits were filtered for hits of *E* value $\leq 10^{-10}$ and $\geq 25\%$ amino acid identities and ranked by the percent of identities multiplied by the ratio of the query length and the Blast pairwise alignment length. In order to render phylogenetic analysis using maximum likelihood (ML) computationally tractable for over >11,000 phylogenies, the number of operational taxonomical units per alignment had to be restricted, hence only the first 3 hits from each phylum were selected for sequence alignment. Because we are addressing the question of which proteins are most closely related to the plant homologues, rather than the question of how proteins from all lineages are related to one another, the selection of the 3 best representatives from each phylum should not influence our estimates. Proteins from the cyanobacterial genomes were investigated (*Anabaena variabilis* ATCC29413, *Gloeobacter violaceus* PCC7421, *Nostoc* sp. PCC7120, *Prochlorococcus marinus* MIT9313, *Synechococcus* sp. CC9605, *Synechococcus elongatus* PCC7942, *Synechococcus* sp. WH8102, *Synechocystis* sp. PCC6803, and *Thermosynechococcus elongatus* BP-1), and other genomes were obtained from GenBank.

Sequence Alignment

Each protein was aligned with its homologues using Muscle (Edgar 2004) with a maximum of 16 iterations. For the tails alignment, the protein sequences were reversed using a PERL script and were aligned again with Muscle using the same parameters. For the comparison between the heads and tails alignments, the tails alignment is reversed and each amino acid is replaced by its serial number in the protein sequence. Gaps in the alignment are converted to “0.” An identical heads and tails column includes exactly the same amino acids from each sequence in both alignments. The columns score (CS) is calculated as the proportion of columns in the heads alignment that had a matching column in the tails alignment. An identical heads and tails pair is a pair of amino acids in 2 different sequences that are aligned together in both alignments. The sum-of-pairs score (SPS) is calculated as the proportion of pairs in the heads alignment that had a matching pair in the tails alignment.

Phylogenetic Trees

The heads and tails alignments were used to reconstruct phylogenetic trees with Phyml (Guindon and Gascuel 2003) using the ungapped positions only and the Jones-Taylor-Thornton (JTT) model (Jones et al. 1992) assuming rate variation across sites according to a Gamma distribution with 8 rate categories with the alpha parameter estimated from the data and invariable sites taken into account and with Neighbor-Joining (NJ) (Saitou and Nei 1987) using JTT distances. The heads and tails trees were compared by counting the identical splits between the trees using Treedist of PHYLIP (Felsenstein 2005). The phylogenetic partitions score (PPS) is the proportion of tree splits (branches) that were reconstructed identically from the heads and tails alignments. The nearest neighbors were identified as taxa contained in the smallest clade that included the homologue of the photosynthetic eukaryote.

Pairwise Analysis

Reciprocal best Blast hits (BBHs) of *Arabidopsis* genes and rice or alternatively *Chlamydomonas* genes were defined as orthologous pairs. Each pair of genes was aligned using ClustalW (Thompson et al. 1994), and protein distances were calculated with Protdist (Felsenstein 2005) using the JTT substitution matrix.

Nuclear Plastid DNA Rescreening

To identify nuclear plastid DNAs (NUPTs) (Richly and Leister 2004) protein-coding sequences from the corresponding chloroplast genomes were Blasted to the RefSeq database version January 2008. Proteins with 100% identical amino acids over the total length were scored as probable NUPTs.

Functional Classification

Proteins were classified into functional categories according to their BBH above 25% amino acid identities found in Swiss-Prot database (Boeckmann et al. 2003). The function of each protein in Swiss-Prot is described by one or more keywords. We manually classified Swiss-Prot keywords into functional categories (the list is available upon request). Each Swiss-Prot protein was assigned to a functional category according to the most frequent category of its keywords.

Results and Discussion

In order to identify genes of cyanobacterial origin in the genomes of 4 photosynthetic eukaryotes containing a primary plastid, we compared the nonredundant set of 83,138 proteins encoded in their genomes to a data set of 851,607 proteins from 9 sequenced cyanobacterial genomes, 13 reference eukaryotic genomes, and 215 reference prokaryotic genomes using Blast. Alignments were constructed for those query sequences that detected at least one homologue in cyanobacteria and homologues in at least 2 other search phyla (supplementary table S1, Supplementary Material online) at an *E*-value threshold 10^{-10} and 25% amino acid identity in the pairwise Blast alignment. Within each phylum, hits were ranked by their similarity and the ratio of the hit length to the query length. The first 3 hits from each phylum were selected for alignment and phylogenetic analysis. A summary of the distribution of hits at this threshold for each phylum is shown in table 1. For each query, we thus obtained a set of sequences that included the best cyanobacterial matches and the best matches from at least 2 other phyla to address the cyanobacterial ancestry of the plant homologue via phylogenetic inference. For those

Table 1
The number of Blast hits for each phylum and the number of trees in which the phylum was included

		Arabidopsis		Rice		*Chlamydomonas*		*Cyanidioschyzon*	
	Phylum (number of genomes)	Hits	Trees	Hits	Trees	Hits	Trees	Hits	Trees
Eubacteria	*Cyanobacteria* (9)	5,199	4,670	3,524	3,186	2,767	2,500	1,363	1,213
	Actinobacteria (17)	5,265	3,760	3,934	2,617	3,437	2,143	1,298	1,056
	Aquificae (1)	1,550	1,387	911	811	857	792	603	564
	Bacteroidetes (4)	3,116	2,431	2,024	1,558	1,702	1,392	988	849
	Chlamydiae (6)	1,884	1,638	1,124	982	946	828	2,126	858
	Chlorobi (3)	2,264	2,105	1,404	1,283	1,325	1,204	813	758
	Chloroflexi (2)	1,232	1,157	882	691	874	742	485	464
	Deinococcus–Thermus (2)	2,598	2,158	1,912	1,455	2,061	1,478	890	770
	Firmicutes (45)	5,188	3,545	3,402	2,412	2,608	1,860	1,214	996
	Fusobacteria (1)	1,328	1,200	742	675	773	700	481	441
	Planctomycetes (1)	2,730	2,177	1,882	1,479	1,483	1,266	788	705
	Proteobacteria (105)	6,857	4,378	4,958	3,009	4,385	2,392	1,585	1,160
	Spirochaetes (5)	2,766	2,317	1,633	1,380	1,415	1,211	851	735
	Thermotogae (1)	1,615	1,368	952	817	854	757	582	530
Archaeabacteria	*Crenarchaeota* (5)	1,941	1,366	1,175	815	974	694	693	491
	Euryarchaeota (18)	4,100	2,902	2,716	1,911	1,982	1,450	1,122	801
	Nanoarchaeota (1)	429	208	248	122	217	131	202	106
Eukaryotes	*Ascomycota* (8)	10,734	3,850	7,853	2,624	4,881	1,819	2,303	922
	Basidiomycota (2)	10,047	3,545	7,685	2,497	4,151	1,619	2,126	858
	Microsporidia (1)	2,849	836	1,594	516	1,182	415	769	284
	Protists (2)	7,187	2,158	4,432	1,509	3,155	1,139	1,415	525

data sets, alignments were constructed and phylogenetic trees were inferred using maximum likelihood and NJ.

On average, about 14% of those archeaplastidan proteins with homologues in cyanobacteria and at least 2 other phyla branched as nearest neighbor to cyanobacterial homologues. However, among the gene trees investigated, we found that almost every phylum sampled appeared as a nearest neighbor of the query species (table 2), which can be due to phylogenetic error or LGT among prokaryotes. Contrasted to the proportion of the proteins from each phylum in our database, the cyanobacterial signal is substantial. For example, 13% of the trees deliver cyanobacterial nearest neighbors in *Arabidopsis*, yet only 3.8% of the genes in our database reside in cyanobacterial chromosomes: the proportions of cyanobacterial nearest neighbors in plant genomes are far above random similarities. If we divide the proportion of nearest neighbors for each phylum by its gene content in our data set, then we find that the next largest proportion of plant gene nearest neighbors stems from the eubacterial genomes in our sample, with a majority of proteobacterial genes (fig. 1 and table 2). These could be attributed either to genes that were acquired from the mitochondrion ancestor, biased phylogenetic signals, and/or the fluid nature of bacterial genomes over time (Esser et al. 2007).

EGT Inference Depends upon Sequence Conservation

The foregoing first approximations for the percentage of cyanobacterial acquisitions represent only the fraction of genes in each archaeplastidan genome for which we can construct trees with at least 4 taxa at the phylum level. They exclude genes with a more narrow distribution across prokaryotic phyla (many photosystem genes, for example) and those that do not fulfill our criteria for inclusion in multiple alignments because of low sequence conservation. Moreover, among those genes with sufficient phyletic distribution and sequence conservation to be included in the tree-building procedure, proteins with higher sequence conservation showed a tendency to display a cyanobacterial origin more often than more divergent proteins did. For example, 22% of the *Arabidopsis* gene trees with a mean branch length of ≤0.2 substitutions per site indicate cyanobacterial origin of the plant nuclear gene, whereas only 6% of the gene trees with mean branch length ≥1 do (fig. 2*a*). The same tendency is observed for rice (fig. 2*b*) and *Chlamydomonas* (fig. 2*c*). This could have either a biological or a methodological cause: 1) genes of cyanobacterial origin in plants might preferentially belong to the most slowly evolving proteins in the genome or 2) because phylogenetic inference is less accurate with more highly diverged

Table 2
The distribution of nearest neighbors of archaeplastidan genes within the different taxa. Mixed clades include members from various taxa

Nearest Neighbor	*Arabidopsis* (%)	*Oryza* (%)	*Chlamydomonas* (%)	*Cyanidioschyzon* (%)	No. proteins
Cyanobacteria	592 (13%)	432 (14%)	356 (14%)	207 (17%)	31,940
Proteobacteria	522 (11%)	308 (10%)	458 (18%)	104 (9%)	360,234
Other eubacteria	882 (19%)	588 (18%)	524 (21%)	257 (21%)	226,314
Archaeabacteria	45 (1%)	38 (1%)	25 (1%)	18 (1%)	56,513
Eukaryotes	2156 (46%)	1498 (47%)	895 (36%)	523 (43%)	176,606
Mixed	473 (10%)	323 (10%)	242 (10%)	106 (9%)	

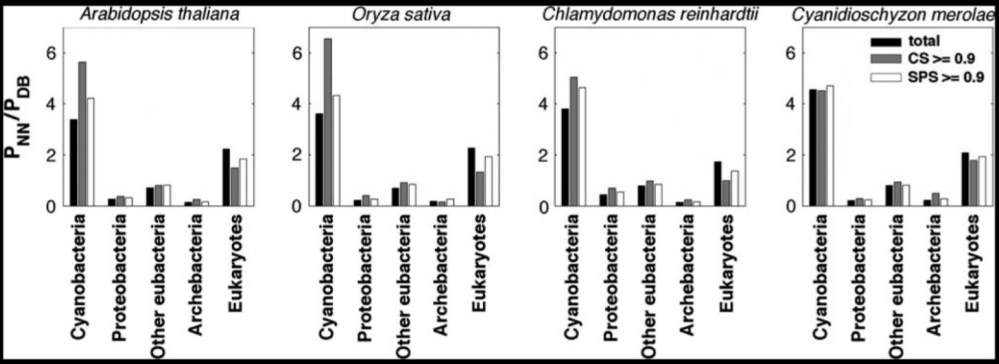


FIG. 1.—The ratio of the proportion of nearest neighbors in different taxa (P_{NN}; table 2) and the proportion of the proteins of the taxa in the database (P_{DB}).

proteins (Nei et al. 1995; Nei 1996), the rate of false negatives—in the present case, trees failing to recover a true cyanobacterial origin of some plant proteins—might increase with increasing sequence divergence.

To test the first possibility, we calculated pairwise protein distances of *Arabidopsis* proteins to their orthologs in rice and *Chlamydomonas* and plotted the proportion of proteins that were inferred as an EGT (or not) as a function of those pairwise distance distributions (fig. 2*e*). Comparison of the protein distances calculated for *Arabidopsis* and rice orthologs showed no significant difference between the protein distance distribution of EGTs and of non-EGT proteins ($P = 0.64$, using Wilcoxon test). The same procedure for *Arabidopsis* and *Chlamydomonas* orthologs showed that the protein distances of EGTs are somewhat smaller than the protein distances of non-EGTs ($P = 0.002$, using Wilcoxon test). Hence, the tendency to infer EGT among the less polymorphic alignments is not clearly attributable to a tendency for EGTs to preferentially occur among the more slowly evolving proteins in the genome, suggesting that phylogenetic inference itself might be preferentially generating an increased frequency of false negatives for EGT candidates among the more highly divergent sequences.

Divergent Proteins Produce Unreliable Alignments

If phylogenetic inference is producing false EGT negatives preferentially for the more highly divergent sequences in our sample, there are, in principle, 2 prime suspects as the possible source of error: Either the tree-building algorithms are failing or the alignments themselves are called into question. It is well known that phylogenetic inference is error prone (Phillips et al. 2004), particularly when sequence divergence exceeds 50% (Nei et al. 1995; Nei 1996), which is very commonly the case in genome-wide phylogenetic studies such as the present one (fig. 2). But the influence that the methodology of sequence alignment itself can have upon phylogeny is less well studied. A phylogenetic tree can hardly be more reliable than the alignment upon which it is based, but objective criteria to assess the quality of alignments in practice are, at best, extremely rare (Kumar and Filipski 2007). We therefore turned our attention to the alignments underlying the phylogenetic analyses in the present study, rather than to the minutiae of tree-building methods themselves, in order to better characterize their influence upon our ability to infer a particular biological process (gene acquisitions from plastids) from contemporary genome data.

For this purpose, we examined the utility of the Heads or Tails (HoT) method (Landan and Graur 2007) that quantifies the dependence of inferred residue homology in an alignment matrix upon the order in which characters are entered into the alignment. In the HoT method, the input sequences are aligned to create a heads (or forward) alignment. For the tails (or reverse) alignment, the input sequences are first reversed, so that the sequences read C- to N-terminus, and aligned independently using the same algorithm and settings, creating a second alignment of the same sequences. The 2 alignments are then compared using the CS (Thompson et al. 1999), which is the proportion of alignment columns that were reconstructed identically in the 2 alignments. When the forward and reverse alignments are nearly identical, the vast majority of columns are reproduced in both alignments and are thus consistent with respect to the character input order and the alignment can be considered reliable (or at least reproducible). But when the CS is low, the alignment is marked as being highly dependent upon an arbitrary variable, namely the order in which the amino acids are subjected to alignment. In this way, both an alignment and trees inferred from it can be identified as questionable or unreliable by virtue of their dependence upon a variable no less arbitrary than a coin toss (Landan and Graur 2007).

When all alignments underlying the trees in figure 2 are examined by HoT analysis, a strong correlation is found in all 4 genome data sets between sequence conservation as estimated by mean branch length in the trees and the proportion of identical columns recovered in the heads and tails alignments (fig. 3). CS among the most conserved sequences (mean branch length ≤ 0.2) ranges between a mean of 0.77 ± 0.17 in *Arabidopsis* and 0.76 ± 0.18 in *Chlamydomonas*, whereas among the most divergent sequences CS falls to a mean of 0.45 ± 0.37 in *Cyanidioschyzon* (mean branch length ≥ 0.8) or 0.11 ± 0.14 in *Arabidopsis* (mean branch length ≥ 1; fig. 3). In other words, for the most highly divergent sequences in the present data, about 80–90% of the columns generated by multiple alignment

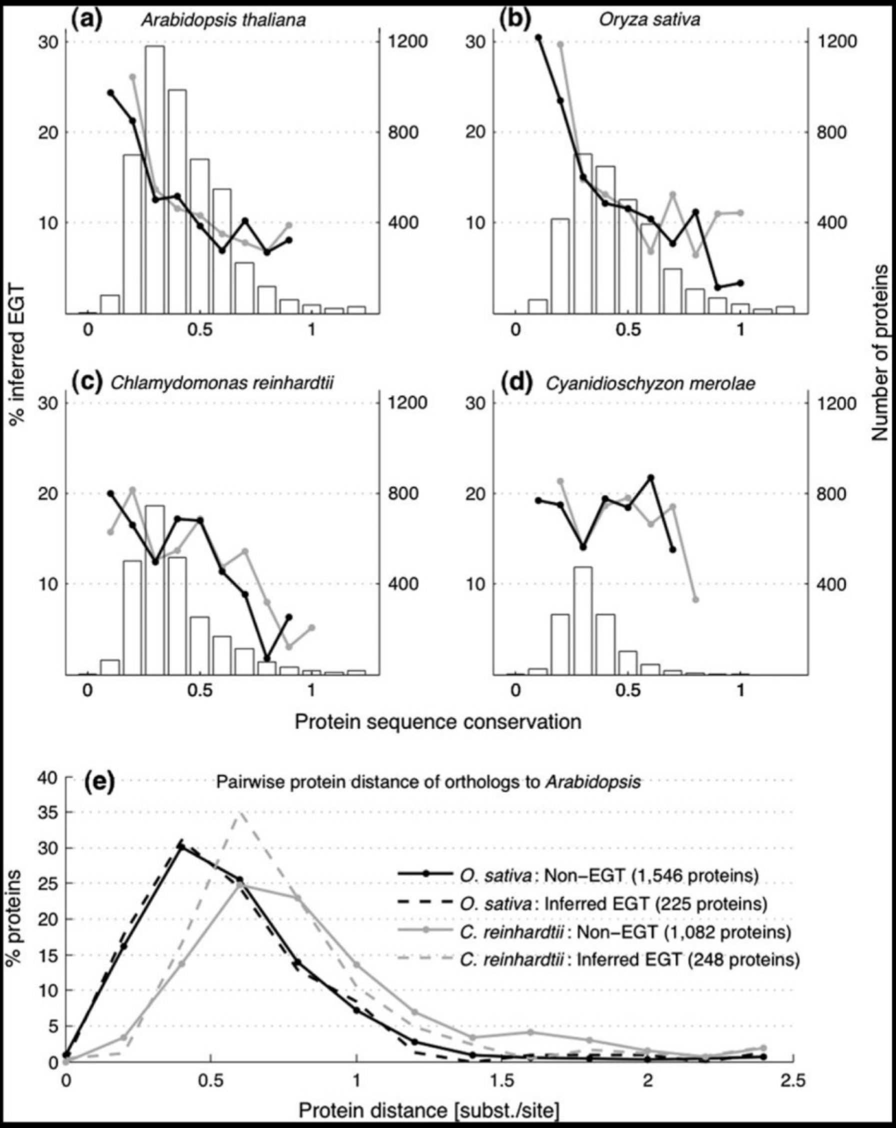


FIG. 2.—(*a–d*) The frequency of EGT as inferred from multiple sequence alignments (MSAs) of varying sequence conservation degrees. The distribution of sequence conservation as calculated by mean branch length of the phylogenetic trees (the sum of branch lengths divided by the number of branches) reconstructed using ML approach is presented in open bars (similar distribution is observed for NJ trees). The frequency of genes inferred as EGT is plotted for both ML tress (black) and NJ trees (gray). (*e*) Distribution of pairwise protein distances for rice and *Chlamydomonas* orthologs of *Arabidopsis* genes.

are dependent upon the order in which the amino acids are aligned. Such columns are irreproducible in the simplest test case and therefore contain unreliable, and possibly misleading, information.

Because EGT inference uses the alignments to reconstruct phylogenetic trees, we used the HoT analysis to construct a separate phylogeny for both the heads and tails alignments of each protein and compared the results. For each sequence set, the heads and the tails alignments were created using Muscle, both were purged of gapped sites, and both were used to infer trees with ML and NJ. The similarity of the heads and tails trees was quantified by the PPS, which is the fraction of internal edges that are common to both trees. The PPS is strongly dependent upon CS for all 4 genome data sets (table 3), and this observation holds for both ML and NJ tree topologies.

The correlation between PPS and CS is positive, except for the *Cyanidioschyzon* data, which probably relates to that genome's small sample size (only 4,762 proteins). But CS is not always a good proxy for PPS in the heads versus tails trees comparison because there are cases (~5% of the data) in which unreliable alignments still produce the same topology. Another measure for alignment reliability using HoT analysis is the comparison of identically reconstructed amino acids pairs (Landan and Graur 2007), in which all of the amino acid pairs between all pairs of sequences are tested for identical reconstruction, yielding a SPS (Thompson et al. 1999). SPS is a less strict measure of alignment uncertainty than CS because if a column differs with regard to only one or a few sequences in the heads or tails comparison, the corresponding HoT columns are scored as proportionately similar using SPS but as

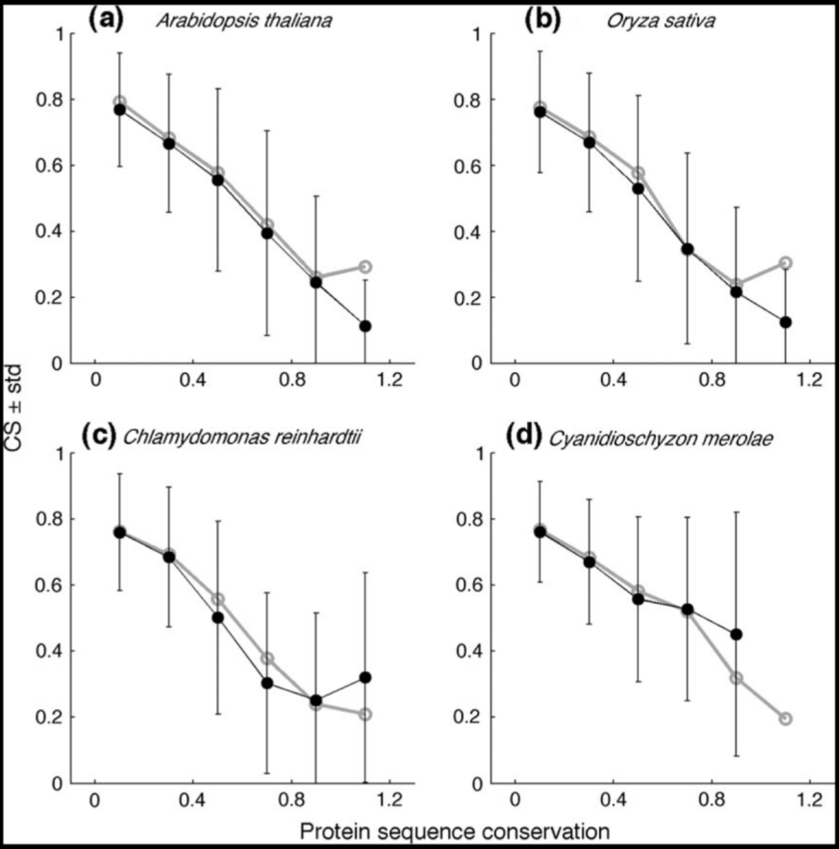


FIG. 3.—Conserved sequences produce more reliable alignments. The mean CS (dots) ± standard deviation is presented for ML trees (black) and NJ trees (gray).

nonidentical (0 of 0/1) using CS. SPS is therefore almost always higher than CS for a given alignment (supplementary fig. 1, Supplementary Material online), but when it is low, the alignment is extremely unreliable because almost all putatively homologous amino acid pairs that are presented to the tree reconstruction procedure are generated in a manner that is dependent upon the heads versus tails parameter of the alignment procedure. Accordingly, SPS correlates better with PPS (table 3), reaching a correlation coefficient of up to $r = 0.77$ in the rice data.

Better Alignments Give Higher Estimates for the Fraction of Acquired Genes

The correlation between sequence conservation and the reliability of the alignment (fig. 3) and its implications for the reliability of the inferred phylogentic tree (table 3), suggest that the inferred frequency of EGT depends primarily upon the alignment reliability and that the tree results are merely a secondary symptom thereof.

If we estimate the frequency of EGT-positive trees from the most conserved sequences only, we get higher estimates than if we also consider the poorly conserved sequences. The latter are, however, producing many false negatives because we observe that EGT-positive trees are not more prevalent among slowly evolving genes than EGT negatives by the measure of pairwise identity of the archaeplastidan homologues (fig. 2*e*). Because alignments determine phylogenetic results via the generation of homology patterns, the most accurate estimate should come from the best alignments.

In figure 4, the fraction of EGT-positive trees is plotted across the distributions for bin intervals of heads-versus-tails alignment quality as estimated by the criteria CS,

Table 3
Correlation coefficient of the PPS and 2 alignment reliability measures: the CS and the SPS. Both *CS* and *SPS* correlate significantly ($P < 0.05$ using the Pearson's correlation) with the proportion of identical PPS as calculated for both NJ and ML trees in all tested genomes

	Arabidopsis thaliana		*Oryza sativa*		*Chlamydomonas reinhardtii*		*Cyanidioschyzon merolae*	
	PPS (ML)	PPS (NJ)	PPS (ML)	PPS (NJ)	PPS (ML)	PPS (NJ)	PPS (ML)	PPS (NJ)
CS	0.6	0.63	0.61	0.66	0.61	0.65	0.42	0.45
SPS	0.71	0.75	0.77	0.79	0.7	0.75	0.48	0.53

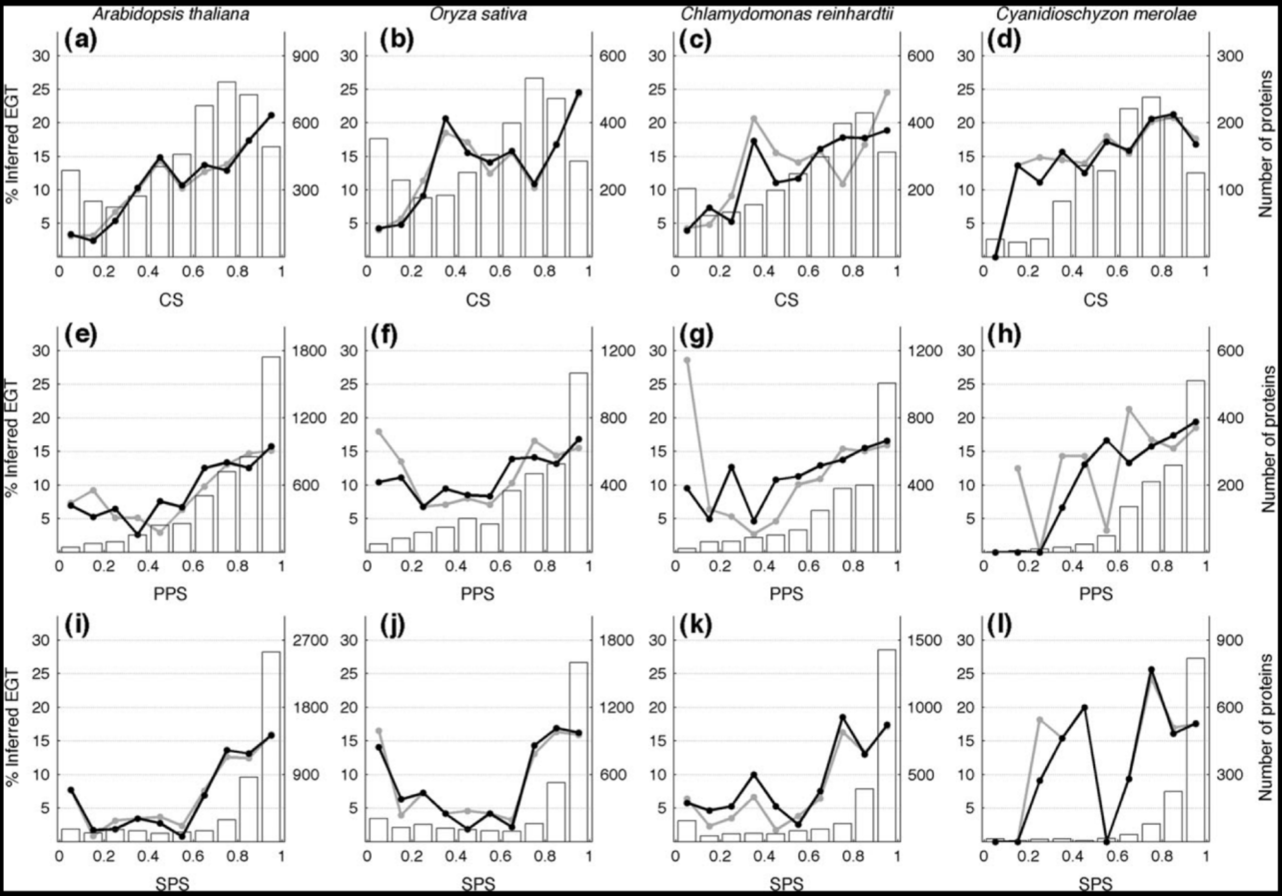


FIG. 4.—The frequency of EGT as inferred from MSAs of varying reliability degrees. The distribution of MSA reliability as estimated by the 3 measures is presented in open bars. The frequency of genes inferred as EGT is plotted above for ML trees (black) and NJ trees (gray).

PPS, and SPS for each archeaplastidan genome data set. For each criterion, the most reliable (reproducible) alignments give the highest estimate of EGT-positive trees, on the order of 17–25% of all trees examined, with a decreasing trend toward poorer alignments, notwithstanding variation among intervals containing small sample sizes. For alignments with CS ≥0.9 (fig. 4*a–d*), the highest estimates are obtained, but between 87% (*Chlamydomonas*) and 91% (*Oryza*) of the gene trees are excluded from consideration by this stringent criterion. For alignments with PPS ≥0.9 (fig. 4*e–h*), over 50% of all trees are excluded, but the estimates for the proportion of EGT-positive trees are very similar to the estimates for SPS ≥0.9 (fig. 4*i–l*), where between 50% (*Oryza*) and 67% (*Cyanidioschyzon*) of the gene trees are included in the estimate. Clearly, excluding alignments in which >20% of the site pairs differ in the toss-up comparison represents a very conservative threshold for excluding phylogenomic data; such data should not be subjected to phylogenetic inference because the phylogeny inference program will be optimizing parameters for site patterns of extremely uncertain homology. If we use the conservative value of SPS ≥0.9 as a cutoff for excluding data that is identified by the HoT method as irreproducibly alignable, hence unreliable, then 16%, 16%, 17%, and 18% of the alignments (gene trees) examined indicate a cyanobacterial origin of the plant nuclear gene for *Arabidopsis*, *Oryza*, *Chlamydomonas*, and *Cyanidioschyzon*, respectively.

How Do Genes Get to the Nucleus and Why Are Any Retained in the Plastid?

The quantitative contribution of EGT inferred here for plant genomes is substantial and raises the question of what gene transfer mechanism is involved. Various lines of evidence favor the view that the mechanism of EGT involves bulk recombination of organelle DNA that is released into the cytoplasm, perhaps by organelle lysis, with DNA of nuclear chromosomes (Timmis et al. 2004). One such line of evidence stems from fragments of organelle DNA that have been relocated to the nucleus and integrated into nuclear chromosomes (Bensasson et al. 2001; Richly and Leister 2004; Behura 2007; Hazkani-Covo and Graur 2007). Nuclear copies of organelle DNA do not preferentially comprise coding regions or particular segments of organelle DNA (Hazkani-Covo and Graur 2007). In *Arabidopsis*, a complete and almost intact 131-kb copy of the chloroplast genome is found near the centromere of nuclear chromosome 2, whereas in rice, a complete and nearly intact copy of the 367-kb mitochondrial genome is found near the centromere of chromosome 10. Both copies share >99%

sequence identity with their homologues in organelles (Huang et al. 2005), indicating that they were transferred intact and recently during evolution. In addition, laboratory studies employing transgenic mitochondria (Thorsness and Fox 1990) or transgenic chloroplasts (Huang et al. 2003) with suitable marker genes have directly demonstrated novel organelle-to-nucleus DNA transfer events involving bulk DNA recombination and at rates that compare to the rate of point mutation (Stegemann et al. 2003). Sequenced eukaryotic genomes are, in general, replete with copies and fragments of organelle DNA that have been relocated to the nucleus and integrated into nuclear chromosomes (Leister 2005). This, together with studies involving transgenic organelles, indicate that gene transfer from organelles to the nucleus via organelle lysis and bulk recombination of organelle chromosomes into nuclear chromosomes is the mechanism of organelle-to-nucleus gene transfer, with recombination, expression, mutation, and selection governing the fate of the transferred genes.

As in a previous study (Martin et al. 2002), the list of genes transferred to plant nuclei from the ancestor of plastids contains sequences encoding virtually all functional categories but mainly biosynthetic and metabolic functions (supplementary table S2, Supplementary Material online). The mechanism of transfer suggests that any gene can be transferred and fixed in the nucleus, so why should any genes be retained in the organelle? A number of suggestions have been offered in this regard (Allen 2003), but the one that most fully accounts for the observations is that individual plastids need to be able to regulate the expression of genes encoding proteins involved in maintaining redox balance in the photosynthetic membrane.

Seeking the Closest Relative of Plastids in Nuclear Genes and Allowing for LGT

In analyses of ribosomal RNA sequences, plastids tend to branch deeply within cyanobacterial phylogeny but to show no specific affinity to any particular cyanobacterial lineage (Turner et al. 1999; Marin et al. 2005). In analyses of alignments of concatenated plastid-encoded protein data, Rodriguez-Ezpeleta et al. (2005) obtained the same result, whereas Sato (2006) found weak phylogenetic evidence to suggest that plastids might stem from within the *Anabaena–Synechocystis* lineage. As outlined in the introduction, the concatenation approach usually assumes that the concatenated sequences in question all share the same history, and this assumption is generally problematic where prokaryotic genes are involved (Bapteste et al. 2007). Initial studies with concatenated sequences of plastid-encoded genes investigated the congruence of signals for individually analyzed plastid proteins (Goremykin et al. 1997; Martin et al. 1998; Vogl et al. 2003), but this step is usually omitted in newer analyses (Rodriguez-Ezpeleta et al. 2005), which would appear problematic, especially given the evidence that cyanobacteria exchange genes (Raymond et al. 2002; Zhaxybayeva et al. 2004, 2006) and that signal loss in plastid-encoded proteins is an issue even during algal phylogeny (White et al. 2007). Taken together, the available observations suggest that there is not enough information present in the ~45 proteins common to the genomes of photosynthetic plastids to confidently ascertain the closest relative of plastids among modern cyanobacteria. Furthermore, the observations suggest that LGT among cyanobacteria requires that the question be approached in such a manner as to avoid the concept of "sister group lineages" among prokaryotes at the whole genome level (Doolittle and Bapteste 2007), whereas for individual genes, the concept of the sister group seems unproblematic.

Given the foregoing, we asked: Which of the 9 cyanobacterial genomes sampled here contains the highest frequency of genes that are scored as being of cyanobacterial origin in plant nuclear genomes? For each of the 4 plant genomes and for each of the 9 cyanobacterial genomes sampled, we tabulated the pairwise amino acid sequence identity for each cyanobacterial protein that was present in our alignments and that was scored as being of cyanobacterial origin in the respective plant genome. Color-coded versions of those tables are presented in figure 5*a*, where it is evident that the 9 cyanobacteria sampled differ with respect to their overall similarity to the collection of cyanobacterial genes that is present in plant nuclear genomes. The differences reflect gene presence or absence on the one hand, with the smallest genome (2,265 proteins in *P. marinus* MIT9313) harboring fewer proteins that share high sequence identity to the plant homologues. But the differences also reflect overall sequence similarity, with *G. violaceus* PCC 7421, a midsized genome with 4430 proteins, harboring many homologues of plant proteins of cyanobacterial origin, albeit with visibly lower sequence identity than the other cyanobacteria in the sample (fig. 5*a*). This finding consistent with the circumstance that *Gloeobacter* is generally regarded a primitive or early-branching cyanobacterium because of its lack of thylakoids and its position in some phylogenetic trees (Sato 2006; Tomitani et al. 2006).

The sequence similarity array in figure 5*a* summarizes many millions of individual sequence comparisons, but it does not deliver phylogenetic specifics. Figure 5*b* shows the frequency with which proteins from the given cyanobacterial genome occurred within the sister group to the plant nuclear gene among the 11,569 ML phylogenies inferred in the present study. Those frequencies are shown for all alignments, for the alignments with SPS ≥0.8, and for the alignments with SPS ≥0.9. In all cases, *Anabaena* and *Nostoc* had the highest frequency of harboring a sister of the plant protein in phylogenetic trees. This result is influenced to some extent by genome size because *Anabaena* and *Nostoc* also have the highest frequency of harboring a homologue of those plant proteins, which found hits in cyanobacteria only (fig. 5*c*), and gene presence/absence is directly affected by genome size. But the size of a genome within which a protein-coding gene resides does not affect sequence similarity in individual protein comparisons (fig. 5*a*) nor does it directly bias the phylogentic results, as seen in the comparison of genome size (fig. 5*d*) and sister group frequency in ML trees (fig. 5*b*) for *Gloeobacter*. In other words, among the 9 cyanobacterial genomes sampled, *Nostoc* sp. PCC7120 and *A. variabilis* ATCC29143 harbor collections of genes that are—in terms of presence/absence and sequence similarity—more like those possessed by

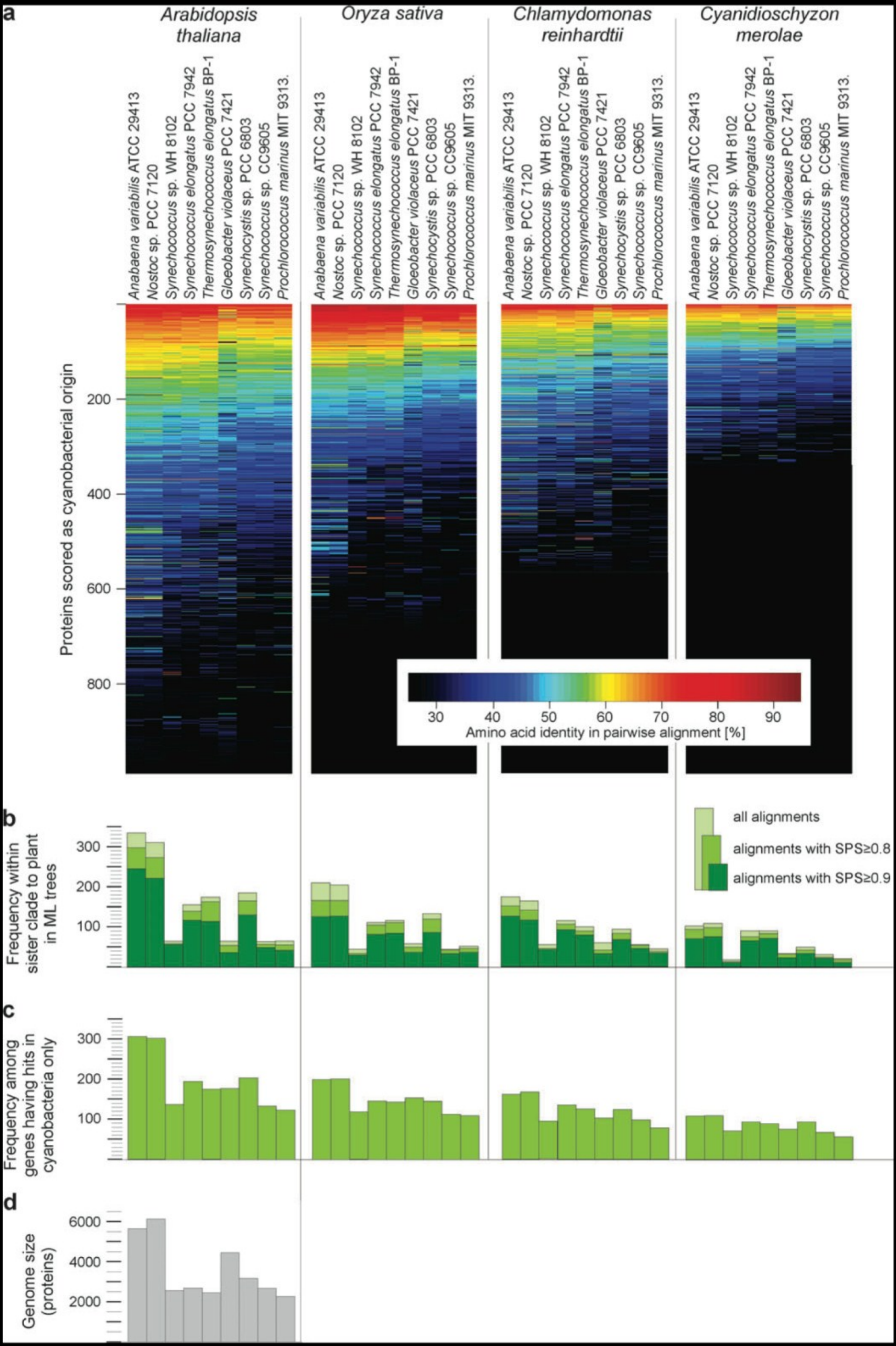


FIG. 5.—Overall similarity of proteins of cyanobacterial origin encoded in plant nuclear genomes to proteins encoded in cyanobacterial genomes. (*a*) For each of the 4 plant genomes, a color-coded table is shown. The rows correspond to the proteins scored as acquisitions from cyanobacteria; this includes cyanobacterial nearest neighbor inference and exclusive Blast hits within the cyanobacterial proteomes. The *Oryza sativa* RefSeq nuclear data includes a total of 51 genes that are annotated as nuclear proteins but probably are NUPTs (see Materials and Methods). Columns correspond to cyanobacterial genomes, the elements of the matrix contain the percent amino acid identity between the plant query, and the best hit in the respective cyanobacterial genome in the pairwise alignment generated by Blast. The scale of amino acid identity is given in the inset at lower right. (*b*) The frequency with which the homologue from the cyanobacterial genome indicated in (a) appeared in the cyanobacterial sister clade to the respective plant protein in ML trees. (*c*) The frequency with which the cyanobacterial genome indicated in (a) contained a homologue of the plant protein for plant proteins that found homologues in cyanobacteria only at the specified threshold. (*d*) Number of proteins encoded in each cyanobacterial genome indicated in (a).

the plastid ancestor than are those in the other 7 cyanobacterial genomes sampled here. With regard to *Nostoc* sp. PCC7120, our result is consistent with an earlier result based on a smaller cyanobacterial genome sample and involving *Nostoc punctiforme* (Martin et al. 2002). With regard to *Nostoc* and *Anabaena*, our result is also consistent with the conclusions of Sato (2006) as they apply to *Anabaena*.

Nostoc, *Anabaena*, Symbiosis, Plastids, and Nitrogen

With the caveat that the present cyanobacterial sample is quite limited, our finding is that *Nostoc* and *Anabaena* harbor more genes than other cyanobacteria of the type that plants acquired from cyanobacteria. *Nostoc* and *Anabaena* are filamentous cyanobacteria that produce heterocysts, differentiated and specialized cells that perform nitrogen fixation without producing oxygen (Rippka et al. 1979; Rajaniemi et al. 2005). Filamentous cyanobacteria that produce heterocysts are grouped in the section IV (without true branching) and section V (with true branching) as recognized in Roger Stanier's cyanobacterial classification (Rippka et al. 1979), which followed the same morphological principles as Geitler's (1932) traditional system. Tomitani et al. (2006) presented evidence to indicate that sections IV and V are the most derived among cyanobacteria, consistent with traditional views (Geitler 1932; Rippka et al. 1979) and furthermore argued that the origin of heterocysts antedates the rise of atmospheric oxygen about 2.3 billion years ago. Were the ancestor of plastids a member of section IV (a heterocyst-forming cyanobacterium), as the present data tend to suggest, that would put a weak bound on the maximum age of plastids at 2.3 Ga, with the minimum age well constrained by the fossil red alga *Bangiomorpha* at 1.2 Ga (Butterfield 2000; Yoon et al. 2004). This conclusion is consistent with the observation that akinetes, large spore-like reproductive cells produced only by some members of sections IV and V (Rippka et al. 1979; Rajaniemi et al. 2005), appear in the palaeontological record >0.4 Ga earlier than *Bangiomorpha* (Tomitani et al. 2006).

If the cyanobacterial ancestor of plastids was a member of section IV, it was able to fix nitrogen, raising the question of what, exactly, the cyanobacterial symbiont was doing for its host during the early stages of symbiosis. It was suggested that the retargeting of a host-derived carbon transporter was the initial step that allowed the cyanobacterium to export reduced carbon for its host (Weber et al. 2006), but normal cyanobacterial cell wall polysaccharides would also perform that function, and how such a transporter would be targeted across the cyanobacterial cell wall into the plasma membrane was not discussed. Other suggestions have also been put forth, namely local oxygen production, either before (Stanier 1970) or after (Martin and Müller 1998) the origin of mitochondria, and nitrogen has recently been considered (Kneip et al. 2007). Indeed, if we look around at the chemical function of cyanobacteria in modern symbiotic endosymbiotic associations where the physiology has been studied, nitrogen supply, in some cases without carbon export, stands in the foreground (Rai et al. 2000; Raven 2002). Well-studied examples of modern cyanobacterial endosymbioses include *Geosiphon pyriforme* (a fungus; Mollenhauer et al. 1996), *Rhopalodia gibba* (a diatom; Prechtl et al. 2004), *Azolla* species (an acquatic fern; Prasanna et al. 2006), coralloid roots in cycads (Costa et al. 2004), and *Gunnera* (a flowering plant; Chiu et al. 2005).

In those examples, the cyanobacterial partner is a member of section IV from the genus *Nostoc* or *Anabanea*, the exception is *Rhopalodia*, where the symbiont is related to *Cyanothece* (Prechtl et al. 2004), a coccoid member of section I. In the majority of modern cyanobacterial endosymbioses, the endosymbionts are diazotrophic (nitrogen fixing), in many cases they provide reduced nitrogen to their host (Rai et al. 2000; Raven 2002; Prechtl et al. 2004), and *Nostoc* is a very common partner. *Richellia* (Nostocales) endosymbionts living within *Hemiaulus* and *Rhizosolenia* provide nitrogen to their diatom hosts (Raven 2002) and are important contributors to marine N availability (Montoya et al. 2004). Also in many ectosymbiotic associations, for example lichens, where *Nostoc* species are typically the cyanobacterial partner (Rikkinen et al. 2002) or in open ocean epiphyite associations of *Dichothrix* (Nostocales) with the brown alga *Sargassum* (Carpenter 1972), nitrogen plays a decisive role. The role of nitrogen in symbiosis has recently been reviewed by Kneip et al. (2007).

In the larger geological context, it was proposed that the current model of anoxic and sulfidic oceans during the Proterozoic would have limited nitrogen availability globally (Anbar and Knoll 2002) starting from about 2.3 Ga up until ~0.58 Ga as newer findings indicate (Fike et al. 2006; Canfield et al. 2007). That is the time during which plastids arose (Butterfield 2000; Yoon et al. 2004) and it corresponds to "*. . . a period of exceptional N stress for the biosphere*" (Anbar and Knoll 2002, p. 1140), owing to the limited marine availability of trace elements—the transition metals Mo, Fe, and V—the limitation arising from the insolubility of the corresponding transition metal sulfides in oceans sulfidic due to biological sulfate reduction (Anbar and Knoll 2002). Taken together, that environmental limitation during the time of plastid origin, the role of nitrogen in modern cyanobacterial endosymbioses (Rai et al. 2000), and our present data linking plastids to heterocyst-forming cyanobacteria among a yet limited sample would be compatible with the view that nitrogen could have played a role in the establishment of the symbiosis that led to plastids (Raven 2002; Kneip et al. 2007).

Conclusion

On average, 14% of the nuclear-encoded proteins in the genomes of 4 photosynthetic eukaryotes having homologues in cyanobacteria and at least 2 other phyla, regardless of their alignment quality, were inferred as gene acquisitions from cyanobacteria. Alignments with better sequence conservation—that is, alignments whose site patterns were independent of the order in which amino acids were aligned—consistently recover a higher proportion of inferred cyanobacterial origin for plant nuclear genes (16–18%; >20% if only very highly conserved sequences are considered) than the alignments of more poorly conserved sequences. That suggests that the latter contain many false negatives and that the conservative average value of 14% is an underestimate, underscoring a quantitatively large contribution of cyanobacteria to the makeup of plant genomes.

It has long been known from simulation studies that sequences sharing <50% amino acid identity perform poorly in phylogenetic reconstruction (Nei 1996). But in the deeper regions of genome phylogenetics, such as plastid origins, sequence pairs sharing ≥50% amino acid identity are comparatively rare. The heads-or-tails approach does not directly identify a demarcation line that may not be crossed in sequence analysis, but it does reveal when phylogenetic results are solely determined by a process no less arbitrary than a coin toss and hence provides a reality check of the data quality underlying phylogenomic analyses.

The present phylogenomic data point to filamentous, heterocyst-forming (nitrogen-fixing) cyanobacteria as the plastid ancestor, an inference that is compatible with the observation that the majority of contemporary, physiologically characterized, cyanobacterial endosymbioses entail nitrogen supply as a benefit from endosymbiont to host. The current model of Proterozoic ocean chemistry (Anbar and Knoll 2002; Canfield et al. 2007) would not only be compatible with the widespread occurrence of anaerobic forms of mitochondria among the major eukaryotic lineages (Theissen et al. 2003; Dietrich et al. 2006; Embley and Martin 2006), but also be highly compatible with the concept of a plastid that could fix and supply nitrogen, owing to the distinct possibility of nitrogen limitation in the Proterozoic ocean.

Supplementary Material

Supplementary figure S1 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Paul Falkowski, Peter Lockhart, and Uwe Maier for stimulating discussions. This work was funded by the Deutsche Forschungsgemeinschaft (SFB-Tr1; W.M.), the German-Israeli Foundation (T.D.) and the National Scientific Foundation (DBI-0543342; G.L.).

Literature Cited

Adl SM, Simpson AGB, Farmer MA, et al. (28 co-authors). 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Euk Microbiol. 52:399–451.

Allen JF. 2003. The function of genomes in bioenergetic organelles. Philos Trans R Soc Lond B. 358:1i9–37.

Allen JF, Raven JA. 1996. Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles. J Mol Evol. 42:482–492.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. 25:338i9–3402.

Anbar AD, Knoll AH. 2002. Proterozoic ocean chemistry and evolution: a bioinorganic bridge. Science. 297:1137–1142.

Archibald JM. 2006. Algal genomes: exploring the imprint of endosymbiosis. Curr Biol. 16:R1033–R1035.

Bapteste E, Susko E, Leigh J, Ruiz-Trillo I, Bucknam L, Doolittle WF. 2007. 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.

Behura SK. 2007. Analysis of nuclear copies of mitochondrial sequences in honey bee (Apis mellifera) genome. Mol Biol Evol. 24:1492–1505.

Bensasson D, Zhang D, Hartl DL, Hewitt GM. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol. 16:314–321.

Bodyl A, Mackiewicz P, Stiller JW. 2007. The intracellular cyanobacteria of *Paulinella chromatophora*: endosymbionts or organelles? Trends Microbiol. 15:295–296.

Boeckmann B, Bairoch A, Apweiler R, et al. (12 co-authors). 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31:365–370.

Bogorad L. 2008. Evolution of early eukaryotic cells: genomes, proteomes, and compartments. Photosynth Res. 95:11–21.

Butterfield NJ. 2000. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiology. 263:386–404.

Canfield DE, Poulton SW, Narbonne GM. 2007. Late-Neoproterozoic deep-ocean oxygenation and the rise of animal life. Science. 315:92–95.

Carpenter EJ. 1972. Nitrogen fixation by a blue-green epiphyte on pelagic *Sargassum*. Science. 178:1207–1209.

Chiu WL, Peters GA, Levieille G, Still PC, Cousins S, Osborne B, Elhai J. 2005. Nitrogen deprivation stimulates symbiotic gland development in *Gunnera manicata*. Plant Physiol. 139:224–230.

Costa JL, Romero EM, Lindblad P. 2004. Sequence based data supports a single *Nostoc* strain in individual coralloid roots of cycads. FEMS Microbiol Ecol. 49:481–487.

Delwiche CW. 1999. Tracing the thread of plastid diversity through the tapestry of life. Am Nat. 154:S164–S177.

Dietrich LEP, Tice MM, Newmann DK. 2006. The co-evolution of life and earth. Curr Biol. 16:R395–R400.

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci USA. 104:2043–2049.

Douglas SE. 1998. Plastid evolution: origins, diversity, trends. Curr Opin Genet Dev. 8:655–661.

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

Embley TM, Martin W. 2006. Eukaryotic evolution: changes and challenges. Nature. 440:623–630.

Esser C, Martin W, Dagan T. 2007. The origin of mitochondria in light of a fluid prokaryotic chromosome model. Biol Lett. 3:180–184.

Felsenstein J. 2005. PHYLIP (phylogeny inference package). Version 3.6. Seattle (WA): Department of Genome Sciences, University of Washington.

Fike DA, Grotzinger JP, Pratt LM, Summons RE. 2006. Oxidation of the Ediacaran ocean. Nature. 444:744–747.

Geitler L. 1932. Rabenhorst's Kryptogamenflora von Deutschland, Österreich und der Schweiz. Vierzehnter Band: Cyanophyceae. Leipzig (Germany): Akademische Verlagsgesellschaft M.B.H., p. 1196.

Goremykin VV, Hansmann S, Martin WF. 1997. Evolutionary analysis of 58 proteins encoded in six completely sequenced chloroplast genomes: revised molecular estimates of two seed plant divergence times. Plant Syst Evol. 206:337–351.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation rate matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Hazkani-Covo E, Graur D. 2007. A comparative analysis of numt evolution in human and chimpanzee. Mol Biol Evol. 24: 13–18.

Huang CY, Ayliffe MA, Timmis JN. 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature. 422:72–76.

Huang CY, Grunheit N, Ahmadinejad N, Timmis JN, Martin W. 2005. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol. 138:1723–1733.

Kneip C, Lockhart P, Voss C, Maier UG. 2007. Nitrogen fixation in eukaryotes—new models for symbiosis. BMC Evol Biol. 7:55.

Kumar S, Filipski A. 2007. Multiple sequence alignment: in pursuit of homologous DNA positions. Genome Res. 17: 127–135.

Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.

Larkum AW, Lockhart PJ, Howe CJ. 2007. Shopping for plastids. Trends Plant Sci. 12:189–195.

Leister D. 2005. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet. 21:655–663.

Lockhart PJ, Howe CJ, Barbrook AC, Larkum AWD, Penny D. 1999. Spectral analysis, systematic bias, and the evolution of chloroplasts. Mol Biol Evol. 16:573–576.

Marin B, Nowack ECM, Melkonian M. 2005. A plastid in the making: evidence for a second primary endosymbiosis. Protist. 156:425–432.

Martin W. 1999. Mosaic bacterial chromosomes—a challenge en route to a tree of genomes. Bioessays. 21:99–104.

Martin W, Brinkmann H, Savona C, Cerff R. 1993. Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. Proc Natl Acad Sci USA. 90:8692–8696.

Martin W, Herrmann RG. 1998. Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol. 118:9–17.

Martin W, Müller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature. 392:37–41.

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA. 99:12246–12251.

Martin W, Schnarrenberger C. 1997. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. Curr Genet. 32:1–18.

Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature. 393:162–165.

Matsuzaki M, Misumi O, Shin-IT, et al. (42 co-authors). 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature. 428:653–657.

McFadden GI, van Dooren GG. 2004. Evolution: red algal genome affirms a common origin of all plastids. Curr Biol. 14:R514–R516.

Merchant SS, Prochnik SE, Vallon O, et al. (117 co-authors). 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science. 318:245–250.

Mereschkowsky C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biol Centralbl. 25:593–604. [English translation in Eur J Phycol. 34:287–295 (1999)].

Mollenhauer D, Mollenhauer R, Kluge M. 1996. Studies on initiation and development of the partner association in Geosiphon pyriforme (Kutz) v Wettstein, a unique endocytobiotic system of a fungus (Glomales) and the cyanobacterium *Nostoc punctiforme* (Kutz) Hariot. Protoplasma. 193:3–9.

Montoya JP, Holl CM, Zehr JP, Hansen A, Villareal TA, Capone DG. 2004. High rates of N_2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. Nature. 430:1027–1031.

Morden CW, Golden SS. 1989. psbA genes indicate common ancestry of prochlorophytes and chloroplasts. Nature. 337:382–385.

Nei M. 1996. Phylogenetic analysis in molecular evolutionary genetics. Annu Rev Genet. 30:371–403.

Nei M, Takezaki N, Sitnikova T. 1995. Assessing molecular phylogenies. Science. 267:253–254.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol. 21:1455–1458.

Prasanna R, Kumar R, Sood A, Prasanna BM, Singh PK. 2006. Morphological, physiochemical and molecular characterization of *Anabaena* strains. Microbiol Res. 161:187–202.

Prechtl J, Kneip C, Lockhart P, Wenderoth K, Maier UG. 2004. Intracellular spheroid bodies of *Rhopalodia gibba* have nitrogen-fixing apparatus of cyanobacterial origin. Mol Biol Evol. 21:1477–1481.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 33:D501–D504.

Rai AN, Söderbäck E, Bergman B. 2000. Cyanobacterium-plant symbioses. New Phytol. 147:449–481.

Rajaniemi P, Hrouzek P, Kastovská K, Willame R, Rantala A, Hoffmann L, Komárek J, Sivonen K. 2005. Phylogenetic and morphological evaluation of the genera *Anabaena*, *Aphanizomenon*, *Trichormus* and *Nostoc* (Nostocales, Cyanobacteria). Int J Syst Evol Microbiol. 55:11–26.

Raven JA. 2002. Evolution of cyanobacterial symbioses. In: Rai AN, Bergman B, Rasmussen U, editors. Cyanobacteria in symbiosis. Dordrecht (The Netherlands): Kluwer Academic Publishers. p. 326–1246.

Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE. 2002. Whole-genome analysis of photosynthetic prokaryotes. Science. 298:1616–1620.

Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D. 2006. Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. Curr Biol. 16:2320–2325.

Reyes-Prieto A, Weber APM, Bhattacharya D. 2007. The origin and establishment of the plastid in algae and plants. Annu Rev Genet. 41:147–680.

Richly E, Leister D. 2004. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice. Gene. 329:11–16.

Rikkinen J, Oksanen I, Lohtander K. 2002. Lichen guilds share related cyanobacterial endosymbionts. Science. 297:357.

Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY. 1979. Generic assignments, strain histories and properties of pure cultures of cyanobacteria. J Gen Microbiol. 111:1–6.

Rodriguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Loffelhardt W, Bohnert HJ, Philippe H, Lang BF. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol. 15:1325–1330.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.

Sato N. 2001. Was the evolution of plastid genetic machinery discontinuous? Trends Plant Sci. 6:151–155.

Sato N. 2006. Origin and evolution of plastids: genomic view on the unification and diversity of plastids. In: Wise RR, Hoober JK, editors. The structure and function of plastids. Dordrecht (The Netherlands): Springer. p. 75–102.

Soll J, Schleiff E. 2004. Protein import into chloroplasts. Nat Rev Mol Cell Biol. 5:198–208.

Stanier RY. 1970. Some aspects of the biology of cells and their possible evolutionary significance. Symp Soc Gen Microbiol. 20:1–38.

Stegemann S, Hartmann S, Ruf S, Bock R. 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc Natl Acad Sci USA. 100:8828–8833.

Stoebe B, Hansmann S, Goremykin V, Kowallik KV, Martin W. 1999. Proteins encoded in sequenced chloroplast genomes: an overview of gene content, phylogenetic information, and endosymbiotic gene transfer to the nucleus. In: Hollingsworth C, Bateman R, Gornall M, editors. Advances in plant molecular systematics. Andover (UK): Francis and Taylor. p. 327–352.

Theissen U, Hoffmeister M, Grieshaber M, Martin W. 2003. Single eubacterial origin of eukaryotic sulfide: quinone oxidoreductase, a mitochondrial enzyme conserved from the early evolution of eukaryotes during anoxic and sulfidic times. Mol Biol Evol. 20:1564–1574.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: 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.

Thompson JD, Plewniak F, Poch O. 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 27:2682–2690.

Thorsness PE, Fox TD. 1990. Escape of DNA from mitochondria to the nucleus in *Saccharomyces cerevisiae*. Nature. 346:376–379.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 5:123–135.

Tomitani A, Knoll AH, Cavanaugh CM, Ohno T. 2006. The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. Proc Natl Acad Sci USA. 103:5442–5447.

Turner S, Pryer KM, Miao VPW, Palmer JD. 1999. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small submit rRNA sequence analysis. J Euk Microbiol. 46:327–338.

Tyra HM, Linka M, Weber AP, Bhattacharya D. 2007. Host origin of plastid solute transporters in the first photosynthetic eukaryotes. Genome Biol. 8:R212.

Vogl C, Badger J, Kearney P, Li M, Clegg M, Jiang T. 2003. Probabilistic analysis indicates discordant gene trees in chloroplast evolution. J Mol Evol. 56:330–40.

Weber AP, Linka M, Bhattacharya D. 2006. Single, ancient origin of a plastid metabolite translocator family in Plantae from an endomembrane-derived ancestor. Eukaryot Cell. 5:609–612.

White WT, Hills SF, Gaddam R, Holland BR, Penny D. 2007. Treeness triangles: visualizing the loss of phylogenetic signal. Mol Biol Evol. 24:2029–2039.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol. 21:809–818.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. 16:1099–1108.

Zhaxybayeva O, Hamel L, Raymond J, Gogarten JP. 2004. Visualization of the phylogenetic content of five genomes using dekapentagonal maps. Genome Biol. 5:R20.

Takashi Gojobori, Associate Editor

Accepted January 20, 2008

7 Lateral gene transfer (LGT) during prokaryote evolution.

7.1 Biases in LGT inference

Roettger M, Martin W, Dagan T: **A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies**. *Mol Biol Evol* 2009, **26**:1931-1939.

(Own contribution: conceived and designed the experiment, analyzed the data, and wrote the paper).

A Machine-Learning Approach Reveals That Alignment Properties Alone Can Accurately Predict Inference of Lateral Gene Transfer from Discordant Phylogenies

Mayo Roettger, William Martin, and Tal Dagan

Institut für Botanik III, Heinrich-Heine Universität Düsseldorf, Germany

Among the methods currently used in phylogenomic practice to detect the presence of lateral gene transfer (LGT), one of the most frequently employed is the comparison of gene tree topologies for different genes. In cases where the phylogenies for different genes are incompatible, or discordant, for well-supported branches there are three simple interpretations for the result: 1) gene duplications (paralogy) followed by many independent gene losses have occurred, 2) LGT has occurred, or 3) the phylogeny is well supported but for reasons unknown is nonetheless incorrect. Here, we focus on the third possibility by examining the properties of 22,437 published multiple sequence alignments, the Bayesian maximum likelihood trees for which either do or do not suggest the occurrence of LGT by the criterion of discordant branches. The alignments that produce discordant phylogenies differ significantly in several salient alignment properties from those that do not. Using a support vector machine, we were able to predict the inference of discordant tree topologies with up to 80% accuracy from alignment properties alone.

Introduction

The phylogenetic approach for lateral gene transfer (LGT) inference from the frequency of incongruent branching patterns in gene trees has so far delivered widely conflicting results, ranging from estimates that as few as 2% (Ge et al. 2005) to possibly 14% of all genes in prokaryote genomes are affected by LGT (Beiko, Harlow, and Ragan 2005). Such divergent estimates using phylogenetic tree comparisons can, in principle, be attributed to many factors including the obvious, such as lineage sampling, the inherent uncertainties of various approaches to phylogenetic reconstruction (Penny et al. 1992; Hillis 1995; Lopez et al. 2002) and the threshold levels of support set to score the presence of genuinely conflicting topologies. But phylogenetic trees of molecular sequences are always inferred from multiple sequence alignments. Nei et al. (1995) and Nei (1996) pointed out early on that alignment of highly diverged sequences may result in erroneous phylogenetic reconstruction. Interest in this aspect of phylogeny has renewed with several reports investigating the alignment step itself as it specifically relates to phylogenetic inference (Landan and Graur 2007; Deusch et al. 2008; Löytynoja and Goldman 2008; Wong et al. 2008)

Here, we wished to examine the extent to which LGT inference by the phylogenetic method might be sensitive to the properties of alignments themselves. For this purpose, we investigated the comprehensive data set compiled and carefully analyzed by Beiko, Harlow, and Ragan (2005), who kindly made their data available. Their data set is highly suitable for the present study 1) because it consists of 22,437 carefully assembled gene families of prokaryotic orthologs, in which paralogs have been sorted out by using a conservative similarity cutoff (Beiko, Harlow, and Ragan 2005, Supplementary Material online), 2) because they used a widely employed filter, Gblocks (Castresana 2000), to exclude poorly aligned regions from their analysis prior to phylogenetic reconstruction, and 3) because they used a very stringent (conservative) threshold for the scoring of discordant phylogenies. In brief, Beiko, Harlow, and Ragan (2005) constructed a consensus supertree for the proteins encoded in 144 prokaryotic genomes and constructed from the same data 22,437 individual phylogenetic trees containing from 4 to 144 sequences each using a Bayesian approach. They inferred LGT only from highly significant (posterior probability ≥ 0.95) discordant tree topologies in comparison to the consensus supertree topology (Beiko, Harlow, and Ragan 2005). For 5,822 of those trees, one or more LGT was inferred on the basis of discordance to the consensus topology, we designate those amino acid sequence alignments as “LGT positive” or LGT for short. The remaining 16,615 of the alignments investigated by Beiko, Harlow, and Ragan (2005) did not produce branches (bipartitions) that were discordant (conflicting) with the consensus supertree topology and are considered here as “vertical gene inheritance” or VGI alignments. We examined the properties of the LGT alignments in comparison to the properties of the VGI alignments.

Methods

For the analysis, we used a data set of 22,437 protein families from 144 prokaryotes for which LGT) was inferred using the phylogenetic method (Beiko, Harlow, and Ragan 2005). The data for each protein family include a multiple sequence alignment yielding the highest score according to the word-oriented objective function (Beiko, Chan, and Ragan 2005) from a set of alignments reconstructed by several different algorithms: ClustalW (Thompson et al. 1994), T-coffee (Notredame et al. 2000), MAFFT (Katoh et al. 2002), POA (Grasso and Lee 2004), and PRRP (Gotoh 1996), a partial alignment of relatively conserved regions constructed with Gblocks (Castresana 2000), and a phylogenetic tree inferred with MrBayes (Huelsenbeck and Ronquist 2001). Bipartitions

Key words: lateral gene transfer, molecular phylogeny, discordant tree topologies, support vector machine, principal component analysis.

E-mail: mayo.roettger@uni-duesseldorf.de.

Mol. Biol. Evol. 26(9):1931–1939. 2009
doi:10.1093/molbev/msp105
Advance Access publication May 14, 2009

in the phylogenetic tree were considered as concordant if they overlap with the reference supertree, or discordant otherwise, which were interpreted as LGT events (Beiko, Harlow, and Ragan 2005).

Multiple Alignment Properties

For each protein family (alignment), we calculated alignment properties as follows: Number of operational taxonomic units (OTUs) is the number of orthologs in the family. Proportion of gaps is the proportion of gap characters in the Gblocks output alignment. Entropy was calculated for each Gblocks output alignment as the average entropy of its sites. For the calculation, we used the Shannon information content normalized by the number of OTUs in the alignment (Valdar 2002):

$$\text{Entropy} = \frac{\sum_{\text{col}=1}^{N_{\text{sites}}} \left(-\lambda_t \sum_{i=1}^{K} p_{i,\text{col}} \log_2 p_{i,\text{col}}\right)}{N_{\text{sites}}},$$

where N_{sites} is the number of alignment sites, K is the alphabet size, which is 21 in this case (20 amino acids plus one gap symbol), and $p_{i,\text{col}}$ the probability of observing the ith character in alignment column col. The λ_t factor is used to scale the entropy into a [0,1] range by the number of OTUs (N_{seq}) and the alphabet size K.

$$\lambda_t = [\log_2(\min(N_{\text{seq}}, K))]^{-1}.$$

Invariant sites are positions in the Gblocks alignment where all sequences contain the same amino acid, and informative sites are defined as alignment columns containing at least two different amino acids, each one observed in at least two sequences at the position. Average pairwise identity is calculated as the proportion of amino acid identities between all sequence pairs and averaged for the protein family as follows:

$$\text{API} = \frac{2}{N_{\text{seq}}(N_{\text{seq}} - 1)N_{\text{sites}}} \sum_{\text{col}=1}^{N_{\text{sites}}} \sum_{i=1}^{N_{\text{seq}}-1} \sum_{j=i+1}^{N_{\text{seq}}} \left\{ \begin{matrix} 1 : a_{i,\text{col}} & = & a_{j,\text{col}} \\ 0 : a_{i,\text{col}} & \neq & a_{j,\text{col}} \end{matrix} \right\},$$

where $a_{i,\text{col}}$ is the observed amino acid in Gblocks-alignment sequence i at position col.

In addition, we tested for alignment reliability using the Heads-or-Tails (HoT) method (Landan and Graur 2007). Tails alignments were obtained by aligning the reversed sequences of each protein family using exactly the same alignment procedure that was used for the original (Heads) alignments. Column score (CS) is calculated as the proportion of identical columns between Heads and Tails alignments, and sum of pairs score (SPS) was calculated as the proportion of identical residue position pairs between Heads and Tails alignments.

Each alignment comprises protein sequences from different species, each sequence named by Beiko et al. with a unique pipeline id. Additional files contained information about the original gi number in the RefSeq database (Pruitt et al. 2005) together with the current gi number in the database and the genome id of the sequence used by the National Center for Biotechnology Information (NCBI). Two hundred and twenty-three proteins in 183 alignments had no information about the sequence except the original gi number in the database. These database entries were replaced or removed from the database. In the calculation of the number of different phyla and the classification of the sequence into the kingdom groups, we discarded these sequences from the alignment. We obtained taxonomical classification information for each sequence from NCBI and counted the number of sequences being classified as different phyla for each cluster. We used the term archaea for clusters that contain only sequences classified as of archaebacterial origin, the term eubacteria for clusters that contain only sequences classified as of bacterial origin, and the term universal for clusters that contain sequences of both kingdoms.

For the comparison of the property distributions between LGT and VGI alignments, the Wilcoxon nonparametric test was used.

Orthologs Pairwise Distances

Protein pairwise distances between orthologs from LGT and VGI families were calculated for several genome pairs that were selected for their high frequency in the data: 1) *Vibrio vulnificus* versus *Yersinia pestis* (1,406 protein pairs where both species were present in the respective protein families), 2) *Brucella suis* versus *Mesorhizobium loti* (1,697 protein pairs), 3) *Agrobacterium tumefaciens* versus *M. loti* (2,205 protein pairs), and 4) *Bradyrhizobium japonicum* versus *M. loti* (1,794 protein pairs), 5) *Staphylococcus aureus* versus *Bacillus cereus* (924 protein pairs), 6) *Nostoc* sp. versus *Pyrococcus furiosus* (114 protein pairs), and 7) *Bacteroides thetaiotaomicron* versus *Sulfolobus solfataricus* (62 protein pairs). Pairwise protein distances were extracted from the distance matrix calculated from the multiple sequence alignments with PROTDIST (Felsenstein 1996) using Jonen-Taylor-Thornton substitution matrix (Jones et al. 1992). In addition, we calculated pairwise distances with the same method after realigning the orthologous sequences using MUSCLE (Edgar 2004).

Classification Procedure

Prediction of LGT (discordant tree bipartition) from alignment properties entailed a support vector machine (SVM) classifier (Christiani and Shawe-Taylor 2000). For the SVM training and classifying procedures, we used the svmtrain and svmclassify functions from the MATLAB 7.6 bioinformatics toolbox with the following parameters: Radial basis function (RBF) kernel, RBFSigmaValue = 1, Mlp_ParamsValue = [1,−1], MethodValue = SMO, BoxConstraintValue = 1, and AutoscaleValue = true. In order to obtain significance levels for the SVM performance, we applied 10-fold crossvalidation in each step using the small 1/10 subset for training and the 9/10 for testing. The LGT/VGI ratio in the training set was adjusted by randomly selecting different numbers of LGT and VGI samples from the preliminary training set to form an equal-sized training set for each validation step.

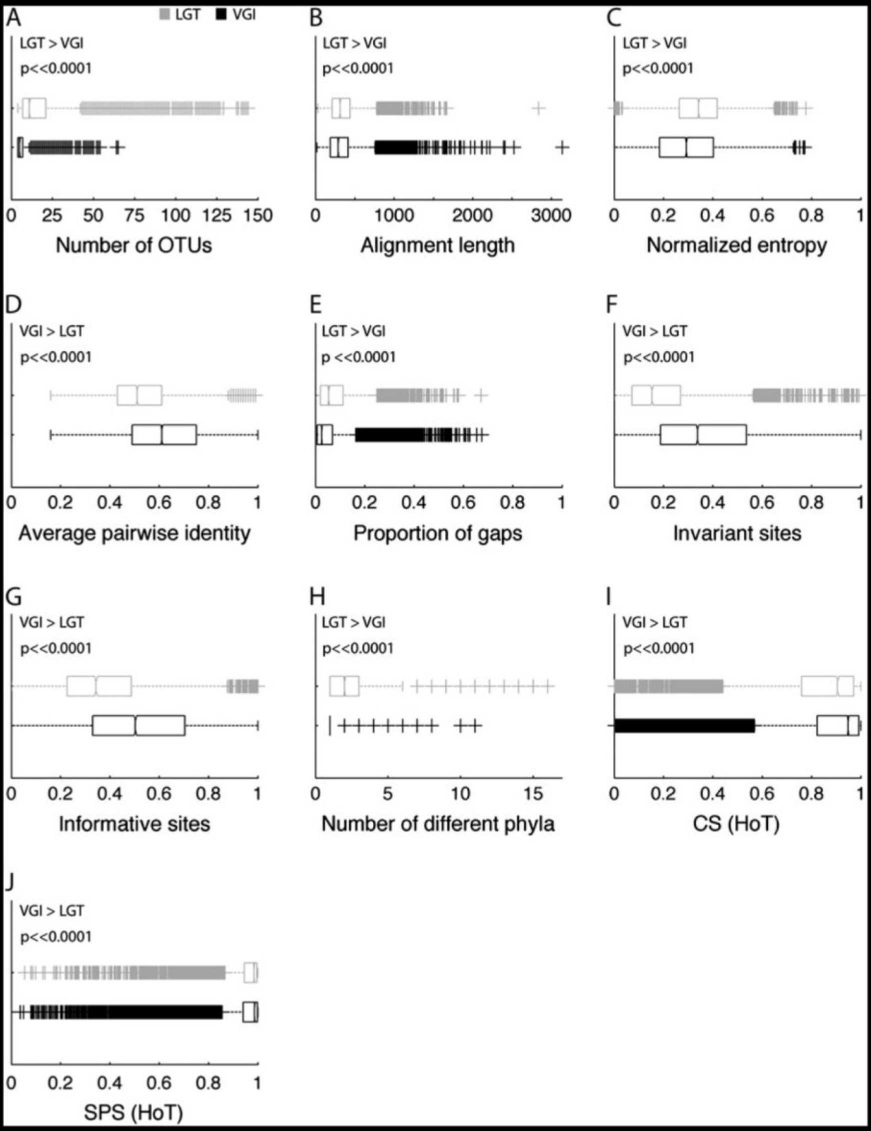


Fig. 1.—Distributions of alignment properties in the LGT and VGI groups. Differences in the distributions of the two groups were tested by the Wilcoxon nonparametric test (*P* values presented at the top of each graph).

SVM performance was evaluated by "accuracy," that is, the proportion of alignments correctly classified as LGT or VGI, "sensitivity," which is the true positive rate or the number of true positives (LGT alignments that are classified as such) divided by the sum of true positives plus false negatives (LGT alignments classified as VGI), and "specificity," as the true negative rate or the number of true negatives (VGI alignments that are classified as such) divided by the sum of true negatives plus false positives (VGI alignments classified as LGT).

To test the performance of the classifier under different LGT/VGI proportions in the training set and in the test set, we used LGT proportions ranging from 25% to 75% while including all 11 alignment properties.

To explore the contribution of the different features and their combinations to the classification performance, we tested all possible 2,047 combinations of the 11 alignment features analyzed in this study using a training set with equal proportions of LGT and VGI alignments.

Multivariate Analysis

We performed principal component analysis (PCA) using the princomp function of MATLAB 7.6. The data for each alignment property were normalized before the analysis, so that all properties had only values ranging from zero to one.

Results and Discussion

It is known that the probability of obtaining incorrect trees increases with the number of sequences (OTUs) analyzed (Nei 1996). The LGT alignments investigated here contained significantly larger numbers of OTUs ($P \ll 0.0001$) than the VGI alignments (fig. 1*A*). No VGI alignment in the present sample contains more than 65 sequences, whereas 5% of the LGT alignments contain ≥65 sequences. It is also known that for a given level of sequence divergence, the probability of obtaining incorrect trees is

Table 1
General Statistics of Protein Family Alignment Properties Grouped by LGT and VGI Categories

	Range (Min–Max)		Mean ± SD		Median	
MSA Parameter	VGI	LGT	VGI	LGT	VGI	LGT
Normalized Shannon entropy	0.000–0.772	0.002–0.777	0.294 ± 0.151	0.343 ± 0.117	0.292	0.341
Average pairwise identity	0.160–1.000	0.160–0.990	0.621 ± 0.176	0.523 ± 0.132	0.610	0.510
Proportion of gaps	0.000–0.675	0.000–0.671	0.052 ± 0.072	0.080 ± 0.084	0.025	0.053
Number of OTUs	4–65	4–144	6.6 ± 4.2	19.0 ± 21.7	5.0	11.0
Alignment length	14–3,135	31–2,837	325.0 ± 203.9	352.0 ± 212.3	287.0	311.0
Proportion of invariant sites	0.000–1.000	0.000–0.992	0.381 ± 0.247	0.191 ± 0.160	0.337	0.153
Proportion of informative sites	0.000–1.000	0.000–1.000	0.524 ± 0.250	0.366 ± 0.188	0.502	0.343
CS (HoT)	0.000–1.000	0.000–1.000	0.856 ± 0.212	0.823 ± 0.213	0.948	0.905
SPS (HoT)	0.037–1.000	0.053–1.000	0.939 ± 0.118	0.943 ± 0.111	0.986	0.984
Number of different phyla	1–11	1–16	1.323 ± 0.708	2.651 ± 2.475	1.0	2.0

NOTE.—The data set contains 22,437 protein family alignments, 5,822 of which are LGT and 16,615 are VGI.

higher when short sequences are analyzed than when longer sequences are analyzed (Nei 1996). However, the LGT protein families investigated here contain sequences that are significantly ($P < 0.0001$) longer than VGI protein families (supplementary fig. S1, Supplementary Material online), producing also longer alignments (mean = 352; table 1) than the VGI alignments (mean = 325; table 1) (fig. 1*B*), suggesting that if incorrect trees are involved in LGT inference in the present data, then short sequences are not the cause.

The probability of obtaining incorrect trees increases when sequence divergence becomes too great (Nei 1996). Several alignment properties can address the issue of sequence divergence. Normalized Shannon entropy provides an estimate for the average number of different amino acids that occur per site in an alignment (Valdar 2002). The mean normalized Shannon entropy of LGT alignments is about 17% higher than for the VGI alignments in the present data (fig. 1*C*), a highly significant difference ($P < 0.0001$). Average sequence identity across all pairwise comparisons is a very simple and robust measure of sequence variability in an alignment. The average pairwise identity of the VGI alignments (mean = 0.621; table 1) is significantly higher ($P < 0.0001$) than in the LGT alignments (mean = 0.523) (fig. 1*D*). In addition, LGT alignments contain on average 50% more gaps than VGI alignments (fig. 1*E*; table 1). Another proxy for sequence divergence in an alignment is the proportion of invariant sites, the mean of which is 2-fold higher ($P < 0.0001$) in the VGI alignments than in the LGT alignments (fig. 1*F*). Furthermore, the proportion of informative sites, defined here as alignment columns containing at least two different amino acids each observed in at least two sequences at the position, is significantly lower ($P < 0.0001$) in the LGT alignments (mean = 0.366; table 1) than in the VGI alignments (mean = 0.524; table 1) (fig. 1*G*).

Thus, several alignment parameters that are known to increase the probability of obtaining incorrect trees—higher numbers of OTUs, sequence divergence exceeding 50% differences on average, and low numbers of informative sites—are significantly different in the LGT and the VGI alignments, and in all cases, LGT alignments are skewed toward the value that increases the probability of obtaining an incorrect tree. This does not directly indicate that the LGT alignments have produced branches that are highly supported but nonetheless incorrect (Delsuc et al. 2003), yet the tendency is consistent.

The proportion of invariant sites, the proportion of informative sites, and average pairwise identity show an inverted trend to the Shannon entropy in the PCA of the total data set (fig. 2). These three measures correlate negatively with Shannon entropy ($r = -0.84$, $r = -0.82$, and $r = -0.97$, $P < 0.0001$, respectively, supplementary fig. S2A–C, Supplementary Material online). This means that the less variable alignments may lack phylogenetic information due to high proportions of invariable sites, where the proportion of informative sites in these alignments will still be high. Yet these correlation coefficients are weaker in the LGT alignments ($r = -0.70$ and $r = -0.77$ and $r = -0.95$, $P < 0.0001$, respectively, supplementary fig. S2D–F, Supplementary Material online), than in the VGI alignments ($r = -0.87$ and $r = -0.84$ and $r = -0.98$, $P < 0.0001$, respectively; supplementary fig. S2G–I, Supplementary Material online) so that even though the LGT alignments are more variable than the VGI alignments, they generally contain not only fewer invariant sites but also fewer informative sites.

High alignment variability in the LGT alignments could be also the result of large numbers of sequences per alignment, as is the case for the LGT group alignments (fig. 1*A*). However, we found no correlation between number of OTUs and normalized entropy ($r = 0.01$, $P = 0.27$), and only weak correlation between number of OTUs and average pairwise identities ($r = -0.11$, $P < 0.0001$), or the proportion of gaps ($r = 0.23$, $P < 0.0001$; supplementary fig. S3, Supplementary Material online). Also, the number of phyla represented in the alignment, another possible source for higher alignment variability, is higher in the LGT groups than in the VGI group (fig. 1*H*). But this measure as well shows no significant correlation with any of the variability measures (supplementary fig. S4, Supplementary Material online). Hence, the high variability of the LGT alignments is not explained by the large number of sequences or the large number of phyla represented in these families.

The more variable the sequences in an alignment are, the more difficult they are to align and the more likely it is that the alignment procedures themselves can produce collections of site patterns that induce topological effects at the tree-building stage (Landan and Graur 2007; Wong et al. 2008). Thus, the LGT alignments, which are more variable than those in the VGI group, might be more error prone at the alignment step than the VGI alignments. To estimate

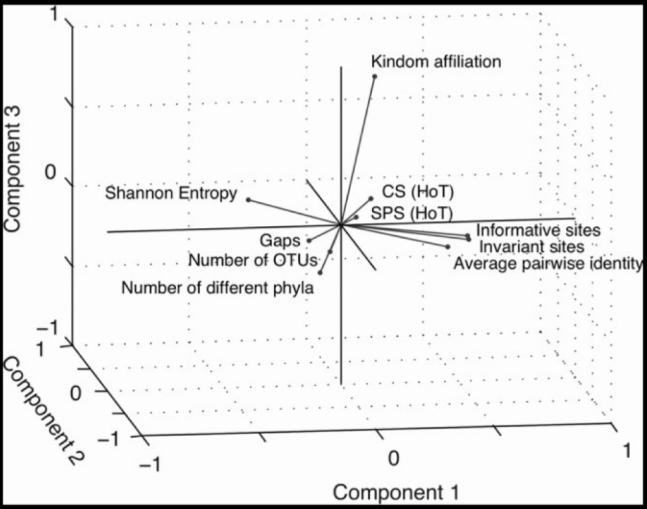


Fig. 2.—Principal component analysis of alignment properties. The axes represent the first three components, explaining 85% of the variability in the data (see supplementary tables S3 and S4, Supplementary Material online for details). Alignment properties are represented as vectors of their principal component coefficients. Alignment length is omitted due to its marginal contribution to the first three principal components. Two-dimensional views of every two respective components can be found in supplementary fig. S7, Supplementary Material online.

this effect, we compared alignment reliability of LGT and VGI alignments using the HoT method (Landan and Graur 2007). For these HoT comparisons, we realigned the original sequences (i.e., before filtering with Gblocks) as kindly provided by Beiko, Harlow, and Ragan (2005) in the C-to-N-direction to form the Tails alignments and compared them with the original (Heads) alignments. Both HoT parameters show an inverted trend to the number of OTUs, number of different phyla, and the proportion of gaps in the PCA analysis (fig. 2). Moreover, we found that the LGT alignments have a significantly ($P < 0.0001$) lower CS, which is the proportion of site columns reconstructed identically in the Heads and Tails alignment (fig. 1*I*), and a slightly but significantly ($P \ll 0.0001$) lower SPS, which is the proportion of identically reconstructed site pairs (fig. 1*J*), than the VGI alignments. Hence, LGT alignments contain significantly more alignment artifacts that are introduced by the sequence alignment process alone, independent of subsequent tree-building procedures. The bias in alignment quality within the LGT set is unlikely to be related with the erroneous guide tree used for the alignment because alignment errors are only marginally affected by the guide-tree quality (Landan and Graur 2008). Beiko, Harlow, and Ragan (2005) used a very conservative rule for inclusion in the LGT set that comprises only trees having at least one highly significant ("posterior probability" ≥ 0.95) discordant branch, whereas all other trees are considered as VGI. This results in abias toward highly supported (though not necessarily true) trees in the LGT set, where the proportion of highly significant branches per tree is 47 ± 28% versus 48 ± 40% (median 41% vs. 33%) in the VGI set. To test if this bias is related to the differences we found in the alignment properties, we deleted from the VGI set those alignments yielding trees with no highly significant branches, leaving 13,811 alignments yielding trees having at least one highly significant (posterior probability ≥ 0.95) concordant branch. This resulted in a set of trees, designated here VGI95, having a much higher proportion of highly significant branches per tree (57 ± 37%, median 50%). A comparison of alignment properties between the LGT and VGI95 sets resulted in identical conclusions to those detailed above for the comparison between the LGT and VGI sets (supplementary fig. S5, Supplementary Material online), so that the bias toward highly resolved trees in the LGT set has no relation to the bias in multiple alignment properties.

The comparison of alignment properties between VGI and LGT alignments summarized so far (fig. 1; table 1) shows that the LGT alignments are more variable than the VGI alignments. It is thus possible that the laterally transferred protein-coding sequences are inherently more variable than vertically inherited ones. To test this possibility, we compared pairwise protein distances between genomes to see if there were differences between LGT and VGI sequences with respect to overall sequence conservation. If so, then orthologous sequence pairs from VGI alignments should have smaller protein distances (i.e., should be more conserved) than orthologous pairs from LGT alignments. We tested that hypothesis for seven frequent genome pairs having proteins in both groups. The contrary was observed: Orthologous pairs from LGT alignments are more conserved (i.e., have smaller protein distances) than orthologous pairs from VGI alignments (fig. 3; supplementary fig. S6, Supplementary Material online). Hence, the higher variability observed in LGT alignments cannot be explained by a systematic bias in protein conservation among inherited versus laterally transferred proteins. This conclusion seems to contradict the bias toward variable multiple sequence alignments in the LGT set. A possible reconciliation between these two findings may be found in a study by Elhaik et al. (2006) showing that conserved proteins have higher probability of being detected by a similarity search, which leads to the composition of larger protein families, hence alignments with more OTUs that are probably more difficult to align. However, we found no correlation between the number of OTUs and the different alignment variability measures in our data set (supplementary fig. S3, Supplementary Material online).

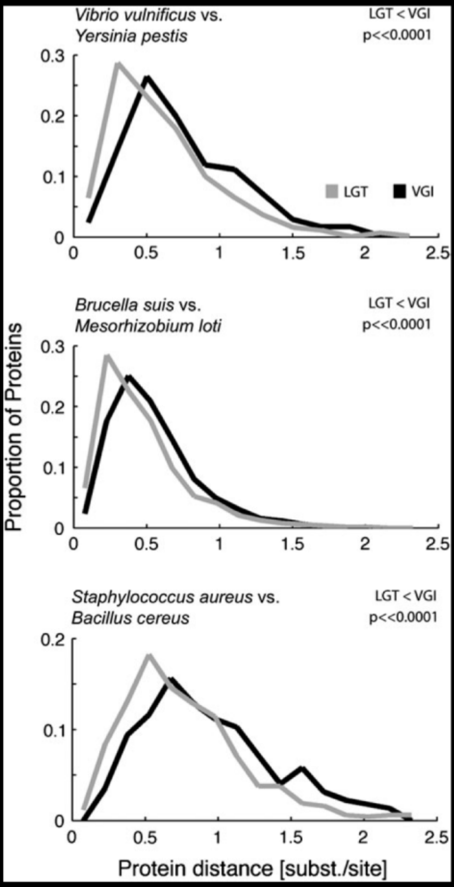


FIG. 3.—Comparison of protein pairwise distances between genome pairs found in LGT and VGI families. Distance distributions were compared using the Wilcoxon nonparametric test (*P* values presented at the top right of the graph).

Clustering of conserved orthologous proteins results not only in bigger families, but also in families having proteins from many taxomonic groups (supplementary fig. S4A, Supplementary Material online). However, the number of phyla alone is unlikely to reflect the relatedness among the sequences in the protein family because a protein family including sequences from eubacteria and archaebacteria is expected to contain more variability than a protein family including sequences from eubacteria only. Therefore, we divided the multiple alignments into those that comprise 1) eubacterial proteins only, 2) archaeal proteins only, and 3) "universal" alignments including proteins from both groups. A comparison of alignment properties among these three categories shows that universal families are much bigger than either eubacterial or archaeal families (fig. 4*A*). Moreover, all variability measures show that the universal alignments are more variable than alignments in the other two categories: Their entropy is higher (fig. 4*C*), their mean pairwise distance is higher (fig. 4*D*), and they contain more gaps (fig. 4*E*) and less invariant sites (fig. 4*F*). Universal alignments also seem to be of lower quality, in that they contain fewer informative sites and their alignments are less reliable (fig. 4*G–I*).

Finally, we tested for dependency between the taxonomical composition of the alignments and their classification as VGI or LGT and found these two properties are significantly dependent ($P < 0.001$, using χ^2 test). LGT alignments comprise about 30% of the archaeal or eubacterial alignments (as in the total data), but they are overrepresented in the universal group where they comprise 57% of the multiple alignments (fig. 4*J*). This leads us to conclude that the clustering of conserved sequences resulted in protein families that are not only large (as predicted by Elhaik et al. 2006) but also have a universal taxonomic distribution that covers much more diverse sequences and that seems to be the reason for their variability.

Our results so far suggest that alignments possessing properties that are known to increase the probability of obtaining incorrect branches are more frequent in the LGT group than in the VGI group. We then asked a slightly heretical question: Can we predict whether an alignment is likely to generate a tree with a strongly supported discordant branch on the basis of alignment properties alone? For this, we used an SVM classifier (Christiani and Shawe-Taylor 2000). In brief, a SVM is an algorithm that, provided with a learning set of features that might or might not correlate to a classificatory decision of the type "yes" or "no," gains experience with the learning set, and then is asked to classify objects, correctly if possible, on the basis of features alone. In the present case, the features correspond to alignment parameters as summarized in figure 1 and table 1, and the desired classification is the proper assortment of the alignment into LGT or VGI groups as predetermined by phylogenetic analysis. The classification performance is evaluated by its accuracy, sensitivity, and specificity (see Methods).

The SVM algorithm was thus trained and queried using the present alignments. In order to calculate SVM performance and standard deviations (SDs), we performed a 10-fold crossvalidation using 1/10 of the data in each step for training and the rest for testing. Accuracy, sensitivity, and specificity of the classifying process are to a vast extent influenced by the ratio of LGT/VGI in the training set. Accuracy and sensitivity are maximal when the proportion of LGT in the training set is equal to the total data (25%) and they decrease when higher LGT proportions are used. The specificity of the SVM classification is minimal at 25% LGT alignments in the training set and increases when higher proportions are used. When LGT proportion in the training set is fixed to 50%, all SVM performance measures are found in equilibrium (fig. 5*A*). The ratio of LGT/VGI alignments in the test set has no influence on the performance of the SVM classifier (fig. 5*B*). In our SVM classification procedure, we used training sets having an LGT/VGI ratio = 1 (see Methods). Performance of the classifying process was evaluated by trying all possible 2,047 combinations of the 11 properties to explore if there is a set of features that, if omitted from the training process, will deteriorate the results, or if there are some features that tend to impair the performance when included in the analysis.

Table 2 shows the combination of features that yielded the top performance values of accuracy, sensitivity, and specificity. We cannot really decide which is the best combination of feature vectors to be included in the training process because widely different combinations of features induce consistent results in the classification performance. But it seems that for equally high performance values for

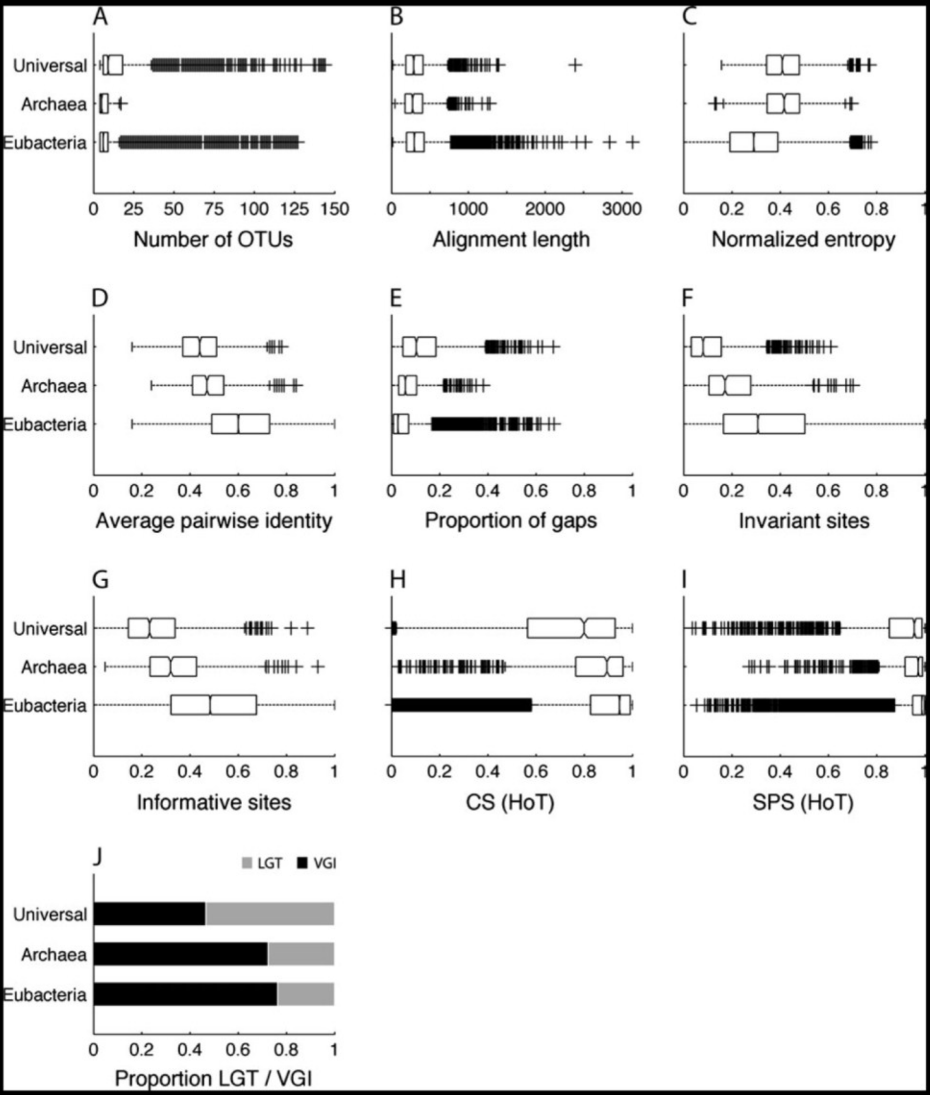


FIG. 4.—Differences in alignment properties for alignments containing only eubacterial sequences, only archaebacterial sequences, or sequences of both kingdoms (universal).

the three parameters (e.g., combinations yielding accuracy = 0.797 or accuracy = 0.796), the number of OTUs, entropy, average pairwise identity, and number of phyla are of particular importance. A complete table with all 2,047 tested combinations of features can be found in supplementary table S1, Supplementary Material online.

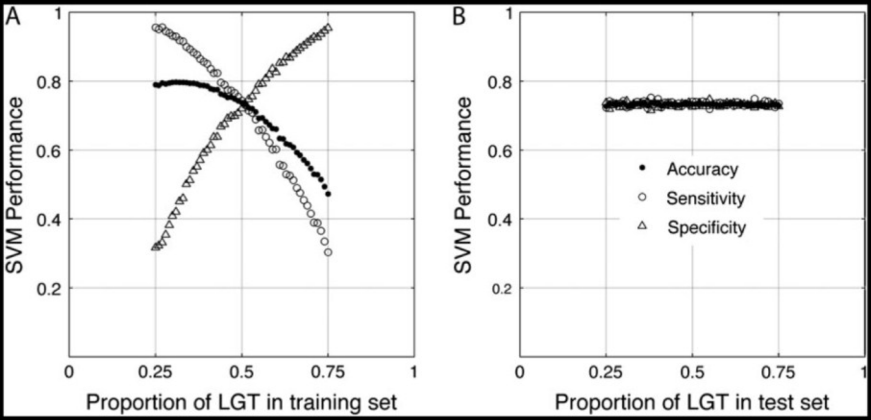


FIG. 5.—Performance of the classifier under different LGT proportions in the training set (*A*) and in the test set (*B*). In (*B*), the LGT/VGI ratio was adjusted to 1.

Table 2
Prediction of LGT/VGI Using an SVM Classifier Trained with Alignment Properties

Combination of Training Parameters													
											Performance		
Number of OTUs	Shannon Entropy	Average Pairwise Identity	Proportion of Gaps	Proportion of Invariant Sites	Proportion of Informative Sites	Alignment Length	CS (HoT)	SPS (HoT)	Number of Different Phyla	Kingdom Affiliation	Accuracy	Sensitivity	Specificity
v	*v*								*v*		0.797 ± 0.009	0.833 ± 0.020	0.692 ± 0.023
v	*v*							v	*v*		0.796 ± 0.009	0.835 ± 0.019	0.685 ± 0.021
v	*v*		*v*						*v*		0.796 ± 0.007	0.834 ± 0.015	0.687 ± 0.021
v	*v*				*v*				*v*		0.796 ± 0.009	0.830 ± 0.017	0.699 ± 0.017
v		*v*									0.794 ± 0.009	0.832 ± 0.020	0.688 ± 0.025
v		*v*			*v*				*v*		0.794 ± 0.012	0.828 ± 0.021	0.698 ± 0.018
v	*v*					*v*			*v*		0.794 ± 0.007	0.833 ± 0.015	0.685 ± 0.019
v	*v*	*v*			*v*				*v*		0.794 ± 0.012	0.827 ± 0.025	0.700 ± 0.025
v					*v*				*v*		0.794 ± 0.009	0.831 ± 0.019	0.687 ± 0.021
v		*v*					*v*		*v*		0.793 ± 0.008	0.829 ± 0.016	0.692 ± 0.020
										v	0.735 ± 0.009	0.931 ± 0.020	0.174 ± 0.022
							v	v	*v*	v	0.743 ± 0.008	0.887 ± 0.087	0.331 ± 0.247
								v		v	0.711 ± 0.041	0.884 ± 0.081	0.218 ± 0.074
			v					v	*v*	v	0.733 ± 0.012	0.873 ± 0.091	0.334 ± 0.229
v							*v*	v	*v*	v	0.765 ± 0.019	0.867 ± 0.093	0.474 ± 0.327
v							*v*	v	*v*		0.784 ± 0.018	0.851 ± 0.061	0.592 ± 0.215
v			*v*				*v*	v			0.771 ± 0.013	0.849 ± 0.066	0.551 ± 0.216
v			*v*					v	*v*		0.778 ± 0.015	0.848 ± 0.063	0.576 ± 0.224
			v				*v*	v	*v*	v	0.736 ± 0.012	0.848 ± 0.091	0.419 ± 0.235
v			*v*				*v*			v	0.762 ± 0.014	0.847 ± 0.082	0.519 ± 0.274
		v									0.544 ± 0.015	0.450 ± 0.031	0.812 ± 0.033
		v			*v*			v			0.598 ± 0.013	0.529 ± 0.025	0.794 ± 0.024
		v						v			0.576 ± 0.011	0.501 ± 0.023	0.790 ± 0.025
		v			*v*	*v*	*v*				0.586 ± 0.013	0.516 ± 0.027	0.787 ± 0.028
				v	*v*	*v*					0.665 ± 0.019	0.623 ± 0.034	0.786 ± 0.025
		v			*v*			v		v	0.601 ± 0.018	0.536 ± 0.037	0.785 ± 0.037
		v			*v*		*v*				0.590 ± 0.012	0.522 ± 0.028	0.785 ± 0.034
		v	*v*		*v*		*v*				0.592 ± 0.019	0.525 ± 0.040	0.783 ± 0.043
		v			*v*						0.582 ± 0.011	0.511 ± 0.024	0.782 ± 0.027
		v					*v*				0.562 ± 0.014	0.484 ± 0.030	0.782 ± 0.029

NOTE.—Alignment properties included in the training process are marked with *v*. The LGT/VGI ratio in the training set was adjusted to 1. Only combinations yielding the best 10 performance values for accuracy, sensitivity, and specificity are shown, respectively. Definitions of accuracy, sensitivity, and specificity can be found in the text. A table presenting the performance of all possible combinations is presented in the Supplementary Material online.

In other words, the alignment properties of the LGT and VGI groups, although having strongly overlapping distributions for all parameters (fig. 1; table 1), are nonetheless sufficiently different in a consistent manner that we can correctly predict about 78% of the time whether a Bayesian phylogenetic inference will generate a branch from a given alignment that is sufficiently discordant to be scored as an LGT. On the strength of this finding and circumstance that for each alignment parameter the LGT alignments were always skewed toward values that are known from simulation studies to generate incorrect branches (Nei 1996), it is likely that reliable construction of phylogenetic trees is affected and incorrectly reconstructed branches may be a possible source of LGT inference. The correlations are consistent with the view (Landan and Graur 2007) that sequence sets problematic at the level of alignments are likely to be problematic at the level of phylogenetic inference as well.

In principle, one could use our trained SVM on other alignment data sets in order to predict which alignments will result in discordant branches comparing with a reference tree. However, one would still have to distinguish between discordant branches stemming from either genuine LGTs or phylogenetic reconstruction artifacts. The results presented here indicate that the latter are more frequent in problematic alignments; hence, alignment quality has high impact on evolutionary inference from phylogenetic trees. A similar observation was recently presented for phylogenetic inference of ancient LGTs during the endosymbiosis of plastids (Deusch et al. 2008). This indicates that it is important to monitor and assess alignment quality in large-scale phylogenetic analyses, particularly those implementing automated or semiautomated phylogeny pipelines.

Supplementary Material

Supplementary tables S1–S4 and supplementary figures S1–S6 (and additional supporting figures) are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Giddy Landan and David Bryant for discussions. This work was funded by the Deutsche Forschungsgemeinschaft (M.R., W.M.), and the German-Israeli Foundation for scientific research and development (T.D.). We would like to thank the central computing facility of Düsseldorf University.

Literature Cited

Beiko RG, Chan CX, Ragan MA. 2005. A word-oriented approach to alignment validation. Bioinformatics. 21:2230–2239.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci. 102:4332–14337.

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

Christiani N, Shawe-Taylor J. 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge (MA): Cambridge University Press.

Delsuc F, Phillips MJ, Penny D. 2003. Comment on "Hexapod origins: monophyletic or paraphyletic?". Science. 301:1482d.

Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Mol Biol Evol. 25:748–761.

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

Elhaik E, Sabath N, Graur D. 2006. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. Mol Biol Evol. 23:1–3.

Felsenstein J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266:418–427.

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

Gotoh O. 1996. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol. 264:823–838.

Grasso C, Lee C. 2004. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics. 20:1546–1556.

Hillis DM. 1995. Approaches for assessing phylogenetic accuracy. Syst Biol. 44:3–16.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl Acids Res. 30:3059–3066.

Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.

Landan G, Graur D. Forthcoming. 2008. Characterization of pairwise and multiple sequence alignment errors. Gene, doi: 10.1016/j.gene.2008.05.016.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol. 19:1–7.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 320:1632–1635.

Nei M. 1996. Phylogenetic analysis in molecular evolutionary genetics. Annu Rev Genet. 30:371–403.

Nei M, Takezaki N, Sitnikova T. 1995. Assessing molecular phylogenies. Science. 267:253–254.

Notredame C, Higgins DG, Heringa J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302:205–217.

Penny D, Hendy MD, Steel M. 1992. Progress with methods for constructing evolutionary trees. Trends Ecol Evol. 7:73–79.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucl Acids Res. 33:501–504.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res. 22:4673–4680.

Valdar WSJ. 2002. Scoring residue conservation. Proteins. 48:227–241.

Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. Science. 319:473–476.

Dan Graur, Associate Editor

Accepted May 5, 2009

7.2 LGT frequency during microbial evolution

Dagan T, Martin W: **Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution**. *Proc Natl Acad Sci USA* 2007, **104**:870-875.

(Own contribution: conceived and designed the experiment, performed the analysis, analyzed the data, and wrote the paper).

Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution

Tal Dagan* and William Martin

Institut für Botanik III, Heinrich-Heine Universität, Universitätsstrasse 1, 40225 Düsseldorf, Germany

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Nova Scotia, Canada, and approved November 21, 2006 (received for review July 25, 2006)

The amount of lateral gene transfer (LGT) that has occurred in microbial evolution is heavily debated. Efforts to quantify LGT through gene-tree comparisons have delivered estimates that between 2% and 60% of all prokaryotic genes have been affected by LGT, the 30-fold discrepancy reflecting differences among gene samples studied and uncertainties inherent in phylogenetic reconstruction. Here we present a simple method that is independent of gene-tree comparisons to estimate the LGT rate among sequenced prokaryotic genomes. If little or no LGT has occurred during evolution, ancestral genome sizes would become unrealistically large, whereas too much LGT would render them far too small. We determine the amount of LGT that is necessary and sufficient to bring the distribution of inferred ancestral genome sizes into agreement with that observed among modern microbes. Rather than testing for phylogenetic congruence or lack thereof across genes, we assume that all gene trees are compatible; hence, our method delivers very conservative lower-bound estimates of the average LGT rate. The results indicate that among 57,670 gene families distributed across 190 sequenced genomes, at least two-thirds and probably all, have been affected by LGT at some time in their evolutionary past. A component of common ancestry nonetheless remains detectable in gene distribution patterns. We estimate the minimum lower bound for the average LGT rate across all genes as 1.1 LGT events per gene family and gene family lifespan and this minimum rate increases sharply when genes present in only a few genomes are excluded from the analysis.

microbial evolution | phylogenomics | gene clusters

Few topics in evolutionary microbiology are as controversial as lateral gene transfer (LGT). Views on the issue span from one extreme that LGT exists but is insignificant in terms of its overall impact on the evolutionary process, such that a tree of microbial phylogeny can be reliably constructed (1–3), to the other extreme that LGT occurs in nature to such an extent that a simple bifurcating tree is an inadequate metaphor to represent the process of microbial evolution (4, 5). Efforts to resolve this debate have focused on attempts to quantify LGT frequency through evolutionary genome comparisons but are impaired by methodological issues.

There are currently three main approaches to quantifying LGT. The first involves identification of codon usage, GC content, or nucleotide-pattern properties within genomes that differ from the genomic norm and hence are likely to represent acquired sequences (6–8). This approach is powerful but can uncover only recent LGT events. The second approach involves gene-tree comparisons in search of incongruent branching patterns. This approach has delivered widely conflicting results, ranging from estimates that up to 60% of all genes are affected by LGT (9) to estimates that as few as 14% (10) or even only 2% are affected (11). The reason for such divergent quantitative estimates is primarily founded in the uncertainties inherent to phylogenetic reconstruction by using real data (12–14) and in differences among investigated gene and genome samples. A third approach entails inference of gene-gain and -loss events (15–18). Estimates using this approach, in which gene losses are weighted against gene acquisitions (LGTs) according to a predetermined loss-to-LGT ratio, suggest that between 40% (16) and 90% (17) of all gene families might be affected by LGT; these discrepancies are caused by different *a priori* specified gain/loss ratios and the genome samples studied.

An additional approach to inferring LGT but hitherto in a nonquantitative manner involves the identification of genes showing patchy distribution patterns across genomes (19, 20). Although differential gene loss can account for patchy distributions in individual instances, it cannot be invoked to account for all such patterns, because the inferred size of ancestral genomes would become unrealistically large. We reasoned that this phenomenon, which Doolittle *et al.* (21) have termed the "genome of Eden," could be used to estimate the rate of LGT. Given the current distribution of genes across genomes and a reference tree, one can calculate ancestral genome sizes under the assumption that all gene distributions are due to gene loss only. If ancient genome sizes become unrealistically large, incremental allowance of LGT should solve the genome-of-Eden problem, and the amount of LGT that causes inferred ancestral genome sizes to assume a size distribution similar to modern ones would be an estimator of the LGT rate.

However, what if a given gene tree is different from the reference tree? Here, we grant each gene the full benefit of all phylogenetic doubt; we assume (*i*) all gene trees are perfectly compatible with the same reference tree, (*ii*) gene loss is unpenalized, and (*iii*) no paralogy; that is, all within-genome duplications for each gene family are assumed to have occurred subsequent to the last divergence for each lineage. Taken together, these three assumptions mean we infer no LGTs from phylogenetic conflicts; hence, our approach delivers conservative lower-bound constraints for the minimum LGT rate during prokaryote genome evolution.

Results

The Distribution of Genes Across Genomes. Using the standard method (22), we clustered all 562,321 protein-coding genes present within 190 prokaryotic genomes [supporting information (SI) Table 4

www.pnas.org/cgi/content/full/0606318104/DC1.

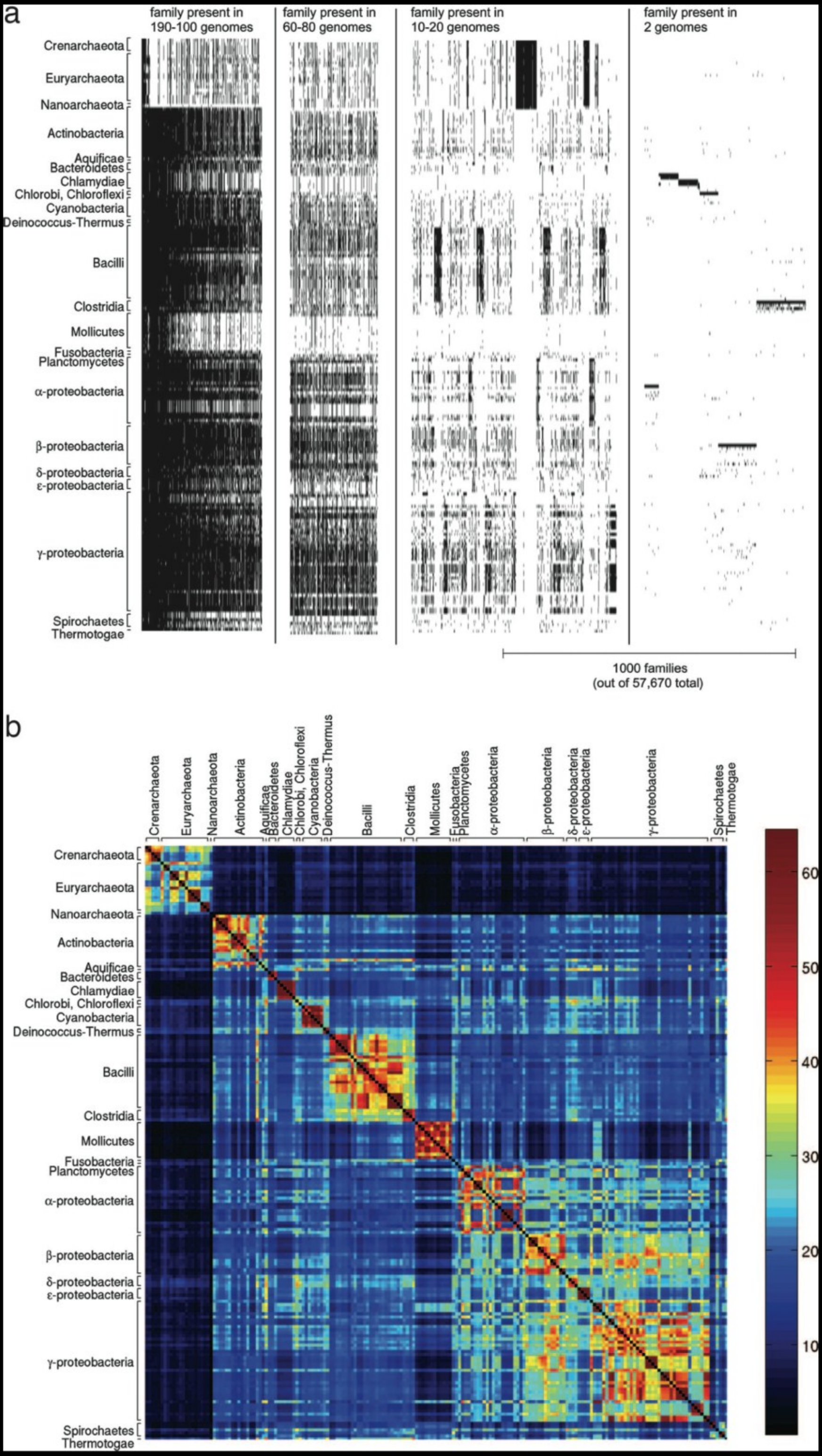


Fig. 1. The distibution of genes across genomes. (*a*) Presence (black) and absence (white) patterns for representative segments of the data comprising widely (present in 100–190 genomes), intermediately (60–80 and 10–20 genomes), and sparsely distributed genes (two genomes). (Note the scale bar.) (*b*) Color-coded matrix of the proportion of shared genes for all genome pairs, with genomes grouped by taxonomical classification. For the same matrix using random genome order, see SI Fig. 4

genomes. Consistent with earlier findings (23), the proportion of shared families among genomes from different prokaryotic groups uncovers components of both vertical and horizontal inheritance (Fig. 1*b*).

Ancestral Genome Sizes Constrain the Average LGT Rate. To estimate the minimum amount of LGT in the present gene-distribution data, we first plotted the PAPs onto a reference tree for the rRNA

operon (SI Fig. 5

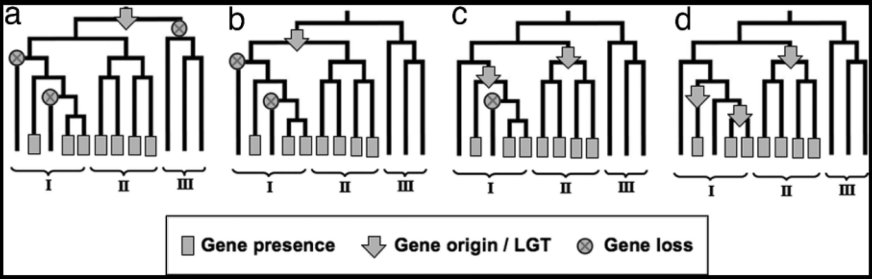


Fig. 2. Gene loss and LGT can both account for patchy gene distributions. Schematic representation of four different LGT allowances. (*a*) In the loss-only model, all genes are assumed to have originated at the root of the tree; PAPs are attributed to gene loss only. (*b*) Introducing a gene origin in the SO model disperses gene origins over internal nodes of the tree according to their first occurrence. (*c*) In the $LGT_{\leq 1}$ model, each gene is allowed to have two origins, where one is an LGT. This model results in further dispersal of gene origins across the tree, hence smaller ancestral genomes. (*d*) Two additional LGTs are allowed in the $LGT_{\leq 3}$ model. Allowances of up to 7, 15, and 31 LGTs were also tested.

by 26-fold and the largest genome in our sample (8,317 families; *Bradyrhizobium japonicum*) by 7-fold (Table 1).

Such burgeoning genome sizes are indeed unrealistic, but so is the notion that new genes do not arise during evolution. Allowing new genes to arise over time according to the single-origin (SO) model (Fig. 2*b*) yields an ancestral prokaryote genome that contains 3,081 genes, those present in both archaebacteria and eubacteria (Fig. 3*a*). However, the SO model does not solve the genome-of-Eden problem; it merely transfers it into the middle ages of microbial evolution, where ancestral genome sizes soar once more to 12,000–14,000 genes, sizes that far exceed those observed among modern organisms (Table 1).

Thus, we either have to embrace the untenable assumption that microbial genome sizes were fundamentally different in the past than they are today or, preferably, we have to allow some amount of LGT. How much LGT is necessary to bring ancestral genome sizes into agreement with the observed contemporary range?

We started by allowing only one LGT event per family (Fig. 2*c*), the $LGT_{\leq 1}$ model. This model allows each gene to have two origins, one of which is an LGT. For 35% of our families, neither LGT nor loss is required; the remaining 65% accept one LGT. This average LGT rate of 0.65 LGT per family (Table 2) brings inferred ancestral genome sizes down to <8,000 genes (Fig. 3*b*), with a maximum of 7,607 and a mean of 2,858, closer to contemporary genomes (Table 1) but still a bit too large.

We tested additional evolutionary models allowing up to 3, 7, 15, or 31 LGTs per gene family. With increasing amounts of LGT allowed, inferred ancestral genome sizes shrink, as do the numbers of inferred gene losses per gene family (Fig. 2*d* and Table 2). Although most gene families do not require more than one LGT to map exactly onto the reference tree (Table 2), they are also quite small and offer little opportunity to observe LGT (SI Fig. 6

Table 1. Modern and last-common ancestor (L-ca) genome sizes under different LGT allowances

		Inferred ancestral genome size (number of families)						
Genome	Modern genomes*	Loss only	Single origin	$LGT_{\leq 1}$	$LGT_{\leq 3}$	$LGT_{\leq 7}$	$LGT_{\leq 15}$	$LGT_{\leq 31}$
$L_{prokaryotic}ca$	2,198	57,670	3,081	2	2	2	2	2
$L_{archeabacterial}ca$	1,573	9,453	3,240	148	91	91	91	91
$L_{eubacterial}ca$	2,297	53,658	3,573	443	35	35	35	35
$L_{proteobacterial}ca$	2,690	35,903	13,652	5,872	3,303	2,119	1,517	1,147
$L_{cyanobacterial}ca$	2,187	5,526	10,509	3,598	1,938	1,306	1,014	886
$L_{actinobacterial}ca$	2,602	11,611	10,233	3,461	1,691	1,044	703	520
$L_{mollicute}ca$	432	1,714	660	557	485	415	352	300
Mean†	2,198	8,142	7,296	2,858	2,234	1,868	1,634	1,472
Ancestral vs. modern‡		<0.01	<0.01	<0.01	0.71	<0.02	<0.01	<0.01

*Average genome size for the group.

†For 190 modern genomes, for 187 ancestral genomes.

‡Probability that the two samples come from distributions of equal medians using the Wilcoxon Mann–Whitney test.

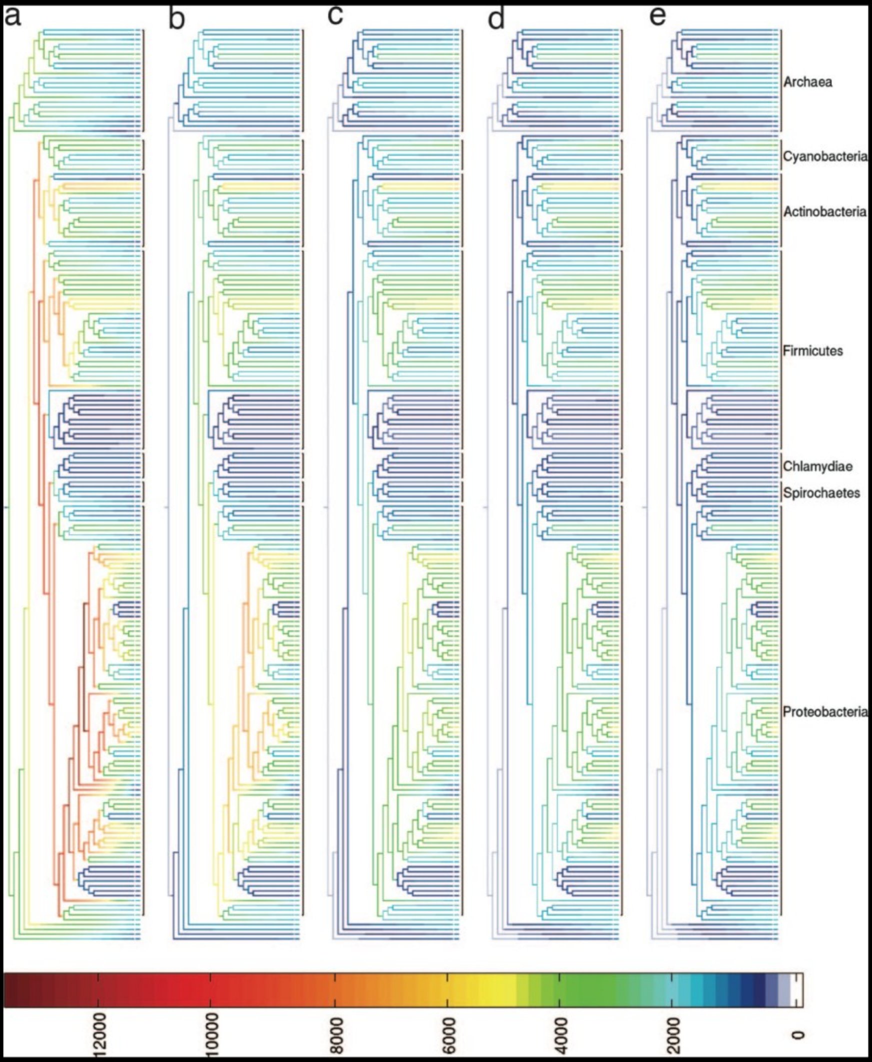


Fig. 3. Ancestral genome sizes reconstructed under the various reconstruction models. The colors of nodes and branches correspond to the inferred ancestral genome size, as indicated in the scale. *a*–*e* correspond to the SO, $LGT_{\leq 1}$, $LGT_{\leq 3}$, $LGT_{\leq 7}$, and $LGT_{\leq 15}$ models, respectively (see SI Figs. 7 and 8

The Tree Is Not Too Important, but Family Size Is. Neither different reference trees (using maximum likelihood with or without a gamma distribution of rate variation across sites or using ribosomal protein sequences) nor alternative rootings (within proteobacteria, actinobacteria, or mollicutes) affected the average LGT rate across all genes by >10% (SI Table 5

SI Fig. 6

SI Fig. 6

SI Table 5

SI Table 5

SI

Fig. 4

SI Fig. 9

SI Table 5

Discussion

The frequency of LGT affects inferred ancestral genome size. No LGT results in untenably large ancestral genomes, whereas too many LGTs result in untenably small ancestral genomes. With the present sample, an average LGT rate on the order of ≈1.1 events per family per family life span (Fig. 3 and Table 1) provides the best fit of inferred ancestral genome sizes to those currently observed in real microbes. This average LGT rate is a very conservative lower bound, because it is based on the assumptions that all families investigated contain orthologs only, and that all gene trees are compatible.

One could argue that ancient genomes were bigger than those of today, and that the amount of LGT inferred here is still not necessary. Indeed, it has been suggested that the vast majority of all LGT occurred before the origin of cells, and that little or none has occurred since (1). However, this suggestion cannot be true, because nucleotide-pattern comparisons indicate that LGT is still an ongoing process today (6–8). One could also argue that ancient genomes were much skimpier than those of today, and that they inflated only recently in a case of evolutionary last-minute shopping, such that higher average LGT rates than those inferred here would be tenable. However, in the absence of evidence to the contrary, Occam's razor would prefer the simpler premise that genome sizes, rates of loss, and rates of LGT in the past, on average, were not fundamentally different from those of today. In the $\text{LGT}_{\leq 3}$ model, genome size is not only similar to the values currently observed among prokaryotes (Tables 1 and 3); it is also far more constant across time than in the other models (Fig. 3). The same is true for gene-origin and -loss frequencies (SI Fig. 10

Above and beyond our full-benefit assumptions, the lower-bound nature of our ≈1.1 LGT per family estimate has two further caveats. First, it is possible that the first origin we infer for each gene is not a birth event, but itself is an LGT from an unsampled genome. Although no genome sample size would exclude that possibility, if we assume that all families were born outside rather than within the lineages sampled, our estimate for the average rate would increase only to ≈2.1. Second, our methods count only observed events; unobserved gene families or events (27) are disregarded.

Our findings indicate that LGT occurs very frequently among prokaryotes in terms of having impact upon individual gene family distributions, in that at least 65% of all families (and given the ultraconservative nature of our full benefit assumptions, probably all) have been affected during the course of evolution. These results can be taken as support for the view that a core set of genes that has remained immune to LGT throughout all of evolution is unlikely to exist (28, 29). The estimates of the average LGT rate reported here represent solely the amounts required to keep ancestral genome-size distributions within realistic bounds; additional contributions from gene-tree comparisons or nucleotide-pattern analyses were not considered. However, despite much LGT, gene-distribution patterns are still nonrandom, as Fig. 1*b* attests. Further specification of the extent of this important process of natural variation among prokaryotes is germane to understanding the evolutionary mechanisms that govern the distribution of genes across genomes.

Materials and Methods

Data. Completely sequenced prokaryotic genomes were downloaded from the National Center for Biotechnology Information (NCBI) web site (www.ncbi.nlm.nih.gov; August 2005 version). For each species, only the strain with the largest number of genes was used. Of 190 genomes (562,321 proteins) in the data, 22 are archaebacterial, and 168 are eubacterial.

Gene Families (PAPs). All proteins in the 190 genomes were clustered by similarity into gene families by using the reciprocal best BLAST hit (BBH) approach. 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 (30) to obtain amino acid identities. Using a cutoff of 30% amino acid identity in the clustering procedure (22), the proteins fell into 57,670 families with two or more members, in addition to 149,894 singletons that were excluded from pattern analysis. The resulting gene families represent a PAP of gene distribution across the prokaryotic genomes. A PAP includes 190 digits; if a gene family includes one or more genes from genome *i*, then digit x_i in its corresponding pattern is "1"; otherwise, it is "0."

Reconstruction of Phylogenetic Trees. For the reference tree, the sequences of the rRNA operon (16S, 23S, and 5S) from all 190 genomes were aligned by using ClustalW (30) and concatenated, and gapped sites were removed. This alignment was used for phylogenetic reconstruction by maximum likelihood (ML) with and without rate variation using dnaml (31), PhyML (32), and Neighbor Joining (33). The tree of concatenated L11 and S11 ribosomal protein sequences was inferred with PhyML (32). Trees were rooted between archaebacteria and eubacteria; additional roots of the dnaml tree (between proteobacteria, actinobacteria, or mollicutes and other genomes) were tested. The random tree was obtained by shuffling species names in the ML tree and rooting on the longest internal edge. All Newick format trees are provided in SI Table 6

8 Phylogenomic networks.

8.1 The cummulative impact of LGT in prokaryote genome evolution

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

Featured in Editor's Choice. *Science* 2008, **321**:747.

(Own contribution: conceived and designed the experiment, performed the analysis, analyzed the data, and wrote the paper).

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 $\leq 40\%$, 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 evident within the gene-sharing network. The α-, β-, and γ-proteobacteria form a nearly discrete module at the 25% amino acid identity threshold (T_{25}), with α-proteobacteria representing a discrete module at T_{50}, the network of which comprises a smaller number of more highly conserved proteins. Some γ-proteobacteria form a module with all β-proteobacteria at T_{55}, but the two modules do not correspond to the rRNA-based taxonomic framework. By contrast, some of the δ- and ε-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 δ-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 α-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. S5

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

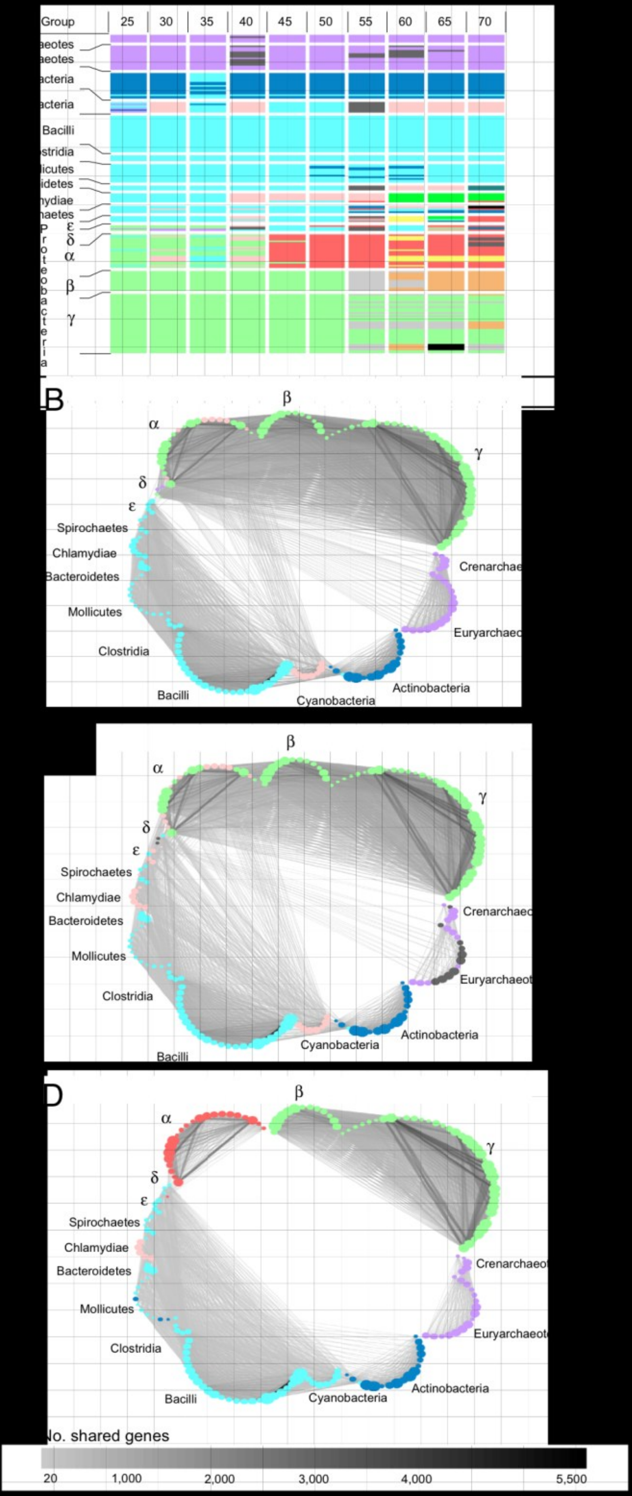


Fig. 1. Modules in networks of shared genes. (*A*) Modules detected (see *Materials and Methods*) are shown as colored boxes within columns for thresholds from T_{25} to T_{70}. Currently recognized higher-level taxonomic groups are indicated in rows for comparison. For example, for the network at T_{25} 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

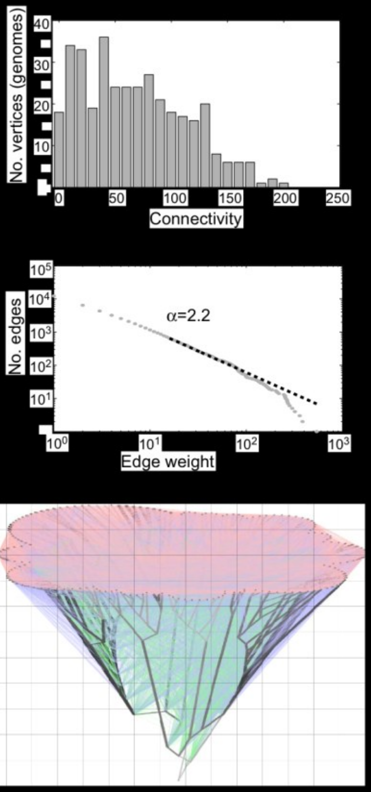


Fig. 2. Properties of the minimal LGT network. Properties are shown for a randomly selected replicate. The coefficient of variation (CV) for the whole data were ≪1% (Fig. S6

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

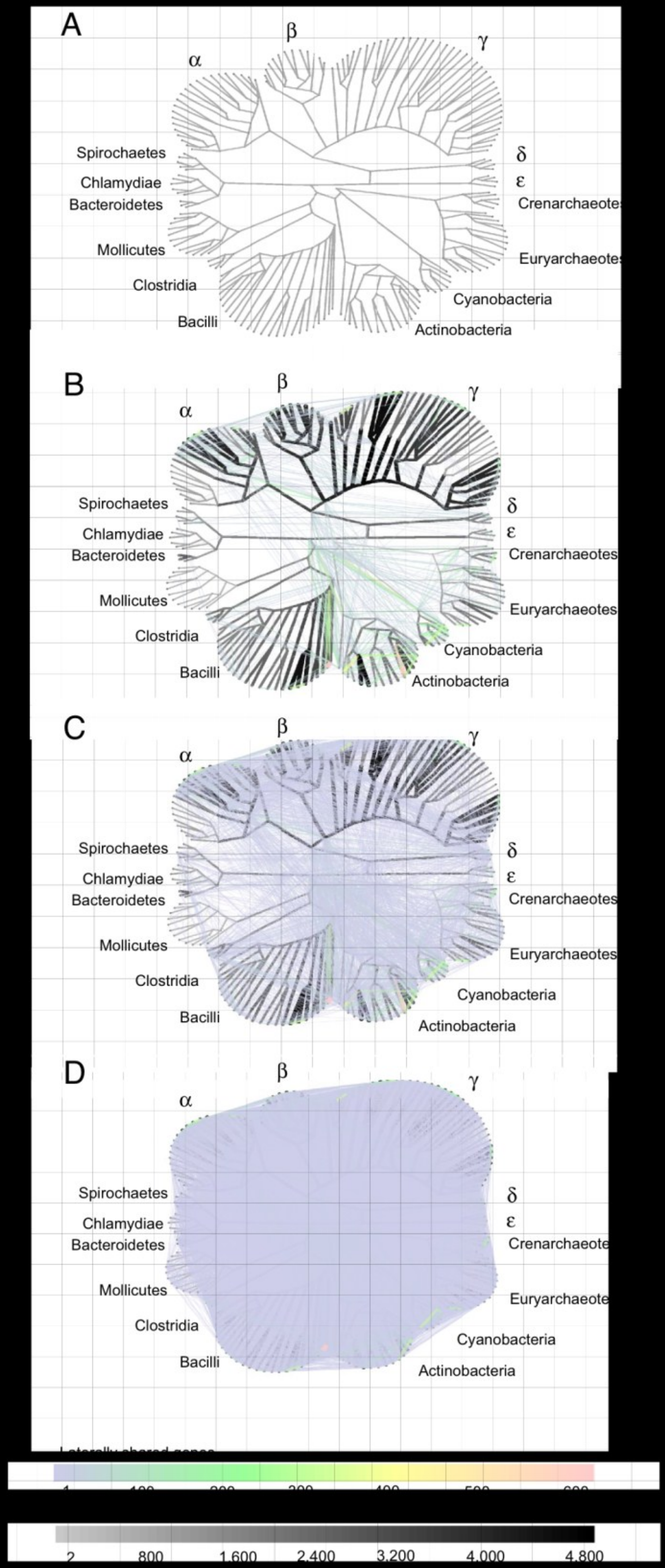


Fig. 3. A minimal LGT network for 181 genomes. (*A*) The reference tree used to ascribe vertical inheritance for inference of the MLN (see *Materials and Methods*). (*B*) The network showing only the 823 edges of weight ≥20 genes. Vertical edges are indicated in gray, with both the width and the shading 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.

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 ≪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

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 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.

8.2 A minimal lateral network of proteobacterial genomes.

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

(Own contribution: conceived and designed the experiment, performed part of the analysis, analyzed the data, and wrote the paper).

Networks of Gene Sharing among 329 Proteobacterial Genomes Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths

Abstract

Lateral gene transfer (LGT) is an important mechanism of natural variation among prokaryotes. Over the full course of evolution, most or all of the genes resident in a given prokaryotic genome have been affected by LGT, yet the frequency of LGT can vary greatly across genes and across prokaryotic groups. The proteobacteria are among the most diverse of prokaryotic taxa. The prevalence of LGT in their genome evolution calls for the application of network-based methods instead of tree-based methods to investigate the relationships among these species. Here, we report networks that capture both vertical and horizontal components of evolutionary history among 1,207,272 proteins distributed across 329 sequenced proteobacterial genomes. The network of shared proteins reveals modularity structure that does not correspond to current classification schemes. On the basis of shared protein-coding genes, the five classes of proteobacteria fall into two main modules, one including the alpha-, delta-, and epsilonproteobacteria and the other including beta- and gammaproteobacteria. The first module is stable over different protein identity thresholds. The second shows more plasticity with regard to the sequence conservation of proteins sampled, with the gammaproteobacteria showing the most chameleon-like evolutionary characteristics within the present sample. Using a minimal lateral network approach, we compared LGT rates at different phylogenetic depths. In general, gene evolution by LGT within proteobacteria is very common. At least one LGT event was inferred to have occurred in at least 75% of the protein families. The average LGT rate at the species and class depth is about one LGT event per protein family, the rate doubling at the phylum level to an average of two LGT events per protein family. Hence, our results indicate that the rate of gene acquisition per protein family is similar at the level of species (by recombination) and at the level of classes (by LGT). The frequency of LGT per genome strongly depends on the species lifestyle, with endosymbionts showing far lower LGT frequencies than free-living species. Moreover, the nature of the transferred genes suggests that gene transfer in proteobacteria is frequently mediated by conjugation.

Key words: horizontal gene transfer, microbial evolution, symbionts.

Introduction

Thomas and Nielsen 2005 Lang and Beatty 2007

Doolittle and Bapteste 2007

McInerney and Pisani 2007 Sorek et al. 2007

Bapteste et al. 2009

Babic et al. 2008

Marraffini and Sontheimer 2008 Horvath and Barrangou 2010

Snel et al. 2002 Beiko et al. 2005 Kunin et al. 2005

Mirkin et al. 2003

Roettger et al. 2009

Ochman et al. 2000 Nakamura et al. 2004

Babic et al. 2008

Dagan et al. 2008

Moran and Wernegreen 2000 Podar et al. 2008

Nakamura et al. 2004 Cordero and Hogeweg 2009

Lawrence and Ochman 1998

Zhaxybayeva et al. 2006 Shi and Falkowski 2008

Dufresne et al. 2008 Shi and Falkowski 2008

Stackebrandt et al. 1988

Brinkhoff et al. 2008

Davidov and Jurkevitch 2009

Kersters et al. 2006

Stackebrandt et al. 1988

Gupta 2006

Markowitz et al. 2010

Mongodin et al. 2005

Kunin et al. 2005

Beiko et al. 2005 Kunin et al. 2005 Fukami-Kobayashi et al. 2007 Dagan et al. 2008 Halary et al. 2010

Lerat et al. 2005 Gupta 2006 Wu et al. 2004 Ettema and Andersson 2009

Nakabachi et al. 2006 Kaneko et al. 2002

Materials and Methods

Data

http://www.ncbi.nlm.nih.gov/

Tatusov et al. 2000

Thompson et al. 1994

Enright et al. 2002

Dagan et al. 2008

Reconstruction of Gene Trees

Thompson et al. 1994

Guindon and Gascuel 2003
Jones et al. 1992

Network of Shared Protein Families

Newman 2006 Dagan et al. 2008

Reconstruction of a Reference Tree

Dagan et al. 2008

Pei et al. 2010

Thompson et al. 1994

Guindon and Gascuel 2003
Hasegawa et al. 1985

Guindon and Gascuel 2003

Felsenstein 1983

Felsenstein 2004

Reconstruction of a Minimal Lateral Network

Dagan and Martin (2007)

Dagan and Martin 2007

Dagan and Martin 2007

Identification of Recently Acquired Genes by Aberrant Nucleotide Pattern

Garcia-Vallve et al. 2000

Nakamura et al. 2004

Benjamini and Hochberg 1995

Results and Discussion

supplementary table S1 Supplementary Material

Poptsova and Gogarten 2010

Fischer and Eisenberg 1999

The Distribution of Shared Proteins among Proteobacteria

Graur and Li 2000

Novichkov et al. 2004 Dagan et al. 2010

table 1 supplementary table S1 Supplementary Material

Table 1.

supplementary table

S2 Supplementary Material

table 1

fig. 1

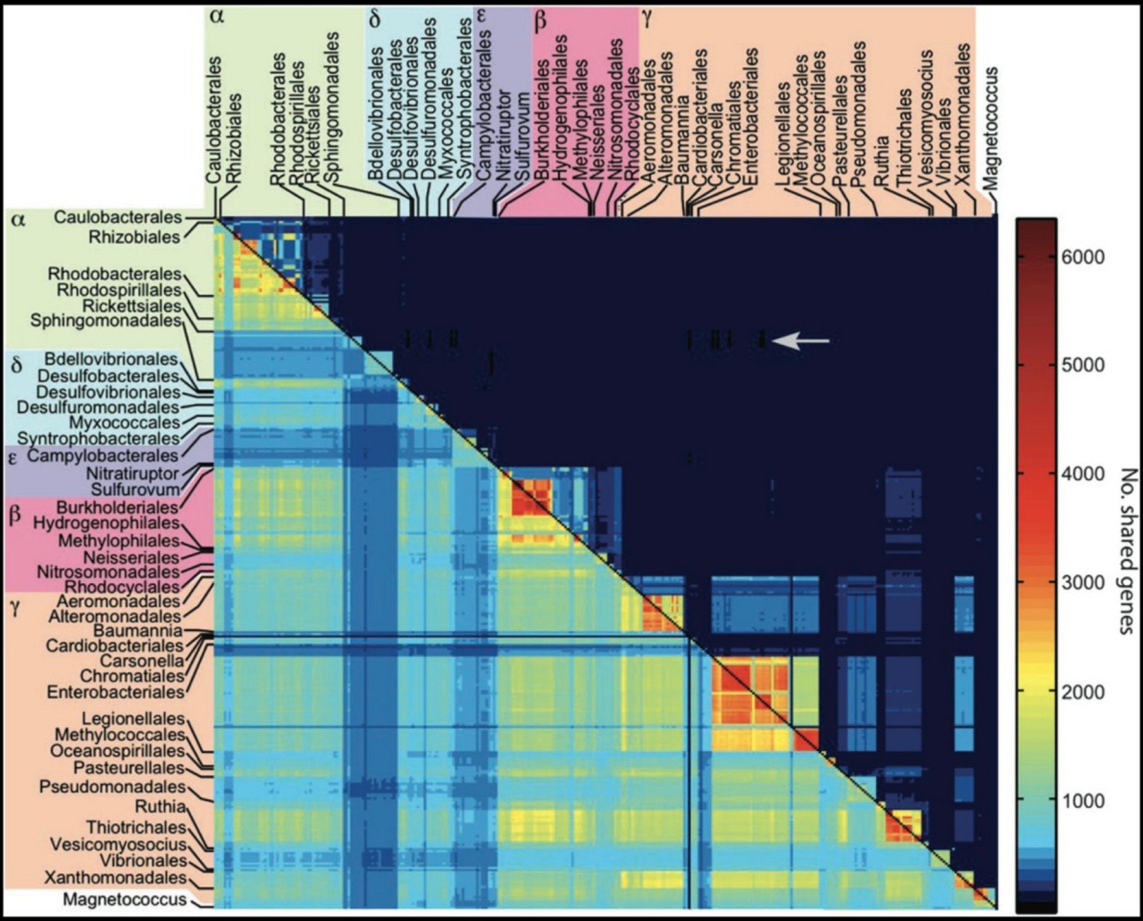


Fig. 1.

fig. 1

supplementary table S3 Supplementary Material

Krylov et al. 2003

Sukdeo and Honek 2008

Halary et al. 2010

supplementary fig. S1 Supplementary Material

supplementary fig. S2 Supplementary Material

Sorek et al. 2007

Chan et al. 2009

Modules within the NSP

Newman 2006 Dagan et al. 2008

fig. 2

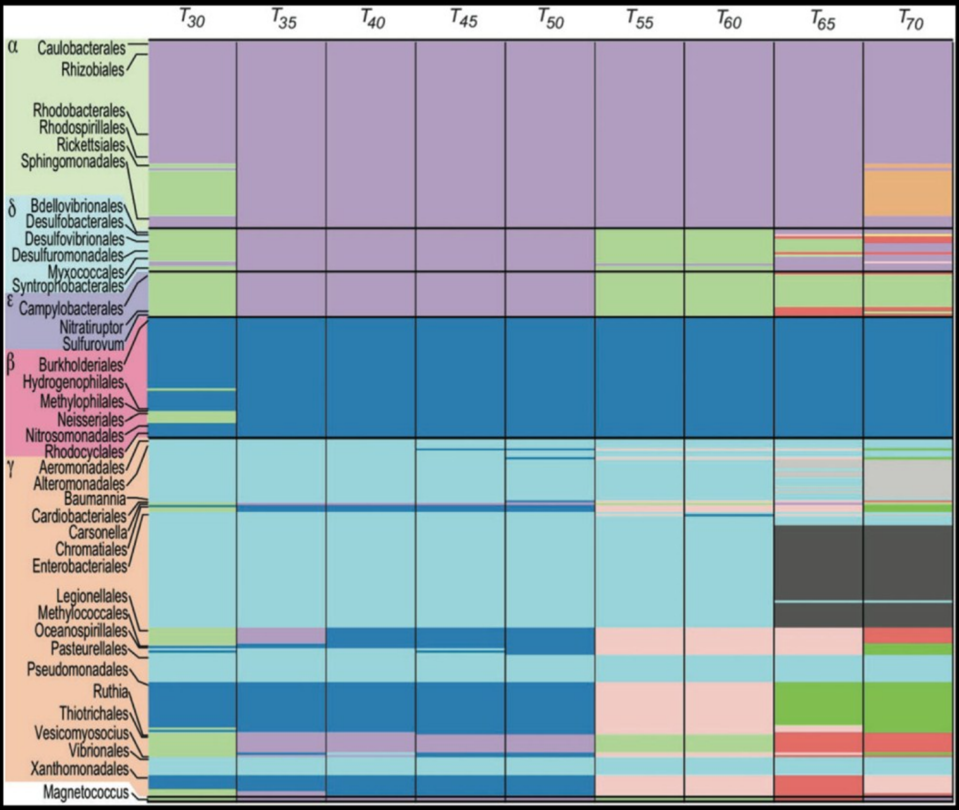


Fig. 2.

supplementary table S4 Supplementary Material

supplementary table S4 Supplementary Material

Doolittle and Bapteste 2007 Galtier and Daubin 2008 Bapteste et al. 2009

Doolittle 1999

Ciccarelli et al. 2006 Galtier and Daubin 2008

Dagan and Martin 2006

fig. 2

fig. 2

Ciccarelli et al. (2006)

fig. 2

Scott et al. 2006

Oyston 2008

Gao et al. 2009

Zar 1999

Pal et al. 2006 Moran 2007

figs. 1 2

Chaffron et al. 2010

Jain et al. 2003

Minimal Lateral Networks

Dagan et al. 2008

Doolittle et al. 2003

Dagan and Martin 2007

Dagan et al. 2008

Zar 1999 fig. 3

table 2A

table 2A

table 2A

Zar 1999 supplementary fig. S3 Supplementary Material

supplementary fig. S3 Supplementary Material

Kunin et al. 2005

Dagan and Martin 2007

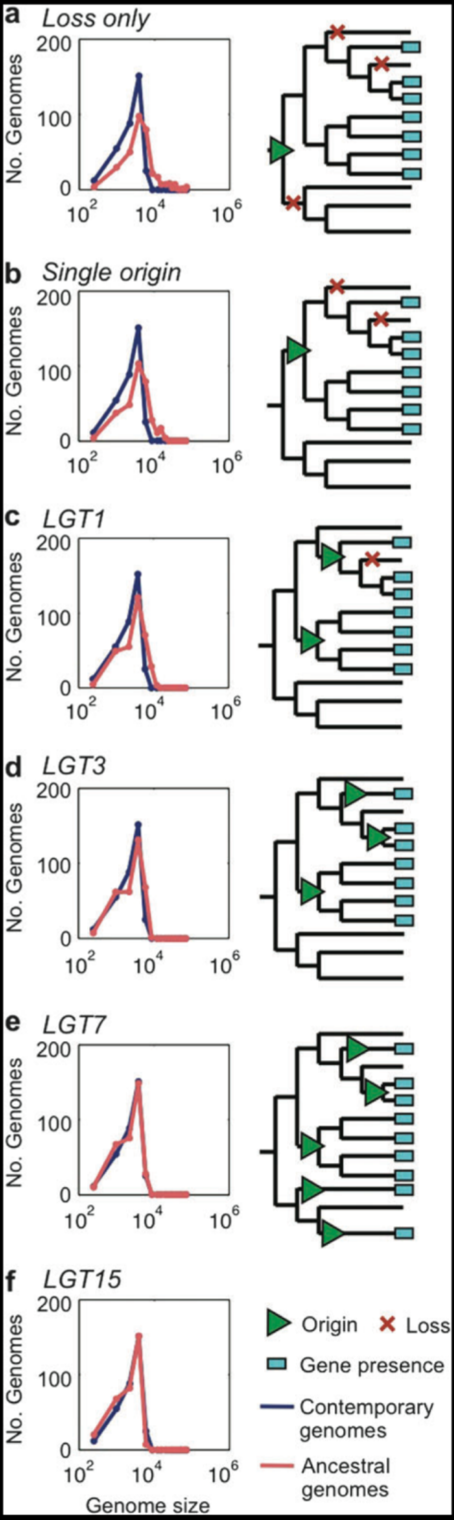


Fig. 3.

supplementary fig. S3 Supplementary Material

LGT Inference against a Gene Content Reference Tree

Dagan and Martin 2007

Bapteste et al. 2009

Snel et al. 1999

Felsenstein 1983

Table 2.

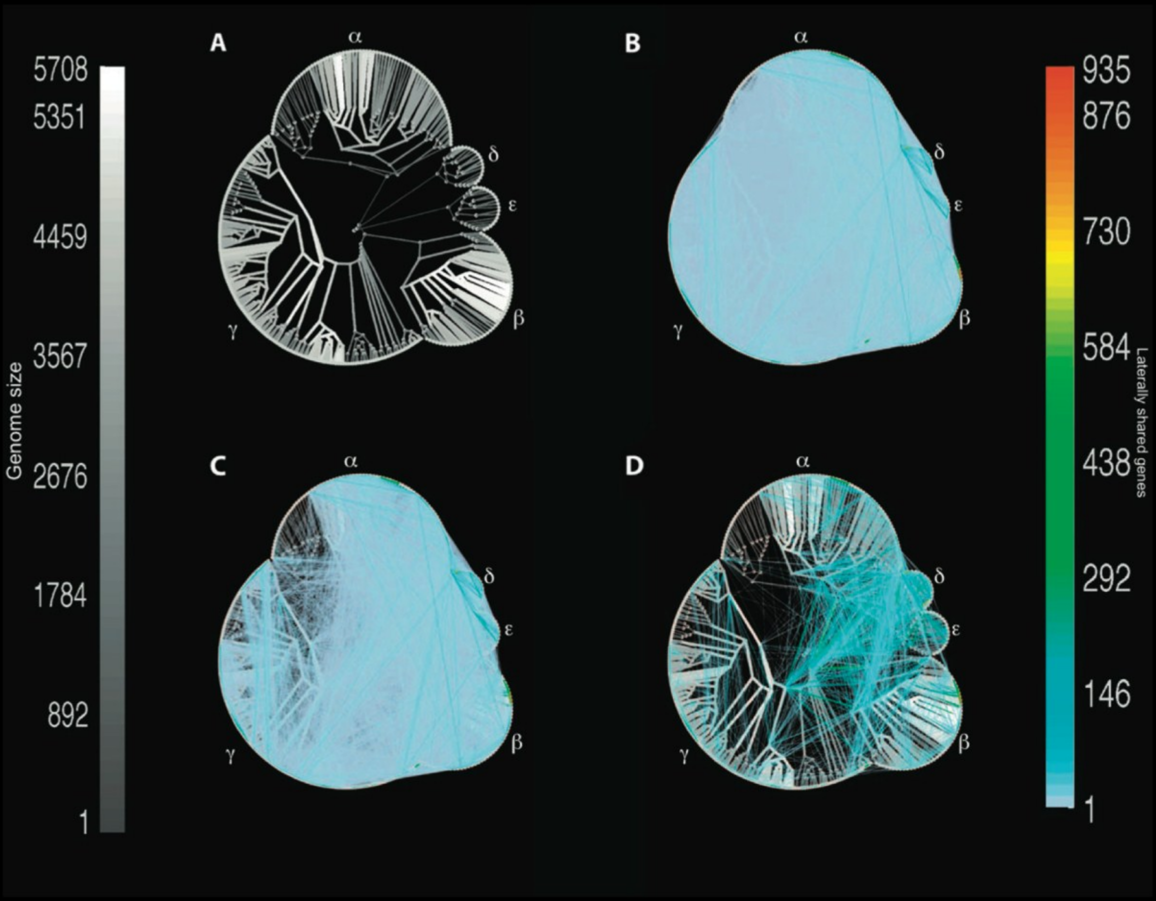


Fig. 4.

supplementary fig. S4 Supplementary Material

supplementary figs. S5 S6 Supplementary Material

table 2B

supplementary fig. S5 Supplementary Material

table 2B

Table 3.

MLN Properties

fig. 4

table 3

fig. 5A

Zar 1999

table 3

fig. 5B

fig. 6A

fig. 5C

fig. 6B

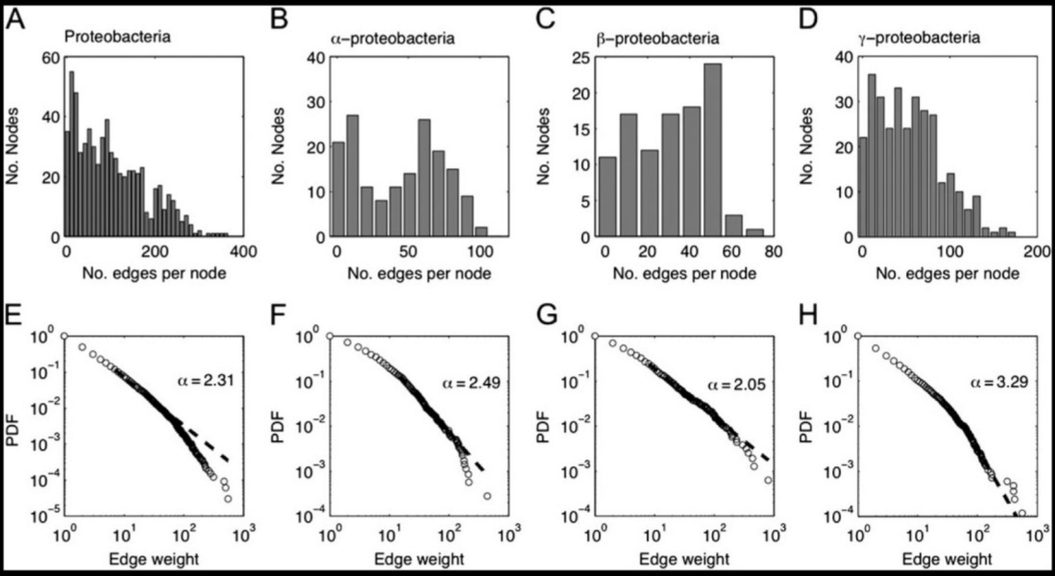


Fig. 5.

table 3

fig. 1

fig. 5*E*

fig. 6*B*

fig. 5*D*

table 3 fig. 5*F*–*H*

table 2

supplementary fig. S7*A*–*C* Supplementary Material

fig. 6*C*

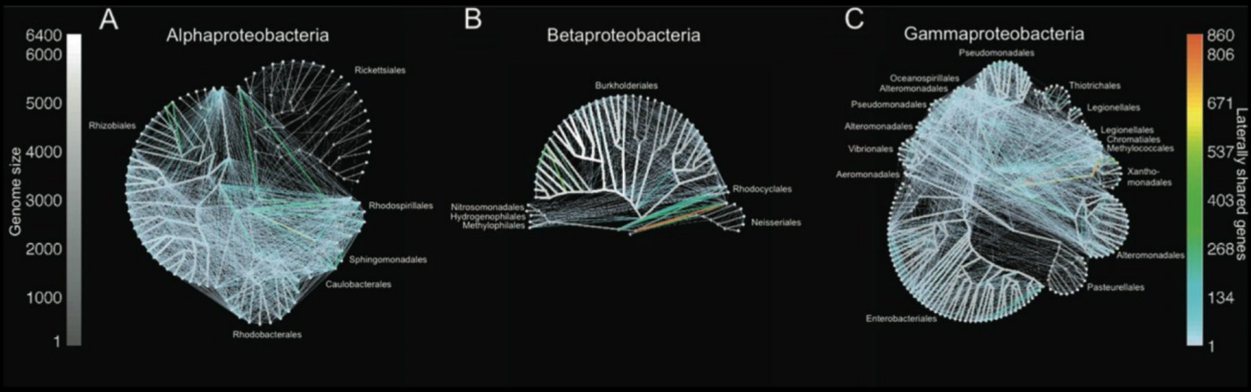


Fig. 6.

Table 4.

supplementary fig. S7D–F Supplementary Material

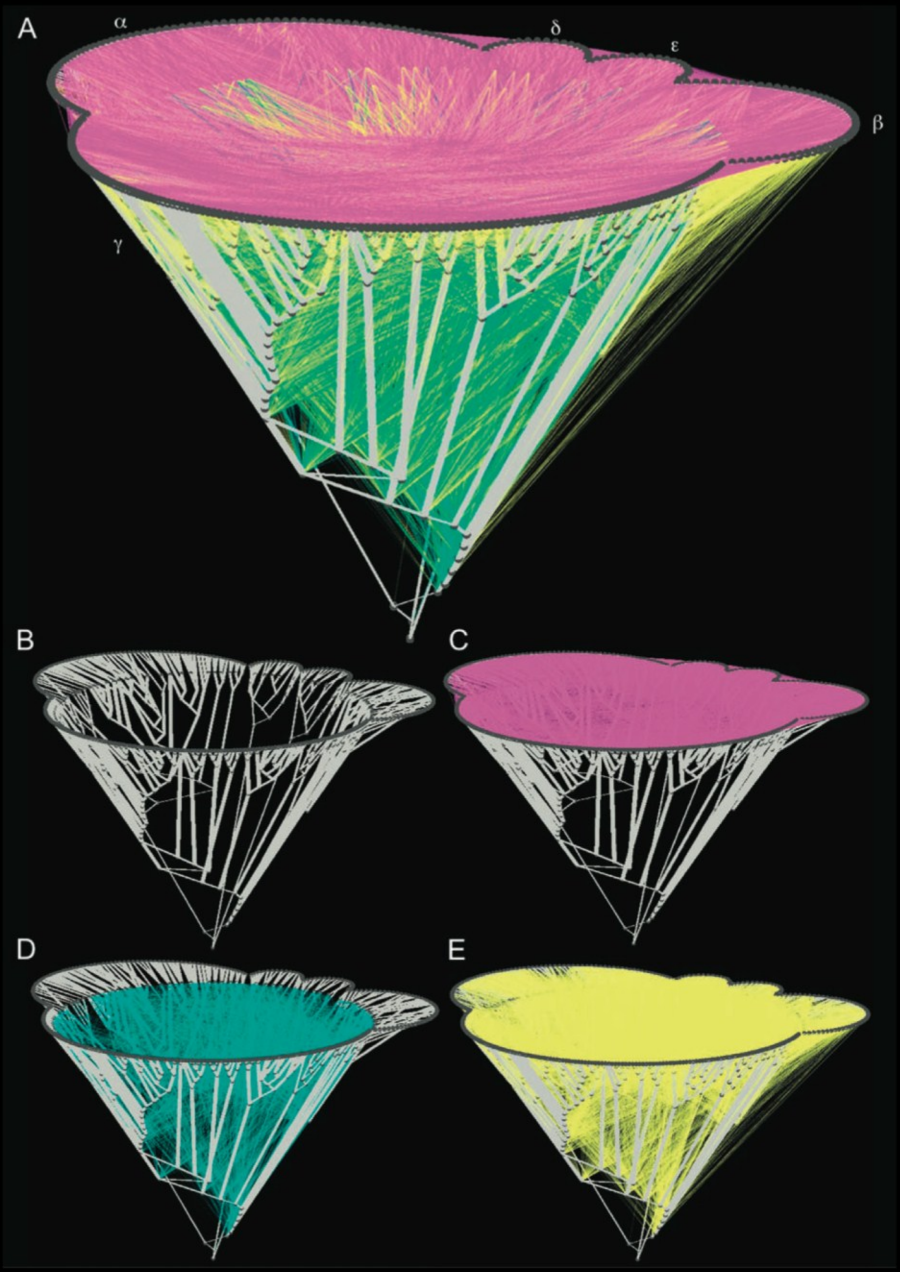


Fig. 7. figure 3

table 3

Andam et al. 2010

Dagan et al. 2008

Lawrence and Ochman 1998 Nakamura et al. 2004

table 5

table 5

fig. 7

Conclusions

Supplementary Material

Supplementary tables S1 S5 figures S1 S8 http://www.mbe.oxfordjournals.org/

Acknowledgments

www.molevol.de/resources

References

8.3 Phylogenomic networks root the tree of life

Dagan T. Roettger M, Bryant D, Martin W: **Genome networks root the tree of life between prokaryotic domains**. *Genome Biol Evol* 2010, **2**:379-392.

(Own contribution: conceived and designed the experiment, performed part of the analysis, analyzed the data, and wrote the paper).

Genome Networks Root the Tree of Life between Prokaryotic Domains

Abstract

Eukaryotes arose from prokaryotes, hence the root in the tree of life resides among the prokaryotic domains. The position of the root is still debated, although pinpointing it would aid our understanding of the early evolution of life. Because prokaryote evolution was long viewed as a tree-like process of lineage bifurcations, efforts to identify the most ancient microbial lineage split have traditionally focused on positioning a root on a phylogenetic tree constructed from one or several genes. Such studies have delivered widely conflicting results on the position of the root, this being mainly due to methodological problems inherent to deep gene phylogeny and the workings of lateral gene transfer among prokaryotes over evolutionary time. Here, we report the position of the root determined with whole genome data using network-based procedures that take into account both gene presence or absence and the level of sequence similarity among all individual gene families that are shared across genomes. On the basis of 562,321 protein-coding gene families distributed across 191 genomes, we find that the deepest divide in the prokaryotic world is interdomain, that is, separating the archaebacteria from the eubacteria. This result resonates with some older views but conflicts with the results of most studies over the last decade that have addressed the issue. In particular, several studies have suggested that the molecular distinctness of archaebacteria is not evidence for their antiquity relative to eubacteria but instead stems from some kind of inherently elevated rate of archaebacterial sequence change. Here, we specifically test for such a rate elevation across all prokaryotic lineages through the analysis of all possible quartets among eight genes duplicated in all prokaryotes, hence the last common ancestor thereof. The results show that neither the archaebacteria as a group nor the eubacteria as a group harbor evidence for elevated evolutionary rates in the sampled genes, either in the recent evolutionary past or in their common ancestor. The interdomain prokaryotic position of the root is thus not attributable to lineage-specific rate variation.

Key words: phylogenies, early evolution, tree of life, microbial genomics, lateral gene transfer.

Introduction

Ueno et al.
2006

Nisbet 2000 Rasmussen 2000 Shen et al. 2001
Brasier et al. 2006 Grassineau et al. 2006

Javaux et al. 2001 Knoll et al. 2006
Butterfield 2000

Fischer
2008 Rasmussen et al. 2008

Rivera and Lake 2004 Embley and Martin 2006 Pisani et al. 2007 Cox et al. 2008 Koonin 2009

Gogarten et al. 1989 Iwabe et al. 1989 Brown and Doolittle 1995

Doolittle 1999 McInerney and Pisani 2007

Lake et al. 2009

Skophammer et al. 2006 Lake et al. 2007

Lake et al. 2008

Di Giulio 2007

Bapteste and Philippe 2002

Bapteste et al. 2009

Gupta 1998 Gupta and Lorenzini 2007

Cavalier-Smith 2006b

Cavalier-Smith 2010b

Cavalier-Smith 2010a

Cavalier-Smith 2006a

Canfield 2006 Ueno et al. 2006

Ventura et al. 2007

Sleep et al. 2004

Butterfield 2000 Javaux et al. 2001

Wong et al. (2007)

Ueno et al. 2006 Zhaxybayeva et al. (2005)

Boussau et al. (2008)

Dagan and Martin 2007 McInerney et al. 2008 Battistuzzi and Hedges 2009 Koonin 2009

Gogarten et al. 1989 Lake et al. 2009

Snel et al. 1999 Graham et al. 2000

Martin and König 1996 Claus et al. 2005 Engelhardt 2007

Chong et al. 2000 Frols et al. 2009

Dimarco et al. 1990 Deppenmeier 2002 Fujihashi et al. 2007

Bell and Jackson 2001

Lake et al. 2008 Cavalier-Smith 2010b

Farris 1972

Materials and Methods

Orthologous Protein Families

http://www.ncbi.nlm.nih.gov/

Tatusov et al. 1997

Thompson et al. 1994

Enright et al. 2002

Splits Network

Huson and Bryant 2006

Midpoint Rooting in Splits Network

Farris 1972

fig. 1*a*

fig. 1*b*

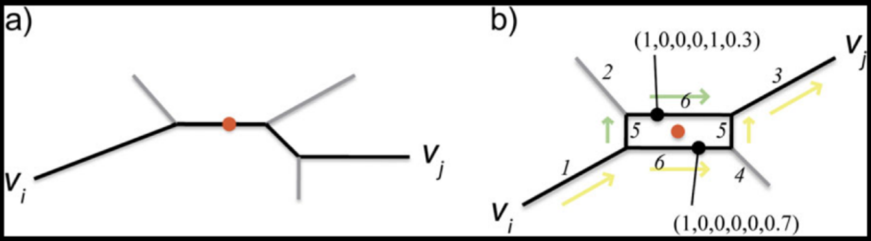


Fig. 1.

fig. 2*a*

Dress and Huson 2004

Thompson et al.

1994

Landan and Graur 2007

fig. 2*b*

Shimodaira and Hasegawa 1999

Felsenstein 1996

Yang 2007

fig. 2*c*

Yoder and Yang 2000

Test of the Global Clock Assumption

Yang

1998

fig. 2*c*

http://www.ncbi.nlm.nih.gov/taxonomy

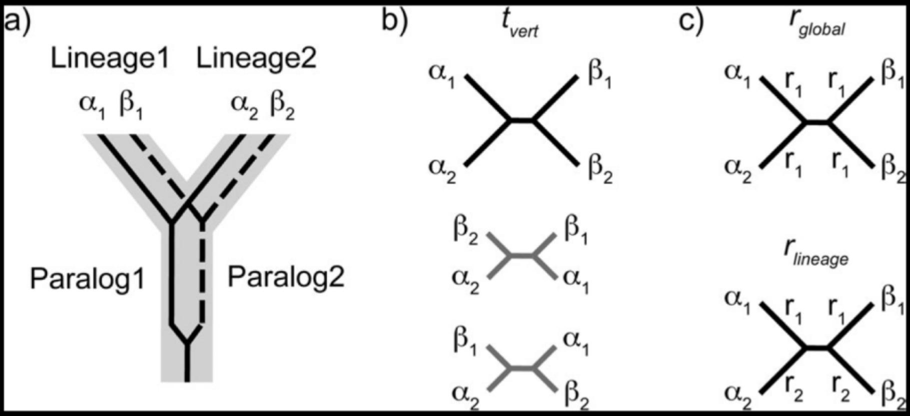


Fig. 2.

Results and Discussion

Splits Networks for Prokaryotic Genomes

Enright

et al. 2002

Bryant and

Moulton 2004

Huson and Bryant 2006

Saitou and Nei 1987

supplementary table S1

fig. 3

fig. 3a

fig. 3b

fig. 3c

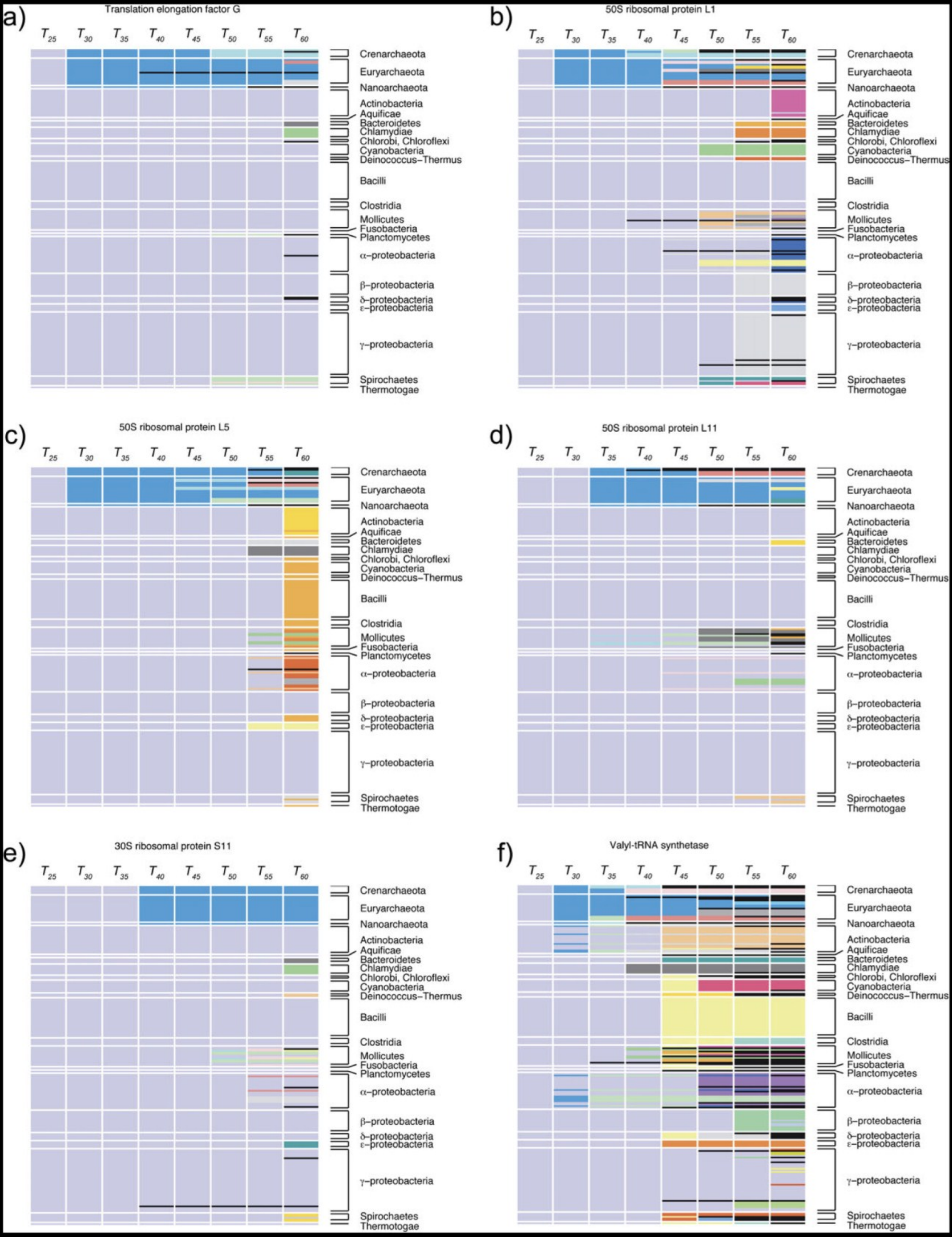

a)
Translation elongation factor G
b)
50S ribosomal protein L1
c)
50S ribosomal protein L5
d)
50S ribosomal protein L11
e)
30S ribosomal protein S11
f)
Valyl-tRNA synthetase
T_{25}
T_{30}
T_{35}
T_{40}
T_{45}
T_{50}
T_{55}
T_{60}
Crenarchaeota
Euryarchaeota
Nanoarchaeota
Actinobacteria
Aquificae
Bacteroidetes
Chlamydiae
Chlorobi, Chloroflexi
Cyanobacteria
Deinococcus–Thermus
Bacilli
Clostridia
Mollicutes
Fusobacteria
Planctomycetes
α–proteobacteria
β–proteobacteria
δ–proteobacteria
ε–proteobacteria
γ–proteobacteria
Spirochaetes
Thermotogae

Fig. 3.

fig. 3*d*

fig. 3*e*

fig. 3*f*

Raoult et al. 2003

Prokaryotic Genome Clusters

fig. 4

supplementary table S1

supplementary fig. S1

fig. 5

figure 5

fig. 6

Lang and Beatty 2007 Dagan et al. 2008

Pierce et al. 2008 Ljungdahl 2009

Thauer et al. 2008

Martin et al. 2008

Müller 2003

Thauer et al. 2008 Biegel et al. 2009

Zhaxybayeva et al. 2009

Chistoserdova et al. 1998

Deppenmeier et al. 2002

Ng et al. 2000

Doolittle and Bapteste 2007

fig. 4

Fukami-Kobayashi et al. 2007 Cox et al. 2008 Puigbò et al. 2009

supplementary table S2

The Root of Prokaryotes

Farris 1972

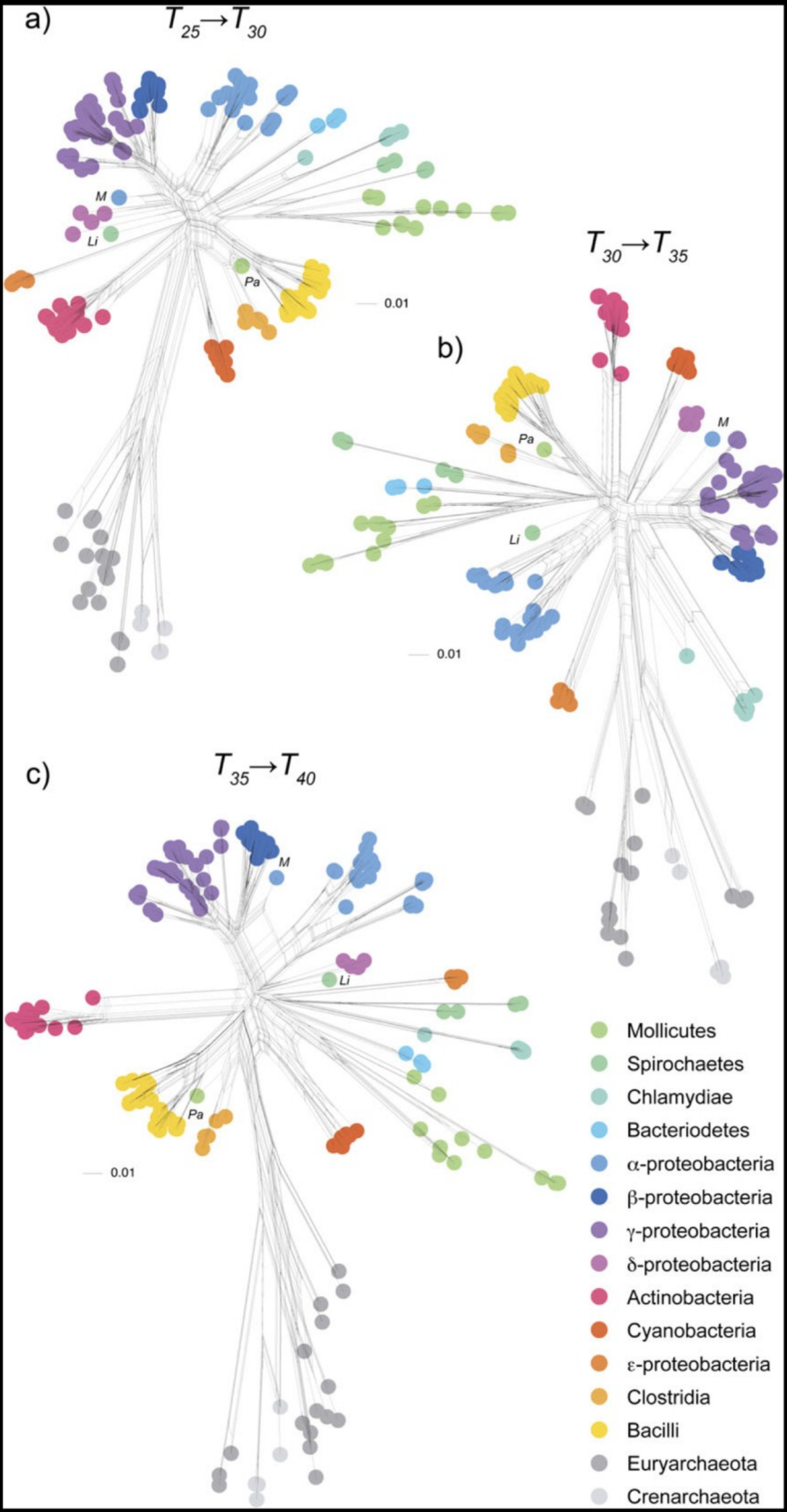


FIG. 4.

supplementary figure S1

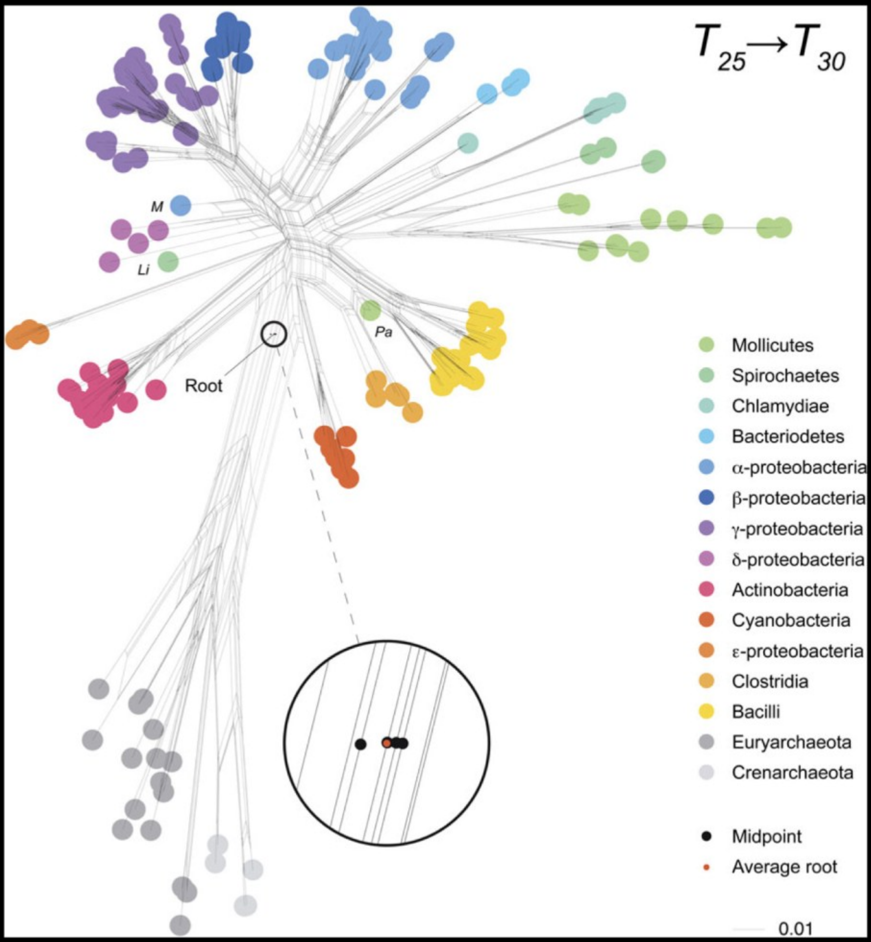


Fig. 5.

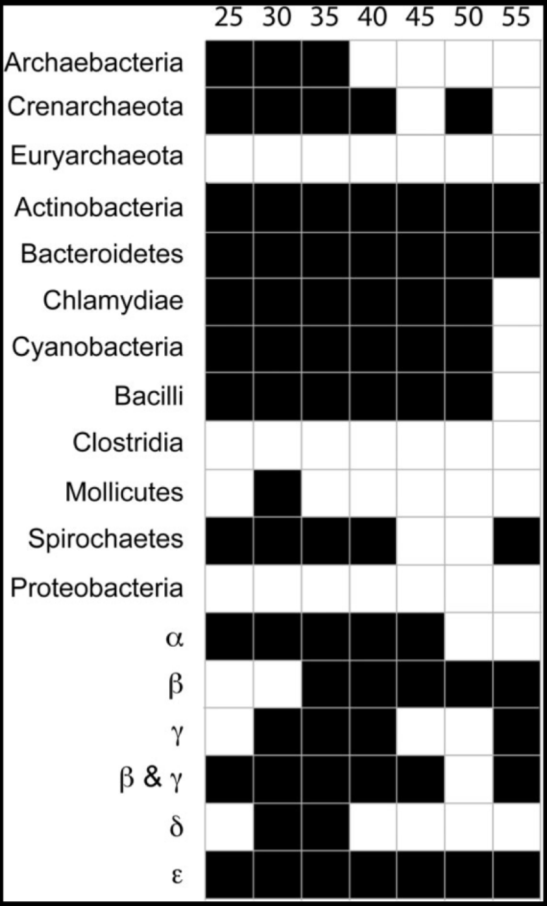
fig. 5

fig. 4 supplementary

fig. S1

supplementary

table S3



	25	30	35	40	45	50	55
Archaebacteria	■	■	■	□	□	□	□
Crenarchaeota	■	■	■	■	□	■	□
Euryarchaeota	□	□	□	□	□	□	□
Actinobacteria	■	■	■	■	■	■	■
Bacteroidetes	■	■	■	■	■	■	■
Chlamydiae	■	■	■	■	■	■	□
Cyanobacteria	■	■	■	■	■	■	□
Bacilli	■	■	■	■	■	■	□
Clostridia	□	□	□	□	□	□	□
Mollicutes	□	■	□	□	□	□	□
Spirochaetes	■	■	■	■	□	□	■
Proteobacteria	□	□	□	□	□	□	□
α	■	■	■	■	■	□	□
β	□	□	■	■	■	■	■
γ	□	■	■	■	□	□	■
β & γ	■	■	■	■	■	□	■
δ	□	■	■	□	□	□	□
ε	■	■	■	■	■	■	■

Fig. 6.

Comparison of Evolutionary Rates among Lineages

fig. 2a

Kollman and Doolittle 2000

supplementary table S4

Zhaxybayeva et al. 2005

fig. 2b

Yang 2007

fig. 2c

Landan and Graur 2007

supplementary table S5

Novichkov et al. 2004

Cavalier-Smith 2006a 2009 2010a 2010b de Duve 2007

Cavalier-Smith 2010b

supplementary table S4

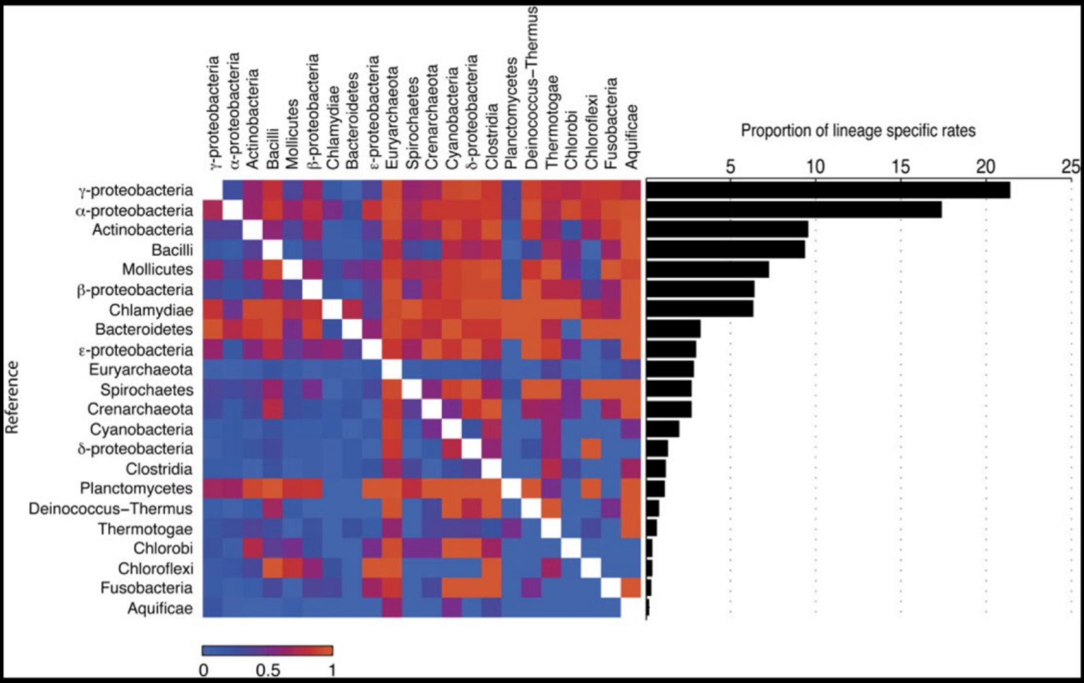


Fig. 7.

supple- mentary table S4

fig. 7

fig. 3

fig. 7

Cavalier-Smith 2006a de Duve 2007

Stetter 2006 Thauer 2007

Life at the Root

Lane et al. 2010

Martin and Russell 2003 Koonin and Martin 2005 Branciamore et al. 2009

Stetter et al. 1990 Pace (1991)

Stetter et al. 1990 Pace 1991

Forterre 1995 1996

Cox et al. 2008 van der Giezen 2009

Gaucher et al. 2003 2008 Boussau et al. (2008)

Amend and McCollom (2009)

Amend and McCollom 2009

Forterre 1996

Constanzo et al. (2009)

Constanzo et al. 2009

Say and Fuchs 2010

Conclusions

Cavalier-Smith 2006b

Lake et al. 2009

Gogarten et al. 1989 Iwabe et al. 1989

Koonin 2009

Pisani et al. 2007

Supplementary Material

Supplementary figure S1 tables S1 S5 http://www.oxfordjournals.org/our_journals/gbe/

Acknowledgments

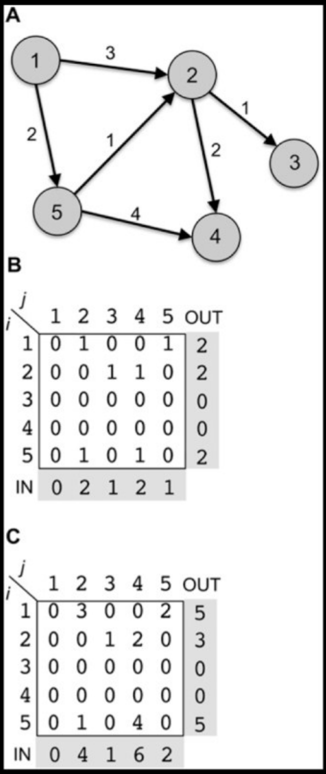
Literature Cited

8.4 Directed networks reveal barriers and bypasses to LGT in prokaryotes

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T: **Directed networks reveal barriers and bypasses to lateral gene transfer in prokaryotes**. *Genome Res* 2011, 21:599-609.

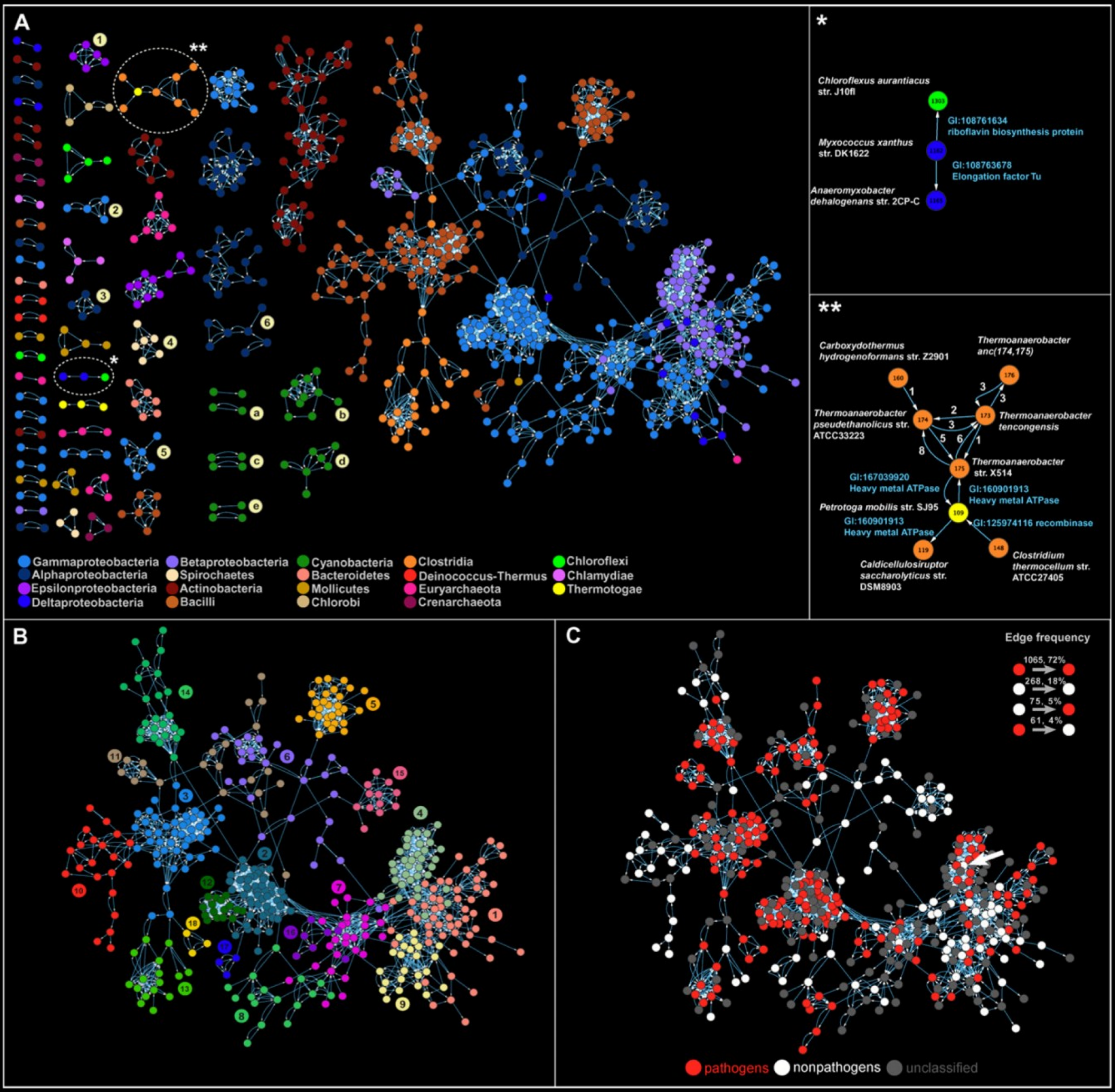
(Own contribution: conceived and designed the experiment, performed part of the analysis, analyzed the data, and wrote the paper).

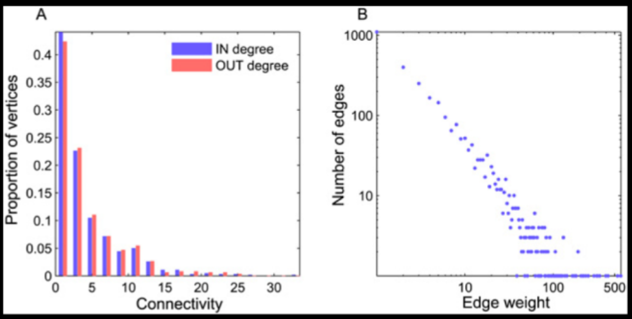
[redacted]



i \ *j*	1	2	3	4	5	OUT
1	0	1	0	0	1	2
2	0	0	1	1	0	2
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	1	0	1	0	2
IN	0	2	1	2	1	

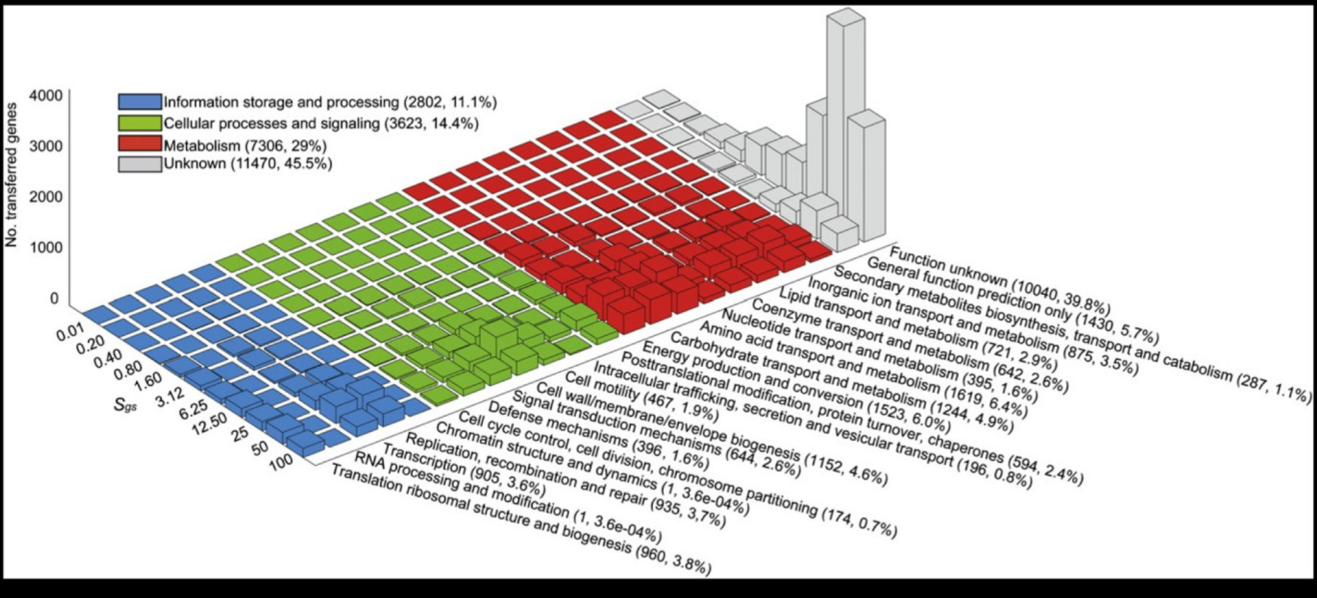
i \ *j*	1	2	3	4	5	OUT
1	0	3	0	0	2	5
2	0	0	1	2	0	3
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	1	0	4	0	5
IN	0	4	1	6	2	


A
1
2
3
4
5
6
a
b
c
d
e
*
**
Gammaproteobacteria
Alphaproteobacteria
Epsilonproteobacteria
Deltaproteobacteria
Betaproteobacteria
Spirochaetes
Actinobacteria
Bacilli
Cyanobacteria
Bacteroidetes
Mollicutes
Chlorobi
Clostridia
Deinococcus-Thermus
Euryarchaeota
Crenarchaeota
Chloroflexi
Chlamydiae
Thermotogae
*
Chloroflexus aurantiacus str. J10fl
1303
GI:108761634 riboflavin biosynthesis protein
Myxococcus xanthus str. DK1622
GI:108763678 Elongation factor Tu
Anaeromyxobacter dehalogenans str. 2CP-C
**
Carboxydothermus hydrogenoformans str. Z2901
Thermoanaerobacter anc(174,175)
160
176
1
3
3
Thermoanaerobacter pseudethanolicus str. ATCC33223
174
2
3
173
Thermoanaerobacter tencongensis
5
6
1
8
175
Thermoanaerobacter str. X514
GI:167039920 Heavy metal ATPase
GI:160901913 Heavy metal ATPase
Petrotoga mobilis str. SJ95
109
GI:160901913 Heavy metal ATPase
GI:125974116 recombinase
119
148
Caldicellulosiruptor saccharolyticus str. DSM8903
Clostridium thermocellum str. ATCC27405
B
14
11
5
6
3
15
4
2
12
7
10
1
18
16
17
13
8
9
C
Edge frequency
1065, 72%
268, 18%
75, 5%
61, 4%
pathogens
nonpathogens
unclassified


A
B
IN degree
OUT degree
Proportion of vertices
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30
Connectivity
Number of edges
1000
100
10
10
100
500
Edge weight

[redacted]

[redacted]


No. transferred genes
4000
3000
2000
1000
0
Information storage and processing (2802, 11.1%)
Cellular processes and signaling (3623, 14.4%)
Metabolism (7306, 29%)
Unknown (11470, 45.5%)
S_{gs}
0.01
0.20
0.40
0.80
1.60
3.12
6.25
12.50
25
50
100
Function unknown (10040, 39.8%)
General function prediction only (1430, 5.7%)
Secondary metabolites biosynthesis, transport and catabolism (287, 1.1%)
Inorganic ion transport and metabolism (875, 3.5%)
Lipid transport and metabolism (721, 2.9%)
Coenzyme transport and metabolism (642, 2.6%)
Nucleotide transport and metabolism (395, 1.6%)
Amino acid transport and metabolism (1619, 6.4%)
Carbohydrate transport and metabolism (1244, 4.9%)
Energy production and conversion (1523, 6.0%)
Posttranslational modification, protein turnover, chaperones (594, 2.4%)
Intracellular trafficking, secretion and vesicular transport (196, 0.8%)
Cell motility (467, 1.9%)
Cell wall/membrane/envelope biogenesis (1152, 4.6%)
Signal transduction mechanisms (644, 2.6%)
Defense mechanisms (396, 1.6%)
Cell cycle control, cell division, chromosome partitioning (174, 0.7%)
Chromatin structure and dynamics (1, 3.6e-04%)
Replication, recombination and repair (935, 3,7%)
Transcription (905, 3.6%)
RNA processing and modification (1, 3.6e-04%)
Translation ribosomal structure and biogenesis (960, 3.8%)

[redacted]

[redacted]

[redacted]

[redacted]

8.5 Minimal lateral network of languages

Nelson-Sathi S, List JM, Geisler H, Fangerau H, Gray RD, Martin W, Dagan T: **Networks reveal abundant hidden borrowing in the evolution of Indo-European languages**. *Proc Roy Soc Lond B* 2011*, in press.*

Featured in Research Highlights at *Nature* 2010 **468**: 735.

(Own contribution: designed the experiment, performed part of the analysis, analyzed the data, and wrote the paper).

PROCEEDINGS OF THE ROYAL SOCIETY B

FirstCite® e-publishing

Proc. R. Soc. B
doi:10.1098/rspb.2010.1917
Published online

Networks uncover hidden lexical borrowing in Indo-European language evolution

Shijulal Nelson-Sathi[1], Johann-Mattis List[2], Hans Geisler[2], Heiner Fangerau[3], Russell D. Gray[4], William Martin[1] and Tal Dagan[1,*]

[1]*Institute of Botany III, and* [2]*Faculty of Philosophy, Heinrich-Heine University Düsseldorf, Germany*
[3]*Institute of the History, Philosophy and Ethics of Medicine, Ulm University, Germany*
[4]*Department of Psychology, University of Auckland, Auckland 1142, New Zealand*

Language evolution is traditionally described in terms of family trees with ancestral languages splitting into descendent languages. However, it has long been recognized that language evolution also entails horizontal components, most commonly through lexical borrowing. For example, the English language was heavily influenced by Old Norse and Old French; eight per cent of its basic vocabulary is borrowed. Borrowing is a distinctly non-tree-like process—akin to horizontal gene transfer in genome evolution—that cannot be recovered by phylogenetic trees. Here, we infer the frequency of hidden borrowing among 2346 cognates (etymologically related words) of basic vocabulary distributed across 84 Indo-European languages. The dataset includes 124 (5%) known borrowings. Applying the uniformitarian principle to inventory dynamics in past and present basic vocabularies, we find that 1373 (61%) of the cognates have been affected by borrowing during their history. Our approach correctly identified 117 (94%) known borrowings. Reconstructed phylogenetic networks that capture both vertical and horizontal components of evolutionary history reveal that, on average, eight per cent of the words of basic vocabulary in each Indo-European language were involved in borrowing during evolution. Basic vocabulary is often assumed to be relatively resistant to borrowing. Our results indicate that the impact of borrowing is far more widespread than previously thought.

Keywords: community structure; lateral transfer; phylogenetics

1. INTRODUCTION

Genome evolution and language evolution have a lot in common. Both processes entail evolving elements—genes or words—that are inherited from ancestors to their descendants. The parallels between biological and linguistic evolution were evident both to Charles Darwin, who briefly addressed the topic of language evolution in *The origin of species* [1], and to the linguist August Schleicher, who in an open letter to Ernst Haeckel discussed the similarities between language classification and species evolution [2]. Computational methods that are currently used to reconstruct genome phylogenies can also be used to reconstruct evolutionary trees of languages [3,4]. However, approaches to language phylogeny that are based on bifurcating trees recover vertical inheritance only [3,5–7], neglecting the horizontal component of language evolution (borrowing). Horizontal interactions during language evolution can range from the exchange of just a few words to deep interference [8]. In previous investigations, which focused only on the component of language evolution that is described by a bifurcating tree [3,5–7], the extent of borrowing might therefore have been overlooked.

Lexical borrowing is the transfer of a word from a donor language to a recipient language as a result of a certain kind of contact between the speakers of the two languages [9]. This is one of the most common types of interaction between languages. Lexical borrowing can be reciprocal or unidirectional, and occurs at variable rates during evolution. Factors affecting the rate of lexical borrowing during evolution include the intensity of contact between the speakers of the respective languages, the genetic or typological closeness of the languages (which facilitates the inclusion of foreign words), the amount of bi- or multi-lingual speakers in the respective linguistic communities, or a combination thereof [10,11]. For example, English has been heavily influenced throughout its history by different languages such as Old Norse and Old French [12], it has been estimated that 8 per cent of its basic vocabulary is borrowed from those languages [13]. Icelandic, on the other hand, has preserved most of its original words [14].

A key part of inferences in historical linguistics is the identification of cognate sets. These are sets of words from different languages that are etymologically related. The words in a cognate set are derived from a single common ancestral form that was present in an ancestral language. Cognate judgement is an arduous enterprise since it includes the complete evolutionary reconstruction of all words in the sampled languages for a certain concept. Historical linguists usually make use of an in-depth analysis of structural resemblances between the

[*] Author for correspondence (tal.dagan@uni-duesseldorf.de).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2010.1917 or via http://rspb.royalsocietypublishing.org.

Received 6 September 2010
Accepted 3 November 2010

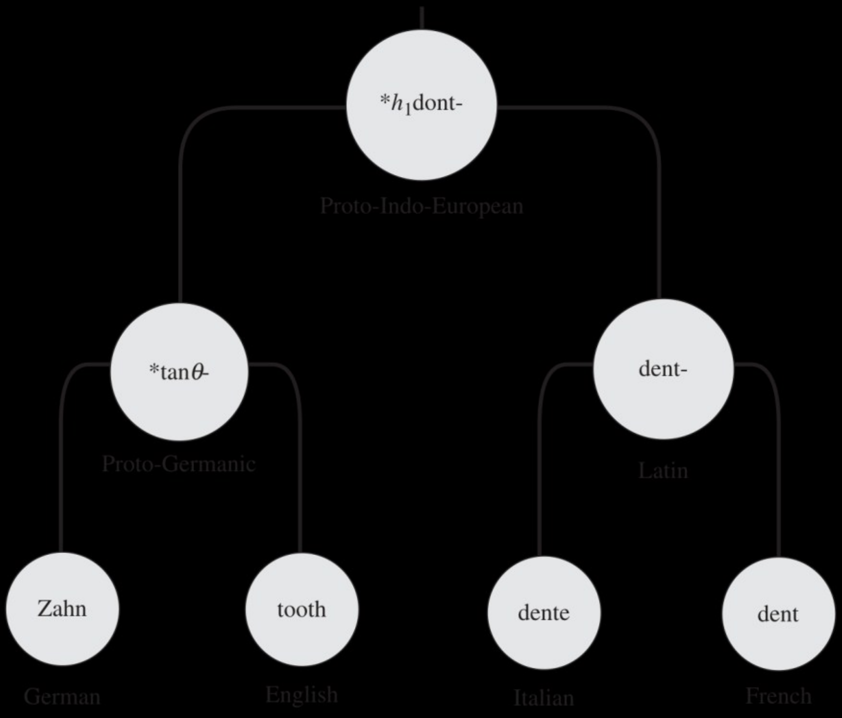


Figure 1. Etymological reconstruction of the concept tooth. The English and German word forms have descended from the Proto-Germanic ancestor [52]. The Italian and French words are descendants of Latin, and the Proto-Germanic and Latin forms stem from Proto-Indo-European [43,53].

word forms, looking for sound correspondences in specific environments. The identification of a cognate is thus much more than just a hunt for resemblant forms or 'lookalikes'. Only a set of words that have regular sound correspondences provide good evidence for genealogical relatedness and thus only these words can be grouped into a single cognate set (COG). For example, the concept 'tooth' has a cognate set that unites English *tooth*, German *Zahn*, Italian *dente* and French *dent* as etymologically related (figure 1). However, similar word forms can arise not only by inheritance, but also by lexical borrowing. Unfortunately, the further we go back in time, the more difficult it becomes to distinguish inheritance from transfer, and reconstructed COGs may include hidden borrowing events that are erroneously coded as vertical inheritance.

Lexical borrowing is a non-tree-like evolutionary event that cannot be reconstructed using phylogenetic trees that are common in evolutionary biology [15,16]. Linguists have long been aware of the problems that borrowing introduces. At about the same time that Darwin suggested the tree metaphor for the evolution of species in 1859 [1], August Schleicher introduced the family tree to linguistics [17]. Few years later, his model was rejected by several scholars arguing against the use of a simple tree model to describe the evolution of languages, which they noted to be reticulated by nature [18,19]. Other non-tree-like models were proposed by linguists to study language evolution—including waves [18,20] and networks [21]—but they lacked either quantitative parameters, historical dimensions or both. At the other extreme, quantitative estimates for language divergence lacked an explicit model to explain language relatedness [22,23]. Apart from some sporadic attempts to visualize language evolution of specific words by a combination of a bifurcating family tree with the non-tree-like component superimposed on it [24], linguists have, for lack of better alternatives, largely stuck to the tree model, while emphasizing its inadequacies.

Phylogenetic methods that were developed to take into account horizontal transfer of genes during microbial evolution offer an alternative model for the horizontal aspects of language evolution. Recent years have witnessed several applications of reticulated trees and split networks to language evolution [25–28], yet none of these have either specifically uncovered borrowing events or delivered an estimate for the borrowing frequency during language evolution. Here, we apply phylogenetic networks to recover the frequency of hidden borrowings during the evolution of Indo-European languages using the criterion of word inventory dynamics over time, proposing a general model for language evolution that includes both vertical and horizontal components of word transfer during evolution.

2. METHODS

(a) *Data*

Here, we used two publicly available cognate datasets: Dyen [29] and Tower of Babel (ToB) [30]. For the analysis, all COGs in both datasets are converted into a binary presence/absence pattern (PAP). A PAP within the Dyen dataset includes 84 digits; if a cognate set includes one or more words from language *i*, then digit x_i in its corresponding pattern is '1'; otherwise, it is '0'. The same conversion method is used for the ToB dataset where the PAPs include 73 digits.

(b) *Shared COGs network*

The number of shared COGs between each language pair is calculated as the number of cognate sets in which both languages are present. A division of the network into modules is based on maximizing a modularity function defined as the number of edges within a community minus the expected number of edges [31]. 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 modules, additional subdivisions are made, each time comparing the contribution of the new subdivision with the general modularity score of the entire network. This process is carried out until there are no additional subdivisions that will increase the modularity of the network as a whole [31].

(c) *Reference trees*

Language trees were inferred by a Bayesian approach using MrBayes [32] as detailed by Gray & Atkinson [3]. In addition, neighbour-joining (NJ) trees [33] were reconstructed from Hamming distances using SplitsTree [34]. A reference tree with English internal to the Germanic clade was produced manually from the Bayesian tree. A randomized reference tree for the Dyen dataset was produced by randomizing the language names in the Bayesian reference tree. Trees are available in Newick format at http://www.molevol.de/resources.

(d) *Borrowing models and the minimal lateral network*

In the loss-only (LO) model, all COGs are assumed to have originated at the root of the reference tree. The loss events for each COG are estimated by using a binary recursive PERL algorithm that scans the reference tree and infers the minimum number of losses [35]. When a COG is absent in a whole clade, a single loss event is inferred in the common ancestor of that clade. In the single-origin (SO) model, each cognate is assumed to have originated at its first occurrence on the reference tree. A binary recursive algorithm scans the reference tree from root to tips to identify the first ancestral node that is the common ancestor of all cognate 'present' cases.

In the BOR1 model, each cognate is allowed to have two word origins, where one is a borrowing. A preliminary origin is inferred as in the SO model, followed by researching for a cognate origin in each of the two clades branching from the preliminary origin node. If the hypothetical taxonomic unit that was inferred as the preliminary origin has no cognate 'absent' descendants, the cognate is inferred to have an SO. Once the nodes of the two origins are set, losses are inferred as in the LO model.

We tested additional models allowing four, eight and 16 origins, where one is an origin, and the rest are borrowings. These are implemented in the same way as in the BOR1 model, except that the origin search is iterated. For example, a search for origins under the BOR3 model entails (i) a search for a preliminary origin (as in the SO model), (ii) a search for the next origin in descendants (as in the BOR1 model) and, (iii) for each next origin, another search. If an origin has no cognate-absent descendants, the number of origins inferred is smaller than the maximum allowed. Ancestral vocabulary size at a certain internal node is inferred as the total COG origins that were inferred to occur at that node. The distributions of ancestral and modern vocabulary sizes were compared by using the Wilcoxon non-parametric test [36].

The minimal lateral network (MLN) [37] is calculated for each dataset by the allowance model that was statistically accepted by the test described above. The MLN comprises the reference tree, with additional information of the vocabulary size in all internal nodes. Lateral cognate sharing among internal and external nodes is summarized in a 167 × 167 matrix that includes all tree nodes, where $a_{ij} = a_{ji}$ = number of laterally shared COGs between nodes i and j. The MLN is then depicted by an in-house script using Matlab.

3. RESULTS AND DISCUSSION

(a) *Community structure in the network of shared cognate sets*

For the study of evolution by borrowing, we analysed two independent, publicly available collections of cognate sets from Indo-European languages. Both datasets comprise words from individual languages or dialects corresponding to concepts that are included in Swadesh lists [38]. Basic concepts are expressed by simple words rather than compounds or phrases and contain names for body parts, pronouns, common verbs and numerals, but exclude technological words and words related to specific ecologies or habitats. Words expressing basic concepts are supposed to exist in all languages and thus may serve as a *tertium comparationis* for language comparison [39]. Moreover, basic concepts are rarely replaced by other words, either through external (lexical borrowing) or internal factors (semantic shift) [13,16].

The Dyen dataset [29] includes word forms for 84 languages (including Greek, Armenian, Celtic, Romance, Germanic, Slavic, Albanian and Indo-Iranian languages) corresponding to 200 basic vocabulary concepts [39] sorted into 2346 COGs [3]. While obvious borrowings were excluded in the original Dyen dataset [29], we used an edited version where 124 marked borrowings are coded into their respective COGs [25]. Detailed reinspection of Romance cognates revealed an additional six hidden borrowings [40] (electronic supplementary material, table S1).

The second dataset is based on etymological dictionaries and Swadesh lists published by the ToB project [30]. It is based on word forms for 110 basic vocabulary items for a total of 98 languages from which we extracted 73 contemporary ones, including languages from the Celtic, Romance, Germanic, Slavic, Albanian and Indo-Iranian branches of Indo-European, sorted into 722 COGs. Detectable borrowings were excluded in the original database; however, a recent detailed screening revealed five undetected borrowings within Romance languages [40].

A network analysis of the distribution of cognate word forms across Indo-European languages should provide new insights into the frequency and distribution of borrowing in Indo-European language history. Networks are mathematical structures used to model pairwise relations between entities. The entities are called vertices and they are linked by edges that represent the connections or interactions between the vertices. A network of N vertices can be fully defined by the matrix $A = [a_{ij}]_{N^* N}$, with $a_{ij} = a_{ji} \neq 0$ if a link exists between nodes i and j, and $a_{ij} = a_{ji} = 0$ otherwise. In the study of Indo-European languages, each language is represented by a vertex, i, whereas the elements of the matrix, A,

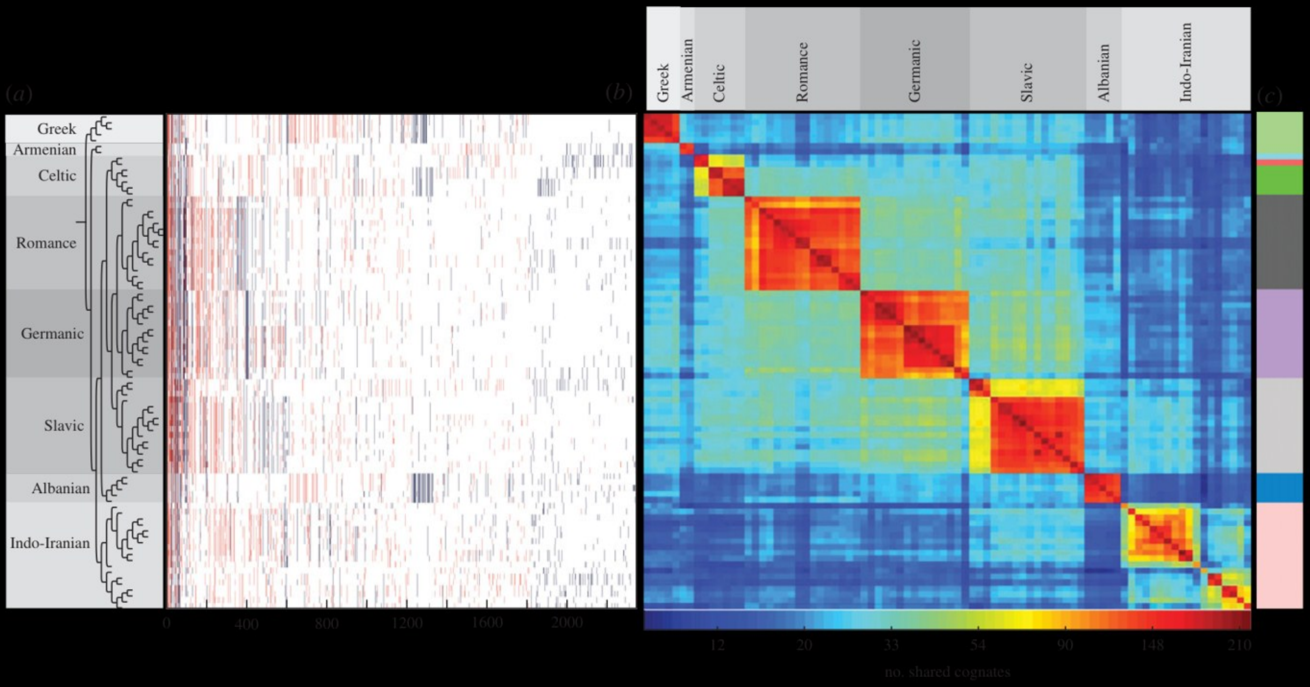


Figure 2. Modules in the shared COGs network. (*a*) A graphic representation of cognate PAPs. Languages are sorted by their order on the reference phylogenetic tree [3]. COGs are sorted by their size in ascending order. A presence case of a certain COG in a certain language is coloured in blue if the COG pattern is congruent with the tree branching patterns and red otherwise. (*b*) A matrix representation of the shared COGs network in Indo-European languages. Cells in the matrix are edges in the network. Edges are colour-coded by the frequency of shared cognate according to the colour bar at the bottom. The languages in the matrix are sorted by order of appearance in the phylogenetic tree on the left. (*c*) Modules within the shared COGs network. Languages included in the same module are coloured in the same colour.

correspond to the number of shared cognate sets between language pairs, a_{ji}. Cognate sharing can result either from vertical inheritance or from borrowing.

For network reconstruction, cognate sets were converted into a binary format of PAPs for each COG in each language [3]. For the 2346 COGs in the Dyen dataset [29], 1169 different PAPs were observed, of which 942 (80%) are unique and 227 are recurring (figure 2*a*). Closely related languages typically share the most frequent PAPs. For example, Panjabi and Lahnda, two Indian languages, share 78 cognates that are unique to both languages. The ToB dataset includes 532 different PAPs, none of which are unique (electronic supplementary material, figure S1). The frequency of shared COGs among languages in the main branches uncovers components of both inheritance and borrowing.

The binary PAPs of the Dyen COGs are readily assorted into an 84×84 matrix representation of the cognate-sharing network that consists of vertices (languages) connected by edges (shared cognates), the edge weights are the number of shared cognates per vertex pair. There are 3486 edges in the network, all vertices of which are connected, thereby forming a 'clique' in network terms (figure 2*b*). Some groups of languages are more strongly interconnected among themselves than with others in the cognate-sharing network, thereby forming communities.

We examined the community structure in the network by division into modules [31,41]. Modules correspond to 'natural' groups within a network, that is, groups of vertices that are more highly connected to each other than they are to other vertex sets. With only two exceptions, the nine modules calculated within the cognate-sharing network correspond exactly to the main branches of Indo-European languages. One exception concerns the Armenian dialects Adapazar (*Armenian List* in Dyen dataset [42]) and eastern modern Armenian (*Armenian Mod* in Dyen dataset [42]), which are grouped with the Greek languages into one module. This is because Armenian shares significantly ($p \ll 0.01$, using the Wilcoxon test) more cognates with the Greek languages (30 ± 2, $n = 5$) than with the other languages (22 ± 3, $n = 79$). This module has been independently recognized by linguists [43]. The other exception is the split of both Irish dialects from Celtic (figure 2*c*). The same network-based analysis of the ToB dataset yields only four modules: (i) Slavic and Albanian; (ii) Armenian, Greek, Celtic, Germanic and Romance; (iii) Indo; and (iv) Iranian (electronic supplementary material, figure S2).

Language communities that do not correspond to monophyletic clades in the tree are the result of patchy COG distributions that could not be reconciled with the phylogenetic tree. For example, Romani, which branches with Indo-Iranian languages, shares 25 COGs with Modern Greek, such as the COGs for 'flower' (Modern Greek: λουλούδι (*louloudi*); Romani: *lulugi*) and 'because' (Modern Greek: επειδή (*epeide*); Romani: *epidhi*). Since the Romani dialect in the Dyen dataset [29] is a variety spoken in Greece [42], these are probably borrowed from Greek to Romani.

(b) *Borrowing frequency during Indo-European language evolution*

In the Dyen dataset, there are 1391 (59%) patchily distributed PAPs that are incongruent with the tree branching pattern (figure 2*a*). In principle, such patchy COG distributions could arise solely through independent parallel evolution, through vertical inheritance from the common ancestor of all languages and differential loss of lexica during language evolution, or via lexical borrowing among languages. The first possibility seems sufficiently unlikely as to exclude *a priori*. There is no clear estimation for the frequency of parallel evolution during language evolution, but we can assume that it is rather rare and cannot, therefore, be used to explain the distribution pattern of all patchy COGs. If we invoke the second scenario to explain all COGs of patchy distribution, then the result is a common ancestral language that includes each and every COG existing in contemporary languages. In order to entertain such a claim, one would have to assume that the proto-language employed many different, but redundant, words for the same basic concepts, far more than every known contemporary language. This runs contrary to uniformitarianism, a key principle in historical sciences such as geology, biology and linguistics, which states that processes in the past should not be assumed to differ fundamentally from those observed today [44,45]. Hence, if ancient and modern languages were of similar nature, then the number of words that were used to express fundamental concepts (basic vocabulary size) in ancestral languages should be similar to that used in contemporary languages. This principle can be used to infer the minimum amount of lexical borrowing in Indo-European languages that is required in order to bring the distribution of basic vocabulary size in ancestral languages into agreement with that of contemporary languages.

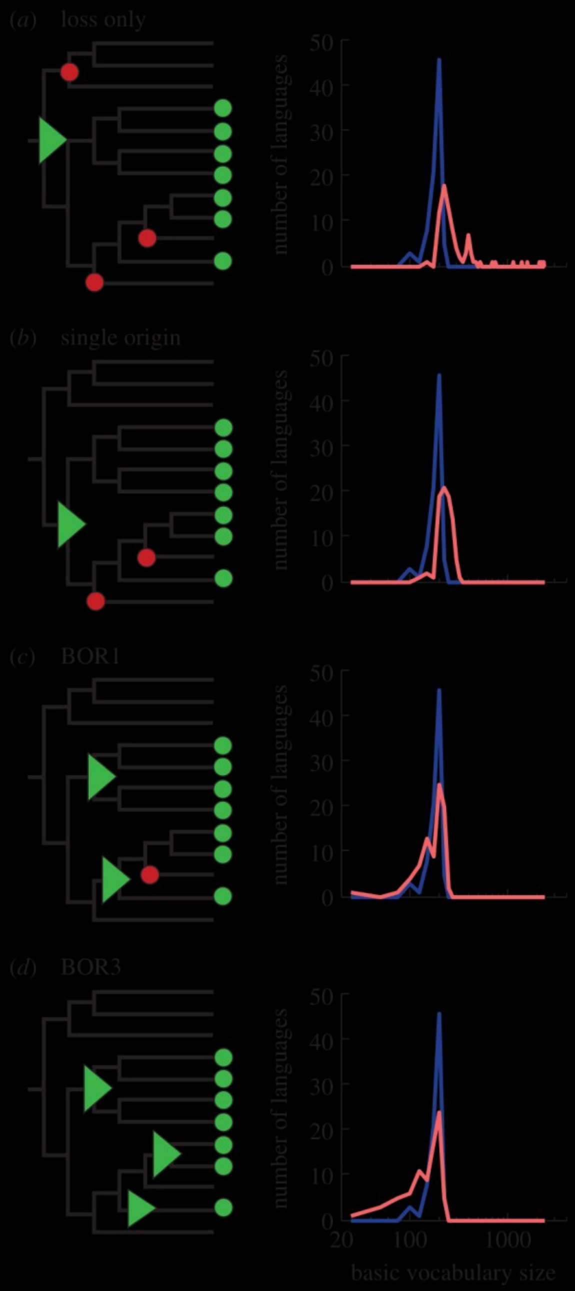


Figure 3. Inference of borrowing frequency by ancestral vocabulary size. (*a–d*) Schematic (left) and dynamics of ancestral and contemporary vocabulary size (right) under the different borrowing models. The fraction of interquartile range ($(\text{Median}_{\text{ancestral}} - \text{Median}_{\text{contemporary}})/\text{IQR}_{\text{contemporary}}$) in the different models is as follows. Loss only: 2.92; origin only: 1.93; BOR1: 0.12; BOR3: −0.86. Green triangles, origin; red circles, loss; green circles, word presence; blue line, contemporary languages; red line, ancestral languages.

This network method to address non-tree-like patterns of shared characters requires the use of a reference tree [37]. Here, we use a phylogenetic tree reconstructed by a Bayesian approach [3]. First, we designate an evolutionary scenario that uses vertical inheritance and LO (model), according to which current COG distribution is governed solely by loss. Each ancestral language contains all cognates present in its descendants, and vocabulary size hence becomes progressively larger back through time (figure 3*a*). Note that a loss event applies only to the sample of basic vocabulary and does not mean a loss from the language as a whole. With the Dyen dataset [29] and the reference tree, the common Indo-European ancestor would have had a vocabulary size of 2346 for basic words, expressing 200 basic concepts. This estimate is 11 times larger than the average basic vocabulary size in our sample ($p = 1.05 \times 10^{-24}$, using the Wilcoxon test). Such large vocabulary sizes are indeed unrealistic, but so is the assumption that new words do not arise during language evolution. In the SO model, we allow new words to arise over time, placing the word origin at the most parsimonious place that is the common ancestor of all COG-present cases (figure 3*b*). This model results in smaller ancestral vocabularies of up to 317 COGs, but these are still significantly larger than the contemporary vocabularies ($p = 1.65 \times 10^{-19}$, using the Wilcoxon test). The SO model entails an average of three losses per COG (electronic supplementary material, table S2).

Thus, we either have to embrace the untenable assumption that ancestral vocabulary sizes were fundamentally different in the past than they are today or, preferably, we have to allow some amount of borrowing during evolution. We start by allowing only one borrowing event per COG, the BOR1 model. This model allows each COG to have two origins in the reference tree, one of which is by borrowing from any source (figure 3*c*). The result of this model is reduced ancestral vocabularies during the early evolution of languages, and an overall ancestral vocabulary size distribution that is not significantly different from that of contemporary languages ($p = 0.61$, using the Wilcoxon test). Of the total Dyen COGs, 918 (39%) are monophyletic, hence their distribution is readily explained by an SO, while the remaining 1373 (61%) are patchy enough to infer two origins (one borrowing event). This frequency translates to an average rate of 0.6 borrowing events per COG during Indo-European language evolution.

If we allow up to three borrowings per COG (the BOR3 model; figure 3*d*), inferred ancestral vocabulary shrinks towards sizes that are again significantly different from modern ones, but this time are smaller than those of contemporary languages ($p = 4.43 \times 10^{-5}$, using the Wilcoxon test); that is, too much borrowing and not enough vertical descent are incurred from the standpoint of ancestral vocabulary sizes. Furthermore, under the BOR3 model, the average number of inferred word losses per COG is less than 1. But loss of COGs within basic vocabulary occurs quite frequently in language evolution [7], hence the BOR3 model is also unrealistic in that sense. Additional models allowing up to 15 borrowings per COG result in even smaller ancestral vocabulary sizes (electronic supplementary material, figure S3). Hence, ancestral basic vocabulary sizes demand borrowings to keep them realistically small, but too much borrowing makes them unrealistically small.

Testing the present evolutionary models with the help of a reference tree that is inferred from the same data might bias the inference of origin and loss events. However, using the Bayesian approach to reconstruct the tree yields the majority signal in the data. If the majority of COGs evolve mainly by vertical inheritance, then the tree is expected to be a reliable representation of the language phylogeny [46]. High frequency of borrowing events may mask the vertical signal and lead to less reliable reconstruction. To test the robustness of our borrowing frequency estimates, we repeated our analysis using various reference trees. Use of an alternative phylogenetic tree reconstructed by NJ [33] results in the same BOR1 model ($p = 0.7$, using the Wilcoxon test; electronic supplementary material, figure S3). In both reference trees, English is basal to the Germanic clade. However, this position is debated among linguists, and traditional classifications put English inside that clade [12,47]. To test the influence of the English position within the tree on our borrowing assessment, we tested all models using a reference tree with English in an internal position. Using that reference tree also yielded the BOR1 model ($p = 0.78$, using the Wilcoxon test), with all other models rejected ($\alpha = 0.05$). Using a random phylogenetic tree eliminates all patterns of vertically inherited COGs and accordingly results in the BOR15 model ($p = 0.16$, using the Wilcoxon test; electronic supplementary material, figure S4).

Performing the same tests on the ToB dataset yielded higher borrowing frequencies, with BOR3 being the only statistically accepted model ($p = 0.59$, using the Wilcoxon test; electronic supplementary material, figure S5). Inference by this model results in 155 COGs of SO, 181 COGs of two origins, 307 COGs of three origins and 79 COGs of four origins. Hence, in 567 (79%) of the 722 COGs, we detected one or more borrowing event. The average rate of borrowing events per COG during language evolution in the ToB dataset is 1.4 (electronic supplementary material, table S2). The higher borrowing rate inferred for the ToB dataset in comparison to the Dyen dataset might have to do with differences in their reconstruction. The cognate judgements in ToB are based on a deeper etymological reconstruction in comparison to the Dyen dataset. This results in more words that are distributed over fewer cognate sets, which leads to patchy COG distribution patterns that are frequently incongruent with the phylogenetic tree.

The sample of languages is crucial for the distinction between COG origin by birth or borrowing because what may seem to be a word birth within a given sample of languages in our data could in fact be a borrowing event from a non-sampled language. How severe is the effect of external borrowing on our results? If we assume the extreme case, for example, that all COGs in the dataset originated by borrowing from external languages, then we have to add one borrowing event to the average rate for each COG. In that case, the average borrowing rate would increase from 0.6 to 1.6 events per COG using the Dyen dataset. However, this extreme scenario is unlikely because it entails the assumption that the Indo-European groups sampled here lacked the wherewithal to invent even one new COG. Nonetheless, external borrowing has almost certainly had an effect on these data. Although we currently lack a dataset that would allow us to quantify the rate of external borrowing, if we assume that it is similar to the internal borrowing rate within our sample, the overall borrowing rate would be double our current estimate. Again we stress that the borrowing frequency inferred from the present sample of languages using our method delivers a minimum value (a conservative lower bound).

Another aspect of the data sample used in our analysis is the collection of cognates. Here, we study the dynamics of vocabulary size during evolution through the proxy of basic vocabulary (i.e. the Swadesh list). However, origin and loss of words in the COGs sample can occur by semantic shift where the word is present in the language but absent from the sample. It is possible that different meaning collections evolve under regimens different from the ones described here. Application of similar methods to study vocabulary size dynamics over time using different cognate datasets will help to clarify this issue.

Notwithstanding certain amounts of cognate misjudgements and parallel evolution [48] resulting in tree-incompatible COG distributions, our inference uncovers abundant, and hitherto unrecognized, borrowing during the evolution of the Indo-European languages.

Scholars usually agree that nouns are more easily borrowed than verbs [49]. When classified according to the English gloss, the Dyen dataset includes 887 (53%) cognate sets corresponding to nouns within basic vocabulary and 766 (46%) cognate sets corresponding to verbs. A total of 503 (53%) nominal cognate sets and 450 (47%) verbal cognate sets were identified as including hidden borrowing events. A comparison of these frequencies shows that there is no significant difference in borrowing frequencies between nouns and verbs ($p = 0.4$, using the *G*-test).

(c) *Minimal lateral networks of Indo-European languages*

COG distributions that do not map exactly onto the phylogenetic tree, with borrowing constrained by ancestral

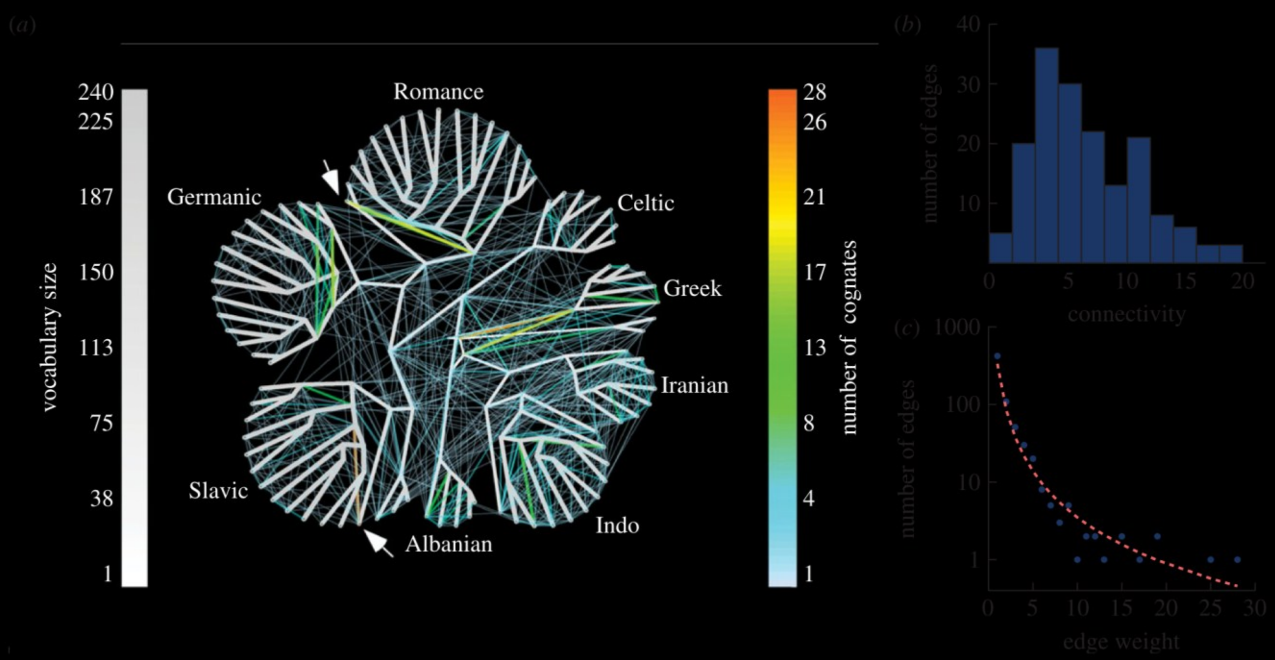


Figure 4. The MLN of Indo-European languages. (*a*) An MLN for 84 contemporary languages reconstructed under the BOR1 model. Vertical edges are indicated in grey, with both the width and the shading of the edge shown proportional to the number of inferred vertically inherited COGs along the edge (see the scale). The lateral network is indicated by edges that do not map onto the vertical component, with the number of cognates per edge indicated in colour (see the scale). Lateral edges that link ancestral nodes represent laterally shared COGs among the descendent languages of the connected nodes, whose distribution pattern could not be explained by origin and LO under the ancestral vocabulary size constraint. The two heaviest edges of Slovene (Slavic) and Romanian (Romance) are marked by an arrow. (*b*) Distribution of connectivity, the number of one-edge-distanced neighbours for each vertex, in the network. (*c*) Frequency distribution of edge weight in the lateral component of the network.

vocabulary size only, constitute the MLN [37]. The MLN reconstructed from the Dyen dataset consists of 167 vertices, of which 84 are contemporary and 83 are ancestral languages (internal nodes in the reference tree). The vertices are interconnected either by the branches of the reference tree, representing vertical inheritance, or by lateral edges, representing horizontal transfer (figure 4*a*).

The internal and external vertices in the MLN for the broad sample of COGs are linked by 666 lateral edges. The connectivity (number of edges per vertex) within the MLN ranges between 0 and 21 edges per language, with a median of 7 (figure 4*b*). The most highly connected node is Ossetic (21 edges), an east Iranian language, which is connected with Indo-Iranian, Greek and Slavic languages. Lateral edges connected to external nodes correspond to comparatively recent borrowing events. On average 8 ± 7% COGs per language are involved in recent borrowing (electronic supplementary material, table S3). This result suggests that English, at 8 per cent borrowing rate [13], is not exceptional; it is merely the most studied language. The clustering coefficient of the MLN is 0.22, and the mean shortest path is 3.128 edges. Combined with the high level of clustering, this means that the MLN forms a small-world network.

The edge weight distribution within the MLN is characterized by a majority of small edge weights. Of the total edges, 422 (63%) are of a single laterally shared COG, while edges of multiple COGs are rare (figure 4*c*). The two heaviest lateral edges include an edge between Slovene and the remaining Slavic languages (28 COGs), and an edge between Romanian and the remaining Romance languages (19 COGs). These lateral edges uncover a certain kind of language change that results from the same evolutionary process. Both Slovene and Romanian, being heavily influenced by neighbouring languages, underwent a process of linguistic revival starting from the early 19th century, in which the original

Table 1. Reconstructed borrowing events. The origin node that includes the reinserted borrowing is shaded in light grey.

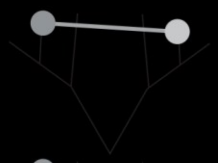
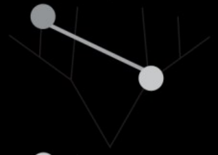
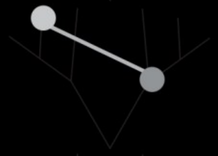
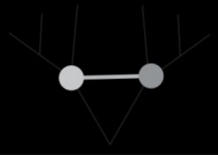
edge type	origin node	number of reinserted borrowings
external–external		1
external–internal		18
		58
internal–internal		40

Table 2. Lateral edge (LE) frequencies between and within groups in the MLN.

		normalized borrowing		median LE weight[b]		H_0:$\text{LE}_{\text{int}} \leq \text{LE}_{\text{ext}}$ frequency[c,d]
group	*n*[a]	int	ext	int	ext	*p*-value
Greek	9	1.22	0.25	2	1	<0.05
Armenian	3	0	0.17	0	1	n.a.
Celtic	13	1.61	0.29	2	1	$\ll 0.05$
Romance	31	2.45	0.36	1	1	$\ll 0.05$
Germanic	29	2.37	0.44	1	1	$\ll 0.05$
Slavic	31	2.35	0.64	1	1	$\ll 0.05$
Albanian	9	1.55	0.18	4	1	$\ll 0.05$
Indic	21	3.33	0.68	2	1	$\ll 0.05$
Iranian	14	2.35	0.75	2	1	$\ll 0.05$

[a] Number of languages within group.
[b] Range of median number of COGs per lateral edge.
[c] One-side Kolmogorov–Smirnov test for lateral edge distribution.
[d] For internal edges (int), number of internal edges per number of nodes within the group; for external edges (ext), number of external edges per number of nodes outside the group.

traits that had been lost during long periods of contact were artificially reintroduced into the languages by the speakers in order to bring them back to a stage of earlier ‘purity’ [50,51]. Before the 19th century, Slovene comprised several dialects spoken in the Alpine provinces of the Austrian Empire, which were dominated by German and Italian. Romanian, on the other hand, was heavily influenced by neighbouring Slavic and Greek varieties, with which it formed the so-called Balkan *Sprachbund*. Along with the nationalist movements in Europe starting from the end of the 18th century, both languages were successively ‘purified’ by replacing the loanwords of non-Slavic or non-Romance origin with ‘native’ words from Slavic or Romance languages, respectively [50,51]. This process is somewhat different from the process of borrowing as it was defined in the beginning of this paper. It nonetheless illustrates additional horizontal complexities in the processes of language evolution that are readily detected in the MLN.

The comparison between the edges reconstructed using the two reference trees that differ in their English position supplies a few interesting observations regarding the applicability of our approach to detect borrowing events. While both reference trees yielded the same borrowing model (i.e. the same overall borrowing rates), there are 23 lateral edges connecting to English in the basal position and only 15 lateral edges connecting to English in the internal position. A closer inspection of the COGs in which the lateral edges connecting to English were detected revealed that seven of the eight COGs detected as borrowings in the basal position could not be verified as borrowings by traditional historical linguistics. Thus, using different reference trees with the same COG distribution patterns does not much affect the resulting borrowing model, but it may increase the accuracy of concrete predictions made by this approach (see electronic supplementary material, table S4 for detailed etymological reconstruction of the COGs). Consequently, the borrowing inference accuracy in our approach is expected to increase with the accuracy of the reference tree.

The MLN inferred from the ToB dataset shows similar network characteristics, with the ancestors of Indian and Iranian clades found also as highly connected nodes and a majority (676; 76%) of single laterally shared COGs (electronic supplementary material, figure S6).

Of the total 666 edges in the MLN reconstructed for the Dyen dataset, 148 (22%) edges connect between two external nodes—that is, between two contemporary languages. The 301 (45%) edges that connect between an internal node and an external node represent COGs that are shared between a group and an outlier. The 217 (33%) edges that connect between two internal nodes represent COGs that are common to two different groups, yet their distribution pattern could not be explained by vertical inheritance alone under the vocabulary size criterion. As a control to see whether our method is inferring spurious borrowing, we examined the edges within cognates that included the 124 reinserted borrowing events. In seven cognates, the algorithm detected no borrowings, while in all other 117 (94%) cognates a borrowing event was inferred. In 59 (48%), the reinserted borrowing language was inferred as an external node. In the remaining 58 (47%), reinserted borrowing languages were inferred within descendants of an internal node (table 1).

The data can address the issue of whether words are exchanged more frequently within than between main branches of Indo-European. We can compare the probability of a certain language to be laterally connected with languages that are either from the same main branch or from different main branches of the Indo-European languages. With the exception of the Armenian branch, the probability for a lateral edge within the branch (internal edge) is considerably higher than between branches (external edge). Furthermore, lateral edge weights are significantly larger in internal lateral edges than in external lateral edges (table 2). Hence, lexical borrowing in Indo-European languages is much more frequent among languages within the same branch in comparison to languages from different branches. This provides new evidence for the existence of certain cultural barriers to lexical borrowing during language evolution [10].

The study was supported by the German Federal Ministry of Education and Research (S.N.S., J.M.L., H.G., T.D. and W.M.) and the European Research Council (W.M.). We are thankful to Frank Kressing, Matthis Krischel, Thorsten Halling and Sven Sommerfeld for helpful discussions, and to Dan Graur for his help in refining the manuscript. We thank Liat Shavit-Grievink for her help in phylogenetic reconstruction.

REFERENCES

1 Darwin, C. 1859 *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. London, UK: John Murray. See http://www.nla.gov.au/apps/cdview/nla.gen-vn4591931.

2 Schleicher, A. 1863 *Die Darwinsche Theorie und die Sprachwissenschaft offenes Sendschreiben an Herrn Dr. Ernst Häckel*, 3rd edn (1873). Weimar, Germany: Böhlau.

3 Gray, R. D. & Atkinson, Q. D. 2003 Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439. (doi:10.1038/nature02029)

4 Pagel, M. 2009 Human language as a culturally transmitted replicator. *Nat. Rev. Genet.* **10**, 405–415.

5 Dunn, M., Terrill, A., Reesink, G., Foley, R. A. & Levinson, S. C. 2005 Structural phylogenetics and the reconstruction of ancient language history. *Science* **309**, 2072–2075. (doi:10.1126/science.1114615)

6 Lansing, J. S. *et al.* 2007 Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc. Natl Acad. Sci. USA* **104**, 16 022–16 026. (doi:10.1073/pnas.0704451104)

7 Pagel, M. & Atkinson, Q. D. 2007 A Meade frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–720. (doi:10.1038/nature06176)

8 Thomason, S. G. 2001 *Language contact: an introduction*. Edinburgh, UK: Edinburgh University Press.

9 Trask, R. L. 2000 *The dictionary of historical and comparative linguistics*. Edinburgh, UK: Edinburgh University Press.

10 Thomason, S. & Kaufman, T. 1988 *Language contact, creolization, and genetic linguistics*. Berkeley, CA: University of California Press.

11 Aikhenvald, A. Y. 2006 Grammars in contact: a cross-linguistic perspective. In *Grammars in contact: a cross-linguistic typology* (eds A. Y. Aikhenvald & R. M. Dixon), pp. 1–66. Oxford, UK: Oxford University Press.

12 Fox, A. 1995 *Linguistic reconstruction: an introduction to theory and method*. Oxford, UK: Oxford University Press.

13 Embleton, S. 2000 Lexicostatistics/glottochronology: from Swadesh to Sankoff to Starostin to future horizons. In *Time depth in historical linguistics* (eds C. Renfrew, A. McMahon & L. Trask), pp. 143–165. Cambridge, UK: The McDonald Institute for Archaeological Research.

14 Bergsland, K. & Vogt, H. 1962 On the validity of glottochronology. *Curr. Anthropol.* **3**, 115–153. (doi:10.1086/200264)

15 Boyd, R., Borgerhoff, M. M., Durham, W. H. & Richerson, P. J. 1997 Are cultural phylogenies possible? In *Human by nature, between biology and the social sciences* (eds P. Weingart, P. J. Richerson, S. D. Mitchell & S. Maasen), pp. 355–386. Mahwah, NJ: Erlbaum.

16 Atkinson, Q. D. & Gray, R. D. 2006 How old is the Indo-European language family? Illumination or more moths to the flame? In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 91–109. Cambridge, UK: McDonald Institute for Archaeological Research.

17 Schleicher, A. 1853 Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Monatsschrift für Wissenschaft und Literatur*, **September**, 786–787.

18 Schmidt, J. 1872 *Die Verwantschaftsverhältnisse der indogermanischen Sprachen*. Weimar, Germany: Hermann Böhlau.

19 Schuchardt, H. 1922 Über die Klassifikation der romanischen Mundarten. In *Hugo Schuchardt-Brevier. Ein Vademekum der allgemeinen Sprachwissenshaft. Als Festgabe zum 80. Geburtstag des Meisters zusammengestellt und eingeleitet von Leo Spitzer* (ed. L. Spitzer), pp. 144–166. Halle, Germany: Max Niemeyer.

20 Hirt, H. 1905 *Die Indogermanen. Ihre Verbreitung, ihre Urheimat und ihre Kultur*, vol. 1. Strassburg, France: Trübner.

21 Bonfante, G. I. 1931 I dialetti indoeuropei. *Annali del R. Istituto Orientale di Napoli* **4**, 69–185.

22 Dyen, I., James, A. T. & Cole, J. W. L. 1967 Language divergence and estimated word retention rate. *Language* **43**, 150–171. (doi:10.2307/411390)

23 Ringe, D. A. 1992 On calculating the factor of chance in language comparison. *Trans. Am. Phil. Soc.* **82**, 1–110. (doi:10.2307/1006563)

24 Southworth, F. C. 1964 Family-tree diagrams. *Language* **40**, 557–565. (doi:10.2307/411938)

25 Bryant, D., Filimon, F. & Gray, R. D. 2005 Untangling our past: languages, trees, splits and networks. In *The evolution of cultural diversity: phylogenetic approaches* (eds R. Mace, C. Holden & S. Shennan), pp. 67–84. London, UK: UCL Press.

26 Nakhleh, L., Ringe, D. & Warnow, T. 2005 Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. *Language* **81**, 382–420. (doi:10.1353/lan.2005.0078)

27 McMahon, A., Heggarty, P., McMahon, R. & Slaska, N. 2005 Swadesh sublists and the benefits of borrowing: an Andean case study. *Trans. Phil. Soc.* **103**, 147–170. (doi:10.1111/j.1467-968X.2005.00148.x)

28 Ben Hamed, M. & Wang, F. 2006 Stuck in the forest: trees, networks and Chinese dialects. *Diachronica* **23**, 29–60.

29 Dyen, I., Kruskal, J. B. & Black, P. 1997 *Comparative Indo-European database: file IEdata1*. See http://www.wordgumbo.com/ie/cmp/iedata.txt.

30 Starostin, G. 2008 *Tower of Babel: an etymological database project*. See http://starling.rinet.ru.

31 Newman, M. E. J. 2003 The structure and function of complex networks. *SIAM Rev.* **45**, 167–256. (doi:10.1137/S003614450342480)

32 Ronquist, F. & Huelsenbeck, J. P. 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574. (doi:10.1093/bioinformatics/btg180)

33 Saitou, N. & Nei, M. 1987 The neighbor-joining method. A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.

34 Huson, D. H. & Bryant, D. 2006 Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267. (doi:10.1093/molbev/msj030)

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. (doi:10.1073/pnas.0606318104)

36 Zar, J. H. 1999 *Biostatistical analysis*, 4th edn. Englewood Cliffs, NJ: Pearson Prentice-Hall.

37 Dagan, T., Artzy-Randrup, Y. & Martin, W. 2008 Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl Acad. Sci. USA* **105**, 10 039–10 044. (doi:10.1073/pnas.0800679105)

38 Swadesh, M. 1955 Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Linguist.* **21**, 121–137. (doi:10.1086/464321)

39 Swadesh, M. 1952 Lexicostatistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proc. Am. Phil. Soc.* **96**, 452–463.

40 Geisler, H. & List, J.-M. 2011 Beautiful trees on unstable ground: notes on the data problem in lexicostatistics. In *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik, Akten der Arbeitstagung der Indogermanischen Gesellschaft Würzburg, 24–26 September 2009* (ed. H. Hettrich). Wiesbaden, Germany: Reichert.

41 Girvan, M. & Newman, M. E. J. 2002 Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **12**, 7821–7826.

42 Dyen, I., Kruskal, J. B. & Black, P. 1992 An Indoeuropean classification: a lexicostatistical experiment. *Trans. Am. Phil. Soc.* **82**, 3–132.

43 Mallory, J. P. & Adams, D. Q. 2006 *The Oxford introduction to Proto-Indo-European and the Proto-Indo-European world*. Oxford, UK: Oxford University Press.

44 Wells, R. S. 1973 Uniformitarianism in linguistics. In *Dictionary of the history of ideas* (ed. P. Wiener), pp. 423–431. New York, NJ: Scribner.

45 Christy, C. 1983 *Uniformitarianism in linguistics*. Amsterdam, The Netherlands: John Benjamins.

46 Greenhill, S. J., Currie, T. E. & Gray, R. D. 2009 Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. B* **276**, 2299–2306. (doi:10.1098/rspb.2008.1944)

47 Lewis, P. M. 2009 *Ethnologue: languages of the world*, 16th edn. Dallas, TX: SIL International. See http://www.ethnologue.com.

48 Garrett, A. 2006 Convergence in the formation of Indo-European subgroups: phylogeny and chronology. In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 139–151. Cambridge, UK: McDonald Institute for Archaeological Research.

49 Hock, H. H. & Joseph, B. D. 2009 Language history, language change and language relationship. In *An introduction to historical and comparative linguistics*, 2nd edn. Berlin, Germany: Mouton de Gruyter.

50 Auty, R. 1963 The formation of the Slovene literary language against the background of the Slavonic national revival. *SEER* **41**, 391–402.

51 Mallinson, G. 1988 The Romance languages in. In *The Romance languages* (eds M. Harris & V. Nigel), pp. 391–419. London, UK: Croom Helm.

52 Orel, V. 2003 *A handbook of Germanic etymology*. Leiden, The Netherlands: Brill.

53 Soukhanov, A. H. 1992 *The American heritage dictionary of the English language*. Boston, MA: Mifflin.

9 Genomic imprints of chaperone-mediated folding.

9.1 Chaperonin-dependent accelerated protein evolution in prokaryotes

Bogumil D, Dagan T: **Chaperonin-dependent accelerated substitution rates in prokaryotes.** *Genome Biol Evol* 2010, **2**:602-608.

(Own contribution: conceived and designed the experiment, performed part of the analysis, analyzed the data, and wrote the paper).

Chaperonin-Dependent Accelerated Substitution Rates in Prokaryotes

Abstract

Many proteins require the assistance of molecular chaperones in order to fold efficiently. Chaperones are known to mask the effects of mutations that induce misfolding because they can compensate for the deficiency in spontaneous folding. One of the best studied chaperones is the eubacterial GroEL/GroES system. In *Escherichia coli*, three classes of proteins have been distinguished based on their degree of dependency on GroEL for folding: 1) those that do not require GroEL, 2) those that require GroEL in a temperature-dependent manner, and 3) those that obligately require GroEL for proper folding. The buffering effects of GroEL have so far been observed in experimental regimens, but their effect on genomes during evolution has not been examined. Using 446 sequenced proteobacterial genomes, we have compared the frequency of amino acid replacements among orthologs of 236 proteins corresponding to the three categories of GroEL dependency determined for *E. coli*. Evolutionary rates are significantly correlated with GroEL dependency upon folding with GroEL dependency class accounting for up to 84% of the variation in amino acid substitution rates. Greater GroEL dependency entails increased evolutionary rates with GroEL obligatory proteins (Class III) evolving on average up to 15% faster than GroEL partially dependent proteins (Class II) and 35% faster than GroEL-independent proteins (Class I). Moreover, GroEL dependency class correlations are strictly conserved throughout all proteobacteria surveyed, as is a significant correlation between folding class and codon bias. The results suggest that during evolution, GroEL-dependent folding increases evolutionary rate by buffering the deleterious effects of misfolding-related mutations.

Key words: genome evolution, misfolding, GroEL, codon usage.

Introduction

Ellis 1987

Young et al. 2004

Rutherford 2003

Queitsch et al. 2002

Fares et al. 2002 Maisnier-Patin et al. 2005

Moran 1996

Moran 1996

Todd et al. 1996 Fares et al. 2002 Queitsch et al. 2002

Maisnier-Patin et al. 2005 Tokuriki and Tawfik 2009

Young et al. 2004

Kerner et al. 2005

Horwich et al. 1993

Lund et al. 2003

Kerner et al. 2005

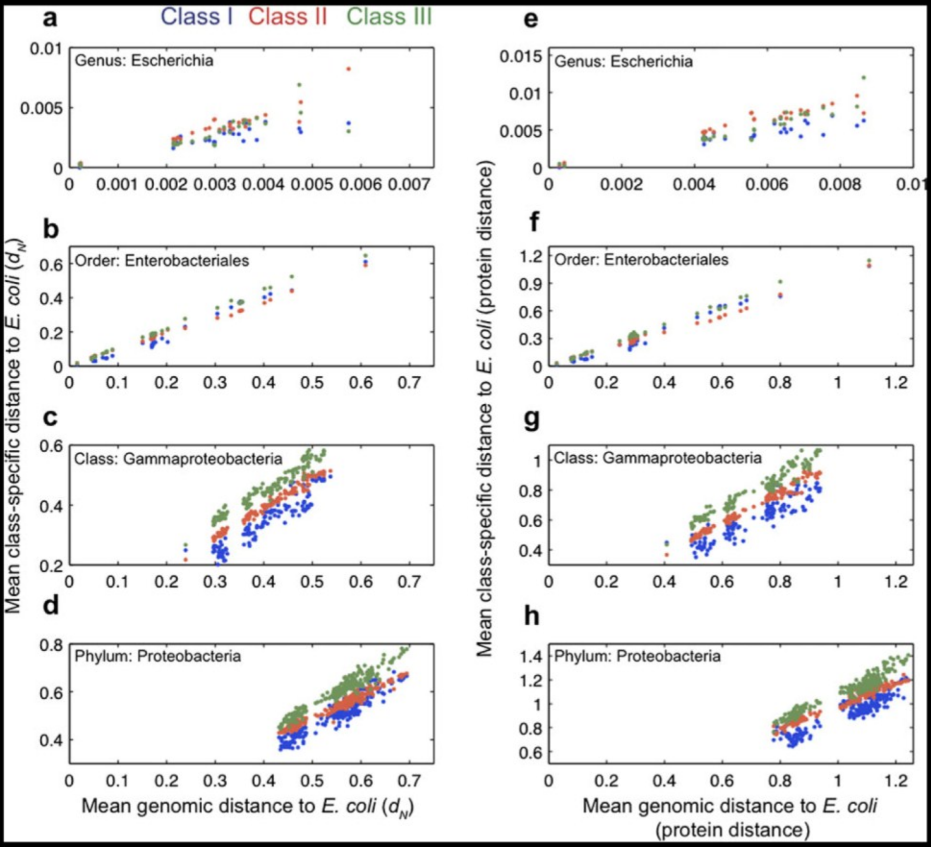


Fig. 1.

Kerner et al. 2005

Fares et al. 2002 Queitsch et al. 2002 Maisnier-Patin et al. 2005 Tokuriki and Tawfik 2009

Materials and Methods

Kerner et al. (2005) Kerner et al. (2005)

http://www.ncbi.nlm.nih.gov/

Kerner et al. (2005) Altschul et al. 1990

Tatusov et al. 1997

Thompson et al. 1994 Landan and Graur 2007

Suyama et al. 2006

Yang 2007

Felsenstein 2005

Jones et al. 1992

Table 1

Sharp and Li 1987

Rice et al. 2000

Results

Thompson et al. 1994

Nei and Gojobori 1986

Yang 2007

fig. 1

fig. 1*a*

Zar 1999

Table 2

Zar 1999

fig. 1

table 1
fig. 1c

table 2

fig. 1*d*
table 1

table 2

fig. 1*e–h*
Jones et al. 1992

table 2

Krylov et al. 2003
Drummond et al. 2006 Pál et al. 2006

Warnecke and Hurst 2010
Lu et al. 2007

fig. 2

table 3

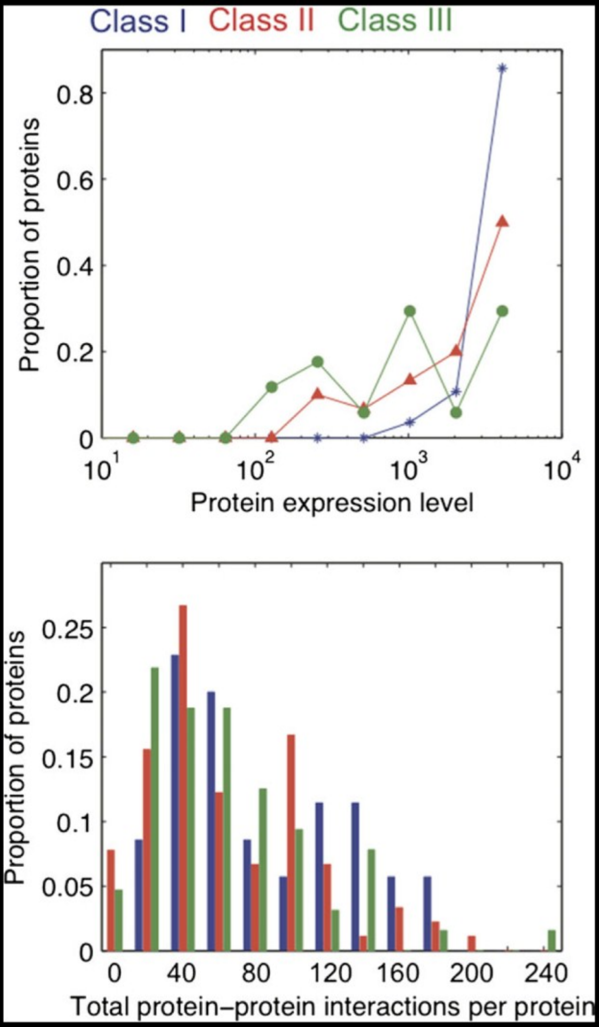


Fig. 2. Lu et al. 2007 Hu et al. 2009

von Mering et al. 2002

von Mering et al. 2002

Hu et al. (2009)

fig. 2
Sharp and Li 1987

Sharp and Li 1987

Table 3

table 1 fig. 3
Warnecke and Hurst 2010

Fujiwara et al. 2010

supplementary table S1 Supplementary Material

supplementary table S1 supplementary fig. S1 Supplementary Material

supplementary table S1 Supplementary Material

supplementary table S2 supplementary figs. S2 S5 Supplementary Material

supplementary table S1 Supplementary Material

Fujiwara et al. 2010

Discussion

Drummond et al. 2005 Drummond and Wilke 2008 Lobkovsky et al. 2010 Warnecke and Hurst 2010

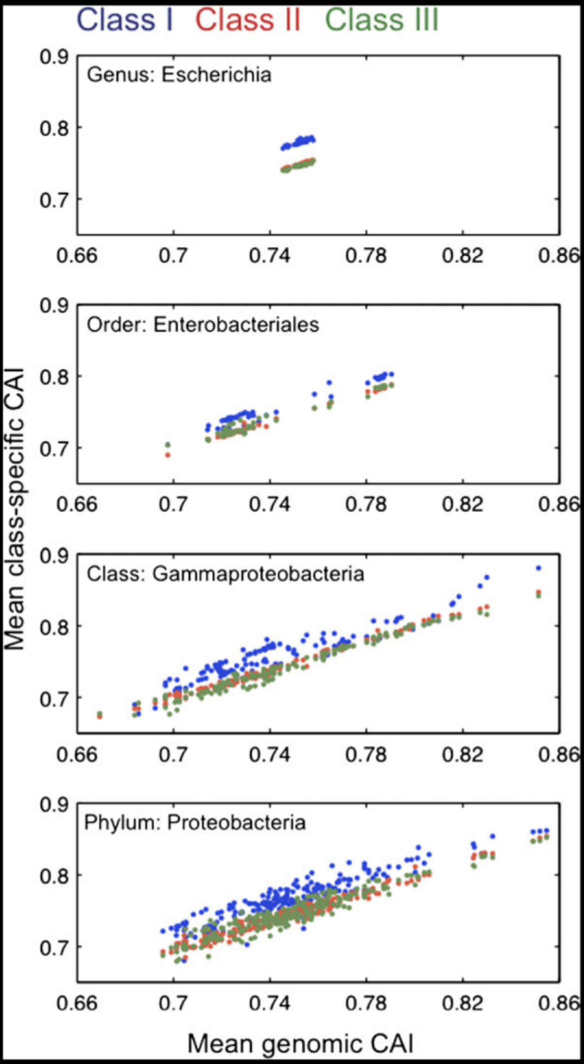


Fig. 3.

fig. 1 fig. 2 Warnecke and Hurst 2010 fig. 3

Pál et al. 2006

Drummond and Wilke 2008 Tuller et al. 2010

Supplementary Material

Supplementary figures S1 S6 tables S1 S2 http://www.oxfordjournals.org/our_journals/gbe/

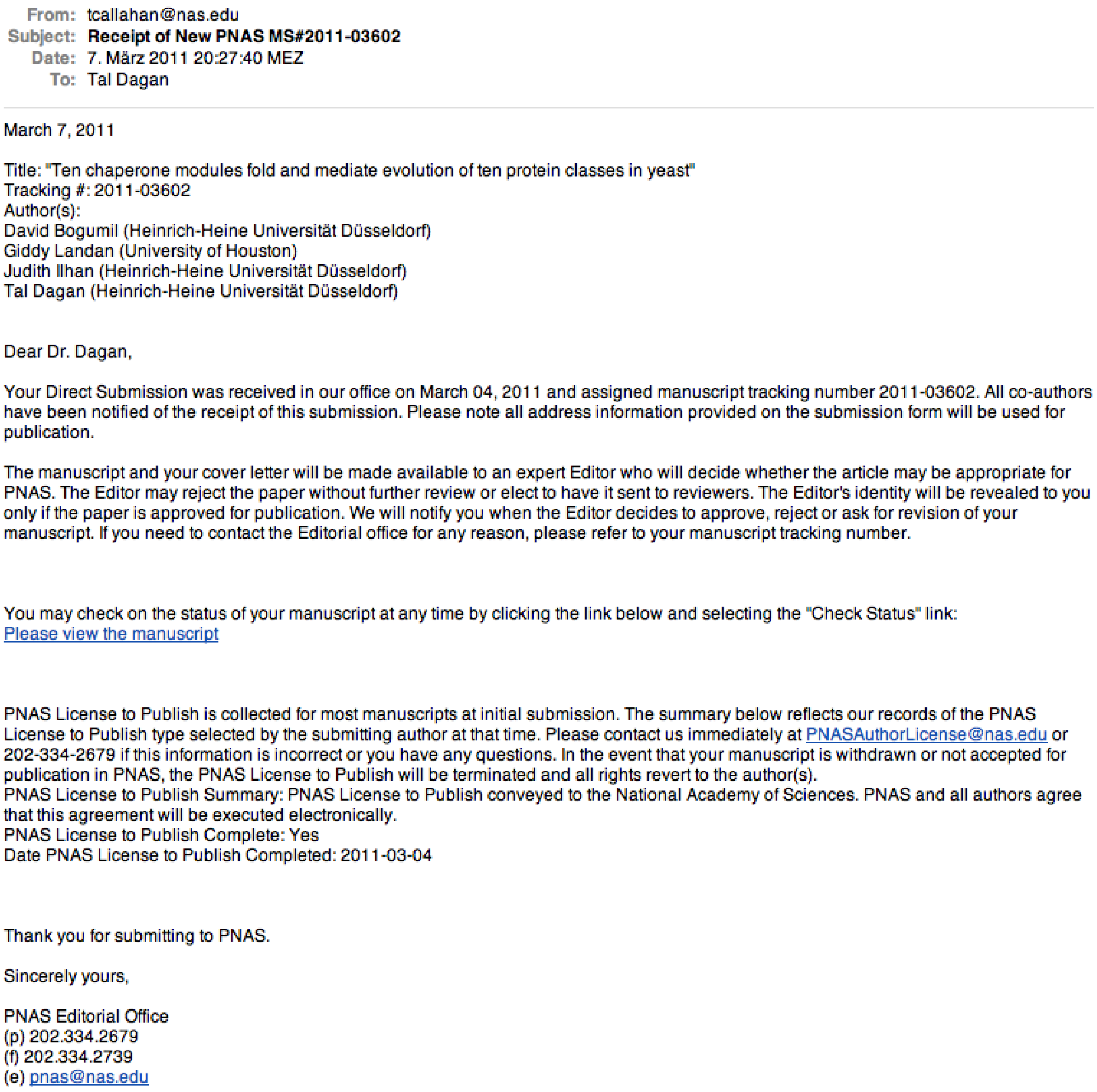
Acknowledgments

Literature Cited

9.2 Chaperone mediated protein evolution in yeast

Bogumil D, Landan G, Ilhan J, Dagan T: **Ten chaperone modules fold and mediate evolution of ten protein classes in yeast**, 2011. *submitted*.

(Own contribution: conceived and designed the experiment, performed part of the analysis, analyzed the data, and wrote the paper).



From: tcallahan@nas.edu
Subject: **Receipt of New PNAS MS#2011-03602**
Date: 7. März 2011 20:27:40 MEZ
To: Tal Dagan

March 7, 2011

Title: "Ten chaperone modules fold and mediate evolution of ten protein classes in yeast"
Tracking #: 2011-03602
Author(s):
David Bogumil (Heinrich-Heine Universität Düsseldorf)
Giddy Landan (University of Houston)
Judith Ilhan (Heinrich-Heine Universität Düsseldorf)
Tal Dagan (Heinrich-Heine Universität Düsseldorf)

Dear Dr. Dagan,

Your Direct Submission was received in our office on March 04, 2011 and assigned manuscript tracking number 2011-03602. All co-authors have been notified of the receipt of this submission. Please note all address information provided on the submission form will be used for publication.

The manuscript and your cover letter will be made available to an expert Editor who will decide whether the article may be appropriate for PNAS. The Editor may reject the paper without further review or elect to have it sent to reviewers. The Editor's identity will be revealed to you only if the paper is approved for publication. We will notify you when the Editor decides to approve, reject or ask for revision of your manuscript. If you need to contact the Editorial office for any reason, please refer to your manuscript tracking number.

You may check on the status of your manuscript at any time by clicking the link below and selecting the "Check Status" link:
Please view the manuscript

PNAS License to Publish is collected for most manuscripts at initial submission. The summary below reflects our records of the PNAS License to Publish type selected by the submitting author at that time. Please contact us immediately at PNASAuthorLicense@nas.edu or 202-334-2679 if this information is incorrect or you have any questions. In the event that your manuscript is withdrawn or not accepted for publication in PNAS, the PNAS License to Publish will be terminated and all rights revert to the author(s).
PNAS License to Publish Summary: PNAS License to Publish conveyed to the National Academy of Sciences. PNAS and all authors agree that this agreement will be executed electronically.
PNAS License to Publish Complete: Yes
Date PNAS License to Publish Completed: 2011-03-04

Thank you for submitting to PNAS.

Sincerely yours,

PNAS Editorial Office
(p) 202.334.2679
(f) 202.334.2739
(e) pnas@nas.edu

BIOLOGICAL SCIENCES: Evolution

Ten chaperone modules fold and mediate evolution of ten protein classes in yeast

David Bogumil[1], Giddy Landan[2], Judith Ilhan[1], Tal Dagan[1]

[1] Institute of Botany III, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany
[2] Department of Biology & Biochemistry, University of Houston, Houston, Texas 77204-5001, USA

Corresponding author: Tal Dagan, Institut für Botanik III, Heinrich-Heine Universität Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany, Tel: +49 211 811 2736, Fax: +49 211 811 3554, e-mail: tal.dagan@uni-duesseldorf.de

ABSTRACT

It has long been known that many proteins require folding via molecular chaperones for their function. Although it has become apparent that folding imposes constraints on protein sequence evolution, the effects exerted by different chaperone classes is so far unknown. We have analyzed chaperone-substrate interaction data in *S. cerevisiae* using network methods. The results reveal a distinct community structure within the network that was hitherto undetectable with standard statistical tools. The 69 yeast chaperones comprise ten distinct modules that are defined by their interaction specificity for their 3,595 polypeptide substrates. The substrate classes defined by their dedicated chaperone modules are distinguished by various physiochemical protein properties, but not by sequence motifs, and are characterized by significantly different amino acid substitution rates, codon usage, and protein expression levels. Although correlations between substitution rate, codon bias, and gene expression level have long been known for yeast, such correlations are, dramatically, two-fold stronger for the chaperone-defined modules that we report here than they are for the whole proteome. This indicates that correlated expression, conservation and codon bias levels for yeast genes are mainly attributable to previously unrecognized effects of protein folding. These results uncover proteome-wide categories of chaperone-substrate specificity as an overriding functional constraint that has been preserved throughout fungal evolution. The data are consistent with the view that aggregation of misfolded proteins imposes fitness costs during evolution and furthermore strongly suggest that codon usage is selected during evolution not for optimal translation efficiency, but for optimal synchronization between protein translation and protein folding so as to avoid accumulation of misfolded protein.

/body

INTRODUCTION

Chaperones are essential in all living cells as they assist protein folding, prevent protein aggregation, and play a crucial role in survival under stress conditions (1-2). Chaperons have been shown to buffer the effects of slightly deleterious mutations, presumably by compensating for decreased folding fidelity of mutated proteins (3-4). The protein-folding pathway in S. cerevisiae comprises 69 molecular chaperones and their co-chaperones that assist the folding or unfolding of proteins in the cell. Most of the proteins encoded in the yeast genome (3,595 out of 5,880) interact with at least one chaperone, many of them (2,952) with two or more chaperones (5). The extent to which chaperone mediated folding of proteins has an effect on their evolutionary dynamics is not yet known.

To address this question, we used a network approach to investigate an extensive dataset of chaperone-substrate interactions assembled by screening for chaperone interactors in yeast (5). A network is a set of entities, or vertices, linked by edges that represent the connections or interactions between these entities. The entities in our network can be either a chaperone or a protein (a substrate). A network of N vertices can be fully defined by a matrix, $A=[a_{ij}]N*N$, with a_{ij} = 1 if chaperone *i* and protein *j* interact and a_{ij} = 0 otherwise. The chaperones and substrates form two disjoint sets of nodes where interactions between substrate nodes are not allowed, because the data reflect the interactions of chaperones with substrate proteins, but not other possible interactions among the substrate proteins. The network is thus semi multipartite, containing 9,194 edges of chaperone-substrate interactions and 332 edges of chaperone-chaperone interactions. Substrate-chaperone interactions reflect chaperone-mediated folding of the substrate, chaperone-chaperone interactions reflect either a chaperone-mediated folding of another chaperone, or an interaction between a chaperone and co-chaperone for folding of a common substrate. Co-chaperones in our network were found to interact almost exclusively with chaperones.

RESULTS AND DISCUSSION

The substrate-chaperone interaction network comprises five highly connected Hsp70 chaperones that are linked to 3,595 substrates in total. The remaining 64 chaperones interact with selected proteins ranging between 2 and 732 substrates per chaperone. Some chaperones interact with a similar set of substrates, thereby forming communities within the network. We examined the community structure in the network by partitioning it into modules (6). For each possible bipartition of the network, a modularity function is defined as the observed number of edges within a community minus the expected number. Maximizing this modularity function using its leading eigenvector yields the modules within the network (6).

The result uncovered ten modules that include a total of 64 chaperones and 2,691 substrates, along with 843 lesser (residual) modules that contain a single protein each. Co-chaperones are grouped into the modules based on their interaction with the chaperones. The modules comprise chaperone groups that interact frequently with common substrates. Five Hsp70 chaperones were not grouped into the ten main modules, forming five single-chaperone modules (Ssa1, Ssa2, Ssb1, Ssb2, and Sse1). These chaperones are characterized by a promiscuous substrate binding and have many substrates in common (*5*) (Fig. 1A). We designate the ten main modules by their most connected chaperone. The modules contain between one (HSP70-Ssa3) and 14 (SMALL-Hsp42) chaperones. The number of substrates folded by each module ranges from 65 (CCT-Cct8) to 485 (AAA+-Hsp78) (Figure 1B).

The majority of nascent polypeptides in the yeast protein-folding pathway interact with the ribosome-associated complex (RAC) that comprises a member of the HSP70 family and a co-chaperone from the HSP40 family (J-proteins). Selected proteins also interact with one or more of the following chaperone classes: prefoldin (PFD), TriC (CCT), and HSP90 (*7*). Module AAA+-Hsp78, for example, includes chaperones from HSP70, HSP40, PFD, and TriC chaperone families (Fig. 1B; Table S1). Three modules (SMALL-Hsp42, HSP90-Hsp82, CCT-Cct8) contain only an HSP40 chaperone lacking the obligatory partner from HSP70 family. However, all substrates in these modules also interact with one or more of the five ungrouped HSP70 chaperones. Two modules, HSP70-Ssa3 and HSP70-Ssa4, include only an HSP70 chaperone lacking an HSP40 partner. Substrates in those two modules interact with various HSP40 chaperones and with the Ydj1, which has no substrate specificity (*8*), as the most common interactor.

Members within the modules are not restricted to a certain cellular localization (Fig. S1), hence the cellular locations of protein folding and protein function do not always overlap[7]. Module HSP90-Hsc82 is however enriched with chaperones localized in the mitochondrion (6 out of 9). The module includes HSP60 and HSP10 that interact to fold proteins in the mitochondrion (*9*). These two chaperones are homologous to the eubacterial GroEL/GroES chaperonin system (*10*). Furthermore, the HSP70 (Ssc1), HSP40 (Mdj2, Zuo1), and HSP90 (Hsc82) chaperones in this module are known to be localized in the mitochondrion (Table S1). Notably, the HSP90-Hsc82 module is lacking both PFD and TriC chaperones, which are homologous to archaeal chaperones (*7*). The evolutionary origin of the chaperones in this module suggests that it is of mitochondrial origin, reflecting a functional eubacterial unit within the yeast proteome (*11*).

To test the impact of chaperone-mediated folding on protein evolution we compared substrate amino acid substitution rate among the modules. Weighted amino acid substitution rate per site was calculated from a pairwise alignment of *S. cerevisiae* substrate proteins with their positional ortholog from among 20 sequenced fungal genomes (*12*). A comparison of amino acid substitution rate distribution among the ten modules revealed significant differences across the modules (p < $2.2\text{x}10^{-16}$, using Friedman test). The same result is obtained for the comparison of amino acid substitution frequency per site (p = $2.58\text{x}10^{-7}$, using Friedman test). Randomizing the module classification of substrates eliminates the differences in evolutionary rate among the modules (p = 0.82, using Friedman test). Using a post-hoc comparison of module amino acid substitution rate distribution, the chaperone-substrate modules were grouped into four evolutionary rate categories: slow, medium,

medium-fast, and fast evolving substrates. A comparison of module ranking across sequenced fungal genomes revealed that module rate category is conserved in all 20 species sampled here. Modules CCT-Cct8, SMALL-Hsp31, and AAA+-Hsp78 have the lowest mean amino acid substitution rate, while modules HSP70-Ssz1, HSP70-Ssa4, and HSP40-Sis1 have the highest rate, regardless of the fungal genome used for the comparison (Fig. 1B; Fig. 2). The same result is observed when using *Debaryomyces hansenii* or *Kluyveromyces waltii*, instead of yeast, as the reference species for the comparison (Table S2). Substrates that interact exclusively with the five ungrouped HSP70 chaperones evolve at an evolutionary rate that is comparable to the medium-fast rate category (Fig. 2). Substrates in the fast rate modules evolve on average 10% faster than substrates in the medium rate modules and 25% faster than substrates in the slow rate categories.

Amino acid substitution rate, protein expression level, and codon adaptation are known to be correlated at the genome level (*13-17*). Theories to explain the significant correlation among these three gene characteristics currently evoke either poorly specified network properties of proteins (*13,18-19*) or the specific effects of amino acid misincorporation during protein translation (*20-21*). Amino acid substitution rate, protein expression level, and codon adaptation are correlated for the subset of yeast proteins included in the substrate-chaperone interaction network (Table 1). A comparison of substrate expression level in yeast revealed significant differences among the modules ($p < 2.2x10^{-16}$, using Kruskal-Wallis test). Moreover, mean substrate expression level and mean amino acid substitution rate are significantly inversely correlated among the modules. The correlation coefficient at the module level is r_s = -0.95 ($p < 2.2x10^{-16}$), a value that is two-fold stronger than the correlation coefficient observed at the proteome level (r_s = -0.45, $p < 2.2x10^{-16}$; Table 1). The correlation between substrate expression level and substitution rate at the module level is significantly different from the expected by random (Table S3). Similarly, amino acid substitution rate and codon adaptation are much more strongly correlated at the module level, and so are codon adaptation and substrate expression level (Table 1). The correlation between substitution rate, expression level and codon adaptation among yeast proteins that are disconnected from the substrate-chaperone interaction network are similar to that of the connected proteins (Table S3). However, disconnected proteins have a significantly lower expression level than connected proteins ($p = 2.8x10^{-62}$, using one-sided Kolmogorov-Smirnov). This suggests that disconnected proteins are folded by chaperones but were so far not detected in surveys for chaperone interactors, possibly due to their low expression level.

However, if the reliance on chaperones for folding is related to protein expression level (*22*), then it is possible that expression level is the determinant of evolutionary rate differences among the substrates in the modules, and not chaperone-mediated folding itself. A comparison of amino acid substitution rates among the ten modules while adjusting for the variability in protein expression level reveals that modules in the slow and fast rate categories have significantly different substitution rates, also when adjusted for variability in substrate expression levels (α=0.05, using ANCOVA; Table S4A). The comparison of amino acid substitution rate among modules within the same rate category or between medium and fast or medium and slow categories shows no significant difference when adjusted for protein expression levels (α=0.05, using ANCOVA; Table S4A). Similar results are obtained when adjusting the variability in amino acid substitution rate for differences in CAI (α=0.05, using ANCOVA; Table S4B). Hence the stark differences observed between modules in the slow and fast rate categories cannot be explained by different selection pressures that are related to expression level or codon adaptation; rather they are attributable to chaperone-mediated folding.

Substrates in the ten modules differ significantly in their physiochemical properties. Amino acid usage of all twenty amino acids as well as secondary structure composition are significantly different among the modules (α=0.05, using Kruskal-Wallis test; Table S3). Valine amino acid usage is negatively correlated with substitution rate at the modules level (r_s = -0.79, p = 0.01). The side chain of valine is involved in hydrophobic interactions that stabilize the protein structure. Exposed hydrophobic side chains in unstructured proteins are a major cause of protein aggregation (*7*). Hence this correlation may be due to selective forces related to either protein folding or disentangling of protein aggregates (*23*). No clear enrichment for substrate functional category, cellular localization, chromosomal location (Fig. S1), protein domain (Table S5), or sequence motif (Table S6) was found among the modules.

Our results suggest that the correlation between evolutionary rate, expression level, and codon adaptation are a manifestation of the protein interaction with chaperones for folding. The question that remains is how protein interaction with the chaperone is related to protein expression level and codon adaptation. We suggest that these stem from the requirement for synchronization between protein translation and protein folding. Recently it was shown that codon usage distribution along the protein sequence plays a role in protein translation speed (*24,25*). Proteins that require chaperones have to be translated at a speed that fits the time required for chaperone recruitment, otherwise the protein will fold spontaneously into the wrong conformation thereby forming aggregates that hinder the cell viability (*26*). Proteins that can fold spontaneously into their functional conformation are free from that constraint and can be translated at a higher speed. However, with increasing translation speed, accuracy becomes more important, so that proteins that are translated at high speed should be more conserved (*21*). Chaperone-mediated folding ensures proper functional conformation, but it costs both time and energy. For example, protein folding by the GroEL/GroES chaperonin system in *Escherichia coli* takes about 10 seconds and consumes seven ATP molecules (*27*). It is therefore probably advantageous to have a subset of the proteome that is less dependent upon chaperones. This subset is probably defined by high expression levels and short response time.

Chaperon interaction with the protein is quite flexible. For example, the mode of interaction with GroEL/GroES chaperonin system in *E. coli* can vary between casual and obligatory substrates. Casual interactors bind to GroEL *in vivo* but can also gain functional activity independent of GroEL *in vitro* (*28*). A recent genomic analysis revealed that casual GroEL substrates have significantly higher expression level than obligatory substrates (*29*). Together with the results presented here, this suggests that protein abundance within the cell largely determines the kind and mode of interaction with the chaperones for folding. Chaperon mediated folding has a profound cumulative impact on genome evolution.

Methods

Modules within the network. Interaction data of *S. cerevisiae* proteins with 69 chaperones were taken from Gong et al. (*5*), which conducted a survey for all yeast chaperones interactors. The data was formatted into a symmetrical matrix $A=[a_{ij}]_{5,880x5,880}$ containing all yeast open reading frames, with a_{ij} =1 if chaperone *i* and protein *j* interact, and a_{ij}=0 otherwise. A division of the network into modules was found by defining a modularity function of each bipartition of the network, as the number of edges within a community minus the expected number of edges in the community. Maximizing this function over all possible divisions using eigenspectrum analysis yields the optimal division of the network (*30*).

Evolutionary rate. Positional orthology assignments within 20 fungal proteomes were obtained from Wapinski et al. (*12*). Open reading frames lacking orthologs (282 in total) were omitted from the substrate-chaperone interaction network. Pairwise alignments of all yeast open reading frames with orthologous sequences were reconstructed with ClustalW (*31*). The frequency of amino acid substitutions was calculated from the pairwise protein alignments. Corrected amino acid substitution rate per site was calculated by PROTDIST (*32*) using the default JTT substitution matrix (*33*). Amino acid substitution rates were compared among the modules using Kruskal-Wallis test (*34*). To classify the modules into rate categories the modules were first sorted by their mean substitution rate in ascending order. Starting with the slowest module (CCT-Cct8), amino acid substitution rate of substrates in the modules were compared to the next module using the Wilcoxon test (*34*). If the substitution rate distribution between the compared modules is not significantly different, then the modules are pooled into the same rate category. The next module in ranking is compared to the pooled modules in the current category. If the substitution rate is significantly different between the compared modules, then the next module is classified into the next rate category. The same procedure was repeated for the modules sorted by their mean substitution rate in descending order. The grouping of all modules except HSP40-Sis1 and HSP70-Ssa3 was independent of the comparison order. These two modules were grouped into the medium rate category in the ascending comparison while in the descending comparison they were grouped into the fast rate category. Consequently HSP40-Sis1 and HSP70-Ssa3 modules were grouped into the medium-fast rate category.

Substrate characteristics. Amino acid usage data, functional assignment (gene ontology information), chromosomal location, frequencies of optimal codons, codon adaptation index (CAI), gravy scores (hydropathy index) and aromaticity scores were obtained from the Saccharomyces Genome Database (*35*) and the Gene Ontology database (*36*). Secondary structure of all proteins was inferred using PsiPred (*37*). For the calculation of secondary structure usage a threshold of probability>0.7 was used. Protein expression data were obtained from Ghaemmaghami et al. (*38*), who calculated the number of protein molecules from yeast under standard conditions. Protein domains were reconstructed by an RPS-BLAST (*39*) search against the database of conserved protein domains (CDD (*40*)). Protein sequence motifs were extracted from the total data by a sliding window algorithm with a window size as the motif length and a single amino acid shift using a PERL script. For the statistical analysis the natural log of protein expression was used. Proteins with no expression level information (107) or with zero expression level (1,665) were omitted from the analysis. Comparison of substrate characteristics among the modules was performed using Kruskal-Wallis test (*34*). Post-hoc comparison among the modules was performed using Tukey test (*34*). All statistical analyses were performed using MatLab© Statistical toolbox.

Correlations. Correlations between protein characteristics at the proteome level were calculated for all 3,595 proteins included in the network. The correlation at the module level was calculated from the variable means. Deviation of the correlation coefficient at the module level from the expected by random was calculated as the percentile of random correlation coefficients smaller (for negative correlation) or larger (for positive correlations) than the correlation coefficient in question. The distribution of random correlation coefficients was calculated from the data using 1,000 permutations of randomized module association of proteins.

ACKNOWLEDGMENTS

This work was supported by the German Science Foundation (DFG) and a European Research Council grant NETWORKORIGINS to W. Martin. Computational support and infrastructure was provided by the Zentrum für Informations- und Medientechnologie (ZIM) at the Heinrich-Heine-University of Duesseldorf. We are thankful to W. Martin, M. J. Lercher, and E. Shakhnovich for their help in refining the manuscript.

REFERENCES

1. Ellis RJ (1987) Proteins as molecular chaperones. *Nature* 328:378-379.

2. Young JC, Agashe VR, Siegers K, Hartl FU (2004) Pathways of chaperone-mediated protein folding in the cytosol. *Nat Rev Mol Cell Biol* 5:781–791.

3. Queitsch C, Sangster TA, Lindquist S (2002) Hsp90 as a capacitor of phenotypic variation. *Nature* 417:618–623.

4. Fares MA, Ruiz-Gonzalez MX, Moya A, Elena SF, Barrio E (2002) Endosymbiotic bacteria: groEL buffers against deleterious mutations. *Nature* 417:398.

5. Gong Y, *et al.* (2009) An atlas of chaperone-protein interactions in *Saccharomyces cerevisiae*: implications to protein folding pathways in the cell. *Mol. Syst. Biol.* 5:275.

6. Newman MEJ, Leicht EA (2007) Mixture models and exploratory analysis in networks. *Proc. Natl. Acad. Sci. U S A* 104:9564-9569.

7. Hartl FU, Hayer-Hartl M (2009) Converging concepts of protein folding in vitro and in vivo. *Nat. Struct. Mol. Biol.* 16:574-581.

8. Kampinga HH, Craig EA (2010) The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nat. Rev. Mol. Cell. Biol.* 11:579-592.

9. Rospert S, *et al.* (1993) Identification and functional analysis of chaperonin 10, the groES homolog from yeast mitochondria. *Proc. Natl. Acad. Sci. U S A* 90:10967-10971.

10. Gupta RS (1995) Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of eukaryotic cells. *Mol. Microbiol.* 15:1-11.

11. Esser C, *et al.* (2004) A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* 21:1643-1660.

12. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449:54-61.

13. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* 9:r43-74.

14. Sharp PM, Li WH (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15:1281-1295.

15. Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158:927-931.

16. Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* 13:2229-2235.

17. Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. *Nat. Rev. Genet.* 7.337-348.

18. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. *Science* 296:750-752.

19. Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. *Nature* 411:1046-1049.

20. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U S A* 102:14338-14343.

21. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341-352.

22. Warnecke T, Hurst LD (2010) GroEL dependency affects codon usage — support for a critical role of misfolding in gene evolution. *Mol. Syst. Biol.* 6:340.

23. Dobson CM (2003) Protein folding and misfolding. *Nature* 426:884-890.

24. Tuller T, *et al.* (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* 141:344-354.

25. Cannarozzi G, *et al.* (2010) A role for codon order in translation dynamics. *Cell* 141:355-367.

26. Geiler-Samerotte KA, *et al.* (2011) Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc. Natl. Acad. Sci. U S A.* 108:680-685.

27. Horwich AL, Apetri AC, W. A. Fenton WA (2009) The GroEL/GroES cis cavity as a passive anti-aggregation device. *FEBS Lett.* 583:2654-2662.

28. Kerner MJ, *et al.* (2005) Proteome-wide analysis of chaperonin-dependent protein folding in *Escherichia coli. Cell* 122:209-220.

29. Bogumil D, Dagan T (2010) Chaperonin-dependent accelerated substitution rates in prokaryotes*. Genome Biol Evol* 2:602-608.

30. Newman MEJ (2010) in Networks: An introduction. (Oxford University Press), pp 345-382.

31. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: 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.

32. Felsenstein J (2005) PHYLIP (phylogeny inference package) Version 3.6. Seattle (WA): Department of Genome Sciences, University of Washington.

33. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation rate matrices from protein sequences. *Comput. Appl. Biosci.* 8:275–282.

34. Zar JH (1999) Biostatistical analysis. (Prentice Hall, Upper Saddle River, NJ) pp 122-230.

35. Cherry JM, *et al.* (1997) Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature* 387:67-73.

36. Ashburner M, *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25:25-29.

37. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292:195-202 .

38. Ghaemmaghami S, *et al.* (2003) Global analysis of protein expression in yeast. *Nature* 425:737-741.

39. Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

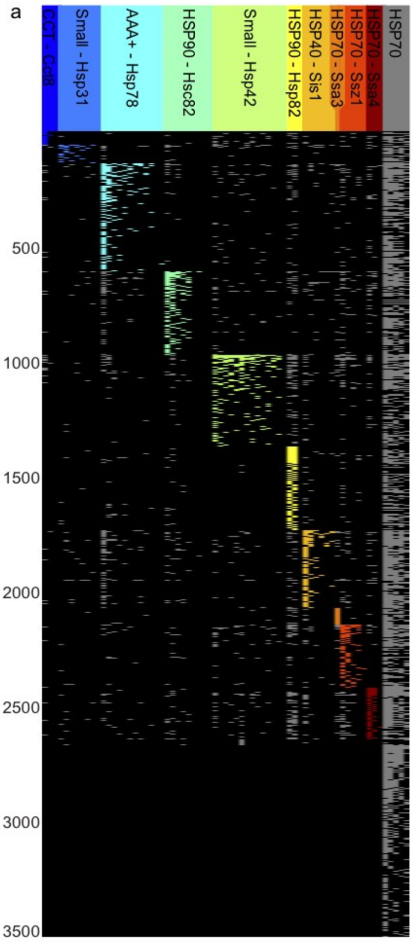
40. Marchler-Bauer A, *et al.* (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* 39:D225-229.

Figure Legends

Figure 1. The network of chaperone-substrate interactions. (a) A graphic representation of the network with chaperones on the x-axis (*i*=1..69) and substrates on the y-axis (*j*=1.. 3,595). Cells in the matrix represent a protein-protein interaction between chaperone *i* and substrate *j*. The cells are colored by the module color if both substrate and chaperone are included in the module, and in grey otherwise. Cells of non-interacting proteins are colored in black. HSP70 group includes the five ungrouped chaperones: Ssb1, Ssa1, Sse1, Ssa2, and Ssb2. (b) Modules within the chaperone-substrate network. Rate categories are coded by S (slow), M (medium), M/F (medium/fast), and F (Fast).

Figure 2. Evolutionary rates of yeast substrates in the ten modules compared to their positional ortholog in 20 fungal species. Hsp70 group includes five ungrouped chaperones: Ssb1, Ssa1, Sse1, Ssa2, and Ssb2. The lines represent rate category medians in each species.

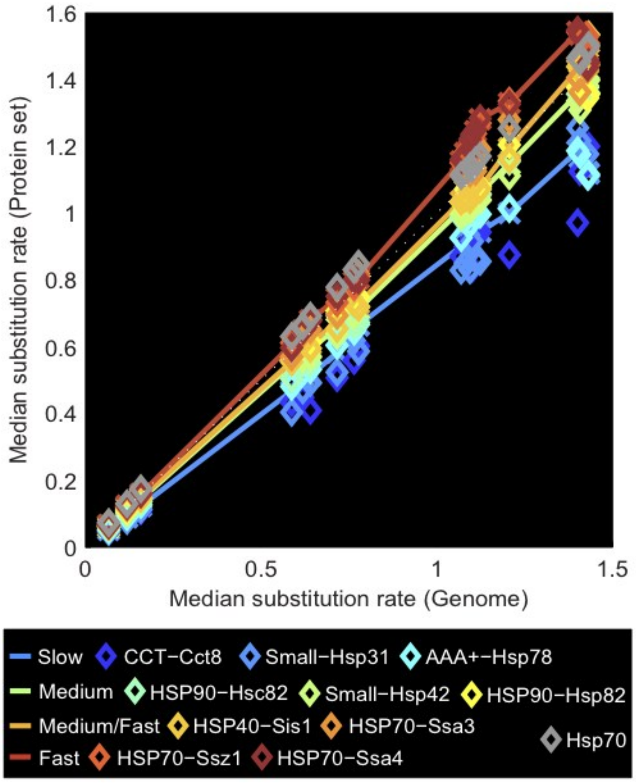
Figure 1



b

Module, No. substrates, No. Chaperones, Rate category	Chaperones
CCT-Cct8 65, 3, S	Cct8, Cwc23, Tfc7
SMALL-Hsp31 86, 8, S	Hsp31, Hsp26, Ecm10, Jjj2, Mdj1, Mcx1, Cct5, Hsp33
AAA+-Hsp78 485, 12, S	Hsp78, Pfd1, Scj1, Cct2, Ssq1, Jem1, Kar2, Tcp1, Cct7, Jid1, Hsp32, Kap114
HSP90-Hsc82 378, 9, M	Hsc82, Hsp104, Ssc1, Zuo1, Hsp60, Hsp10, Mdj2, Cdc34, Ntg2
SMALL-Hsp42 415, 14, M	Hsp42, Gim3, Djp1, Cct4, Apj1, Yke2, Sec63, Cct3, Gim4, Hsp12, Xdj1, Gim5, Jjj3, Tdh3
HSP90-Hsp82 366, 3, M	Hsp82, Ydj1, Spl2
HSP40-Sis1 347, 6, M/F	Sis1, Caj1, Cct6, Lhs1, Erj5, Hlj1
HSP70-Ssa3 76, 1, M/F	Ssa3
HSP70-Ssz1 282, 5, F	Ssz1, Swa2, Pac10, Jjj1, Jac1
HSP70-Ssa4 255, 3, F	Ssa4, Sse2, Sno4

Figure 2


1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Median substitution rate (Protein set)
0
0.5
1
1.5
Median substitution rate (Genome)
Slow
CCT−Cct8
Small−Hsp31
AAA+−Hsp78
Medium
HSP90−Hsc82
Small−Hsp42
HSP90−Hsp82
Medium/Fast
HSP40−Sis1
HSP70−Ssa3
Hsp70
Fast
HSP70−Ssz1
HSP70−Ssa4

10 Acknowledgments

First, I would like to thank Prof. Bill Martin for the opportunity to conduct my research in a very supportive and stimulating environment.

I am grateful to the many students with whom I worked over the years for the part they took/take in my research: Dr. Oliver Deusch, Dr. Nicole Grünheit, Dr. Nahal Ahmadinejad, Dr. Christian Esser, Dr. Gabriel Gellius-Dietrich, Dr. Verena Zimorski, Mayo Röttger, Annette Kürten, Christian Hurtz, Judith Ilhan, Mattis List, Ovidiu Popa, Thorsten Klösges, Shijulal Nelson-Sathi, and David Bogumil.

I thank my colleagues and collaborators for the fruitful discussions over the years: Prof. Martin Lercher, Prof. David Bryant, Prof. Russell Gray, Prof. Heiner Fangerau, Prof. Hans Geisler, Prof. Dan Graur, Prof. Tal Pupko, Prof. Rotem Sorek, Dr. Katrin Henze, Dr. Sven Gould, Dr. Einat Hazkani-Covo, Dr. Eric Bapteste, and Dr. Giddy Landan,