Directed phylogenomic networks of lateral gene transfer during prokaryotic evolution

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Ovidiu-Nicolae Popa

Kiel, 2. November 2015

Die hier vorgelegte Dissertation habe ich eigenständig, ohne unerlaubte Hilfe und unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft angefertigt. Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Teile der kumulativen Dissertation sind bereits veröffentlicht bzw. zur Veröffentlichung eingereicht. Diese sind als solche gekennzeichnet. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Kiel, 2. November 2015

Ovidiu Popa

Referentin: Prof. Dr. Tal Dagan

Koreferent: Prof. Dr. William F. Martin

Tag der mündlichen Prüfung: 18.01.2016

Zum Druck genehmigt: Kiel, den 18.01.2016

Der Dekan

Content

Abstract

Gene acquisition by lateral (or horizontal) gene transfer (LGT or HGT) is an important mechanism for natural variation among prokaryotes with a substantial impact on microbial genome evolution. This thesis presents the development of phylogenomic directed networks as a novel tool to study genome evolution by gene transfer in the prokaryotic domain. The networks comprise genomes connected by lateral gene transfer events where the edges are directed from the donor to the recipient. Research of network topology and node connectivity pattern enables the study of trends and barriers to gene transfer during microbial evolution. The first directed LGT (dLGT) network presented here was reconstructed from 657 fully sequenced prokaryotic genomes. The network topology uncovered the presence of several barriers to LGT in prokaryotes. Most of the transferred genes in the network encode for proteins of metabolic functions. These were transferred mostly between closely related species living in same ecological niche. Specific network analysis further suggests that DNA repair mechanisms can assist the integration of acquired DNA and facilitate gene transfer between distantly related donors and recipients. A second, more complex dLGT network in this study was reconstructed from 3,982 prokaryotic genomes, with a special focus on phage-mediated gene transfer via transduction. Most of the links in the network were restricted to closely related donors and recipients. A considerable number (9%) of transduction events were depicted as gene duplications. Gene duplication via mobile DNA vectors was proposed to be termed *'autology'* instead of paralogy. Network structure and connectivity shape revealed lysogenic interaction as highly species specific whereas host range for lytic interaction can be much wider. LGT by transduction is restricted by stringent phage-host specificity along with genetic barriers. This tight constraint is occasionally relaxed enabling long range LGT. The structural properties of a directed, phylogenomic networks open up fundamentally new insights into microbial gene and genome evolution.

Zusammenfassung

Der Generwerb durch lateralen (oder horizontalen) Gentransfer (LGT oder HGT) ist ein sehr wichtiger Mechanismus für die Variation innerhalb prokaryotischer Organismen. Dieser Mechanismus weist einen erheblichen Einfluss auf die bakterielle Genomevolution aus. Diese Dissertation stellt die Entwicklung eines phylogenomischen, gerichteten Netzwerkes vor, welches als ein neues Instrument für die Analyse der Genomevolution mittels lateralen Gentransfers innerhalb der prokaryotischen Domäne benutzt werden kann. Das Netzwerk beinhaltet Genome, die durch laterale Gentransferereignisse miteinander verbunden sind, sowie gerichtete Kanten, die Donor und Empfänger spezifizieren. Untersuchungen der Topologie und der Konnektivitätsart des Netzwerkes erlauben es, gewisse Richtungen und Barrieren des lateralen Gentransfers während der mikrobiellen Genomevolution zu erforschen. Das erste, in dieser Arbeit vorgestellte gerichtete Netzwerk (dLGT) wurde aus 657 vollständig sequenzierten Genomen rekonstruiert. Die Auswertung dieses Netzwerkes offenbarte verschiedene Barrieren in Prokaryoten, denen der LGT-Mechanismus unterliegt. Die Mehrheit der detektierten, lateral transferierten Gene kodieren für Proteine, die in metabolischen Prozessen involviert sind. Diese werden vermehrt zwischen sehr nah verwandten Spezies übertragen, welche oft die gleiche ökologische Nische besiedeln. DNA-Reparaturmechanismen, die auch eine Integrationsfunktion der erworbenen DNA aufweisen, können die bestehenden Barrieren aufheben und die Übertragung des genetischen Materials zwischen entfernt verwandten Spezies ermöglichen. Ein zweites, komplexeres phylogenomisches Netzwerk, das in dieser Arbeit vorgestellt wird, wurde aus 3982 voll- und teilsequenzierten Genomen rekonstruiert, um den lateralen Gentransfer als Folge eines Transduktionsereignisses besser zu untersuchen. Die Mehrheit der Verknüpfungen im Netzwerk besteht zwischen Genomen, in denen Donor und Empfänger nah verwandt sind. Eine beachtliche Zahl (9%) von Transduktionsereignissen sind Genduplikationen. Hierfür wurde

vorgeschlagen, diese Art der Duplikation ‚*autology'* anstatt ‚*paralogy'* zu bezeichnen. Das Vernetzungsmuster des Netzwerkes zeigte einen deutlichen Unterschied zwischen den lysogenen Interaktionen, die stark speziesspezifisch sind, und den lytischen Interaktionen, die einen größeren Umfang von Speziesarten für Phagen-Infektionen aufweisen. Die strukturelle Eigenschaft des gerichteten, phylogenomischen Netzwerkes eröffnet fundamentale, neue Einblicke in die mikrobielle Gen- und Genomevolution.

.

Introduction

Lateral gene transfer (LGT) plays a central role in prokaryotic genome evolution (1, 2), affecting almost all genes with only few exceptions (1-4). Currently recognized mechanisms enabling lateral gene transfer among microbial species include transformation, conjugation and transduction. Additional, more taxa-specific mechanisms include gene transfer agents (GTAs), cytoplasmic bridges, nanotubes and outer membrane vesicles (OMVs). Modelling prokaryotic genome evolution requires the incorporation of LGT events into a phylogenetic context. Standard approaches based on phylogenetic trees only are therefor insufficient by lacking the ability to combine the LGT information with the vertical signal. Networks are able to visualize the complexity of prokaryotic genome evolution in a more sophisticated way. This thesis especially focuses on directed networks used to analyse the impact of LGT to prokaryotic genome evolution on several levels. The results reveal strong genomic barriers for LGT and mechanisms that can relax them.

Lateral gene transfer mechanisms

Griffith first demonstrated the concept of **transformation** in 1928. Without knowing the exact mechanisms behind it, he was able to transform an attenuated and non-encapsulated *Pneumococcus* (type R) into a fully encapsulated and virulent *Pneumococcus* cells (type S) (5). Sixteen years later in 1944, Avery *et.* al. (6) could demonstrate that the transforming material was composed of deoxyribonucleotide acids. From this result they concluded that DNA was the transforming material of *Pneumococcus* type-R.

Transformation involves the uptake of naked DNA from the environment. The uptake of raw DNA in transformation is enabled during a competence state that involves 20-50 proteins, including the type IV pillus and type II secretion system proteins (7). Experiments in DNA uptake of fluorescence labeled DNA in *Helicobacter*

pylori revealed that the length of imported DNA fragments is on average 10Kb, and that these are imported at a mean velocity of 1,230bp per second (8).

The process of gene recombination as a result of direct DNA transfer between microbial cells, termed **conjugation**, was discovered in 1946 by Lederberg and Tatum(9, 10). This DNA transfer mechanism is mediated by a proteinaceous cell-to-cell junction, forming a tunnel that connects the donor and recipient cells through which the DNA is transferred (7). Conjugation typically comprises conjugative plasmids that encode the proteins required for the conjugation tunnel formation and DNA replication. These mobile circular DNA elements are transferred through a duplication and insertion process where at the end of the transfer both donor and recipient harbor an identical copy of the plasmid (11). Additional DNA elements that are transferred by conjugation are integrative conjugative elements (ICEs), which are transferred by an excision and insertion procedure (12). These are mobile genetic elements whose genetic repertoire includes the genes required for conjugation, excision from the donor genome and integration into the recipient genome (13). Genetic elements transferred by conjugation may be integrated into the recipient chromosomes.

The success rate of gene acquisition by conjugative plasmids was estimated in *Escherichia coli* under laboratory conditions (14). In an experiment with yellow fluorescence protein (YFP) labeled DNA, it could be demonstrated, that 96% of the recipients recombined the acquired plasmid into the chromosome and inherited it to the next generation. (14). The percolation of an acquired DNA within the population can be extremely fast in species where the cells are arranged in chains such as *Bacillus subtilis*. Tracking the spread of a green fluorescence protein (GFP) labeled integrative and conjugative element (ICE) under the microscope showed that in 43 (81%) out of 53 cases a recipient cell turned into a donor and transconjugated the ICE to the next cell in line, often within 30 minutes (15).

An archaeal mating system where DNA is transferred via **cytoplasmic bridges** was documented in the haloarchaeon *Haloferax volcanii* by Mevarech and

colleagues in the mid 1980's (16, 17) (Figure **1a**). Early experiments revealed that a plasmid encoding for selectable properties could be transferred between *H. volcanii* cells. Like in conjugation, the transfer required that the cells be viable and in physical contact, yet unlike conjugation, the DNA transfer was bi-directional where the cells could function either as a donor or a recipient (16). Scanning electron microscopy revealed intercellular bridges between the cells where the membrane and envelope of the connected cells fuse to form what seems to be a cytoplasmic continuity. The bridge length was estimated to reach up to 2 mm and their diameter was up to 0.1 mm (17).

A similar transfer mechanism was discovered in the eubacterial domain and was named **nanotubes** (18) (Figure 1b). These are tubular protrusions composed of membrane components that can bridge between neighboring cells. The nanotubular structure facilitates the transfer of cytoplasmic material resulting in exchange of nutrients and distribution of functions within bacterial communities (19). Nanotubes are between 30 and 130 nm wide and up to 1 μm long. The tube dimension is correlated with the distance between the connected cells. Proximal cells are commonly connected by several small nanotubes, while thicker tubes connect distant cells. The rate of transfer via the nanotubes correlates negatively with the size of the transferred substance. Cellular interconnection mediated by nanotubes is not species specific. However, morphology and diameter of the tubes seem to depend on characteristics of the connected cells (18).

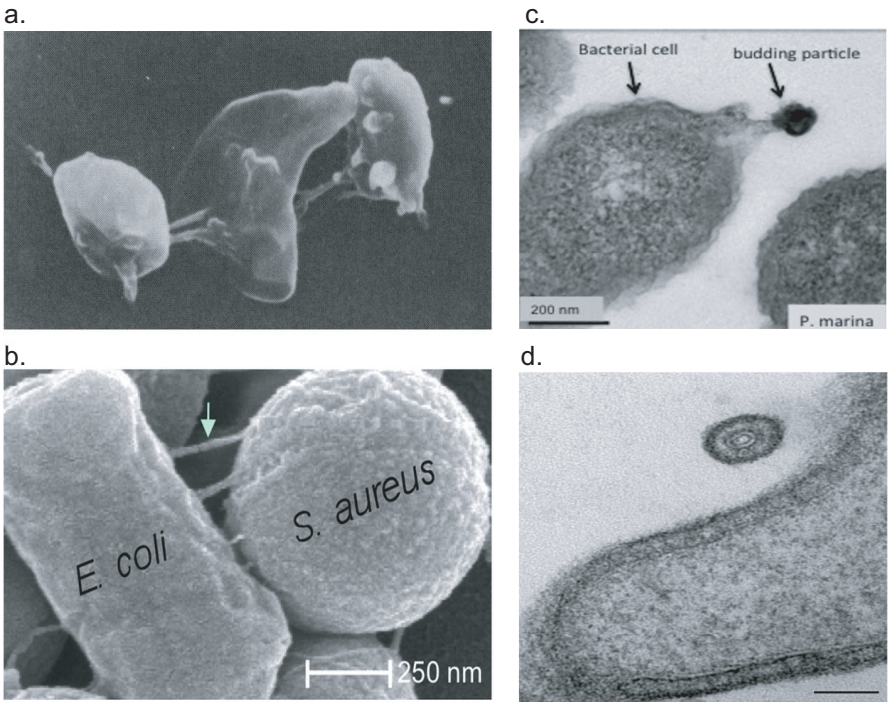


Figure 1. **Intercellular DNA transfer**. (a) Intercellular bridges connecting *Halobacterium volcani* cells (adopted from (16)). (b) Nanotubes connecting *Staphylococcus aureus* (PY79) and *Escherichia coli* (MG1655) cells (adopted from (18)) (c) A *Pseudoalteromonas marina* cell in the process of releasing OMV by budding (adopted from (20)). (d) an OMV released by the archaebacterium *Thermococcus gammatolerans*

Transduction is DNA acquisition following a phage infection (21). This DNA transfer mechanism was discovered by Zinder and Lederberg during a study of recombination in *Salmonella* strains (21). In their study, they have demonstrated that no physical contact was required for the recombination using a U-shaped tube in which the donor and recipient were cultured in common medium but at opposite ends that were separated by a filter dense enough to block the passage of cells. That approach revealed small particles that were suspected as the recombination agents, later identified as bacteriophage PLT-22 (see (22) for a detailed perspective).

Temperate (or lysogenic) phages multiply via the lysogenic cycle, which is established by an integration of the phage genome into the host chromosomes, creating a prophage within the host genome. The phage typically remains dormant within the host and is replicated with the host until the lytic cycle is induced. In the lytic cycle new phages are produced using the host metabolism and are released during the

host cell lysis. The excision of phage DNA from the host genome and the production of phages may be accompanied by packing of host DNA into the phages, which can then transfer it to the next host (23). Specialized transduction occurs when the phage integrases cleave, in addition to the prophage, bacterial genes that are encoded at the prophage flanking regions. These are packed with the phage DNA into the phages. Generalized transduction occurs when random bacterial DNA is packed into the phages (24, 25).

The frequency of gene acquisition via transduction within pure cultures of marine bacteria has been estimated to range between $1.33x10^{-7}$ and $5.13x10^{-9}$ transduced cells per plaque forming unit (PFU). Gene acquisition was observed also in colonies that did not form a plaque; these however occurred at a significantly lower frequency ranging between $6.8x10^{-10}$ and $2.6x10^{-11}$ transductans per colony forming unit (CFU). Applying the same approach on a mixed marine microbial population yielded similar ranges between $1.58x10^{-8}$ and $3.7x10^{-8}$ transduced cells per PFU (26). Nevertheless applying more sensitive methods to detect transferred marker gene sequences yielded frequencies that are up to five orders of magnitude higher (27).

Gene transfer agents (GTAs) are phage-like DNA-vehicles that are produced by donor cells and released to the environment (28-31) (Figure 1d). Marrs described 1974, small phage-like particles that were released into the medium by *Rhodopseudomonas capsulatus* strains and could mediate the transfer of antibiotic resistance genes to a non-resistant *R. capsulatus* strains. As the newly discovered particles were significantly smaller than any known phage and did not induce plaque formation or cell lysis they were recognized as a novel gene transfer mechanism (28). The proteins required for GTA synthesis are encoded within an operon of 15-17 genes. Several genes resemble typical phage genes hence GTAs might be related to phages (30). The DNA stored in GTAs is imported into the recipient in a generalized transduction process mediated by the cellular *RecA* recombination system (32). The mechanism of DNA packing and capsule release

from the cell is still unknown. GTA systems have been documented also in the Deltaproteobacteria *Desulfovibrio desulfuricans* (33) and the Euryarchaeota *Methanococcus voltae* (34).

Many microbial species secret **outer membrane vesicles** (OMVs) that are used for various extracellular functions as well as intercellular communication (35) (Figure 1c,d). Earlier reports implying a role of OMVs in DNA transfer were published by Kahn and colleagues who focused on the study of DNA transfer mechanisms in *Haemophilus influenzae* (36) (37). Earlier studies of *H. influenza*e showed that this bacterium is naturally competent and is able of DNA acquisition (38), however is was not clear how the transferred DNA is protected from restriction enzymes during the transfer from donors to recipients. By using electron microscopy Kahn et al. (36) discovered small membrane vesicles that were attached to the recipient cell membrane. Later studies revealed that double stranded DNA is transferred within the membrane vesicles that protect it during the transfer (37). These OMVs were named *Transformasomes* (37).

The transformation success rate of antibiotic resistance genes encapsulated in OMVs between various strains of the gammaproteobacterium *Acinetobacter baumanii* has been shown to be close to 100 and is strictly dependent on the integrity of the OMVs (39). The DNA capacity of OMVs has been estimated as ~20Kbp in the marine alphaproteobacterium *Ahrensia Kielensis* (20) and ~600bp in the arctic gammaproteobacterium *Shewanella vesiculosa* $M7^T$ (40). DNA bearing OMVs have been reported also in the archaeal domain. The hyperthermophilic archaeon *Thermococcus kodakaraensis* produces OMVs that can carry a shuttle plasmid and transform plasmid-lacking cells (41) (Figure 1b).

Phylogenomic networks of microbial genome evolution

The evolutionary history of a species is most commonly depicted as a bifurcating phylogenetic tree comprising nodes and branches. The external nodes in the tree

correspond to contemporary species while the internal nodes correspond to ancestral species. The branches represent vertical inheritance linking ancestors with their descendants. Since the early 2000s, the amount of fully sequenced genomes is steadily increasing with 2,789 fully and closed prokaryotic genomes available in 2015 (NCBI DB: http://www.ncbi.nlm.nih.gov/). Investigating whole genome data to reconstruct evolutionary history enabled the practice of **phylogenomics**, that is, the study of phylogenetic relationships at the whole genome level (42). The evolutionary reconstruction of gene phylogenies from many genomes at once allows a more accurate reconstruction of evolutionary events such as gene loss, gene gain, and gene duplication (43). However, the widespread occurrence of LGT means that a tree model that takes only vertical inheritance into account fits only a fraction of the bacterial genomic repertoire (44).

A **network** (or a graph) is a mathematical model of pairwise relations among entities. The entities (vertices or nodes) in the network are linked by edges representing the connections or interactions between these entities (Figure 2a). Vertex connectivity (or the degree of a vertex) is the number of vertices connected to the vertex. In a weighted network the edges can also have a certain weight that signifies the strength of the connection between the vertices. In directed networks the edges are oriented from one vertex to another (Figure 2b) and can be either unweighted or weighted. Vertex connectivity in a directed network is calculated depending on the edge direction. The '*out*' and '*in*' degrees of any given vertex are defined as the number of edges that are directed from or into the vertex respectively (45) (Figure 2b). Vertex connectivity in directed networks is calculated separately for outgoing and incoming edges.

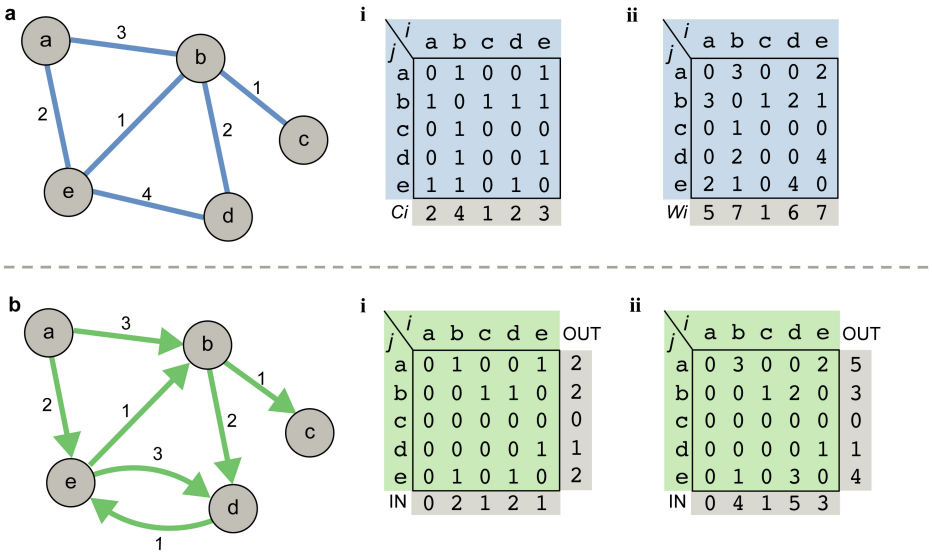


Figure 2. An introduction to networks. (a) A network composed of vertices (circles) and edges (lines). (i) An unweighted network of n vertices can be fully defined by a matrix, A = [a_{ij}]n*n, with a_{ij} = 1 if an edge is connecting between vertex *i* and vertex *j*, and a_{ij} = 0 otherwise. Vertex centrality (C_i) is calculated as the sum of vertices linked to the vertex. (ii) A weighted matrix representation of the network. Cells of connected vertices *i* and *j* contain the edge weight linking the vertices. Vertex connectivity (w_i) is the sum of edge weights of the edges connected to the vertex. **(b)** A directed network comprising vertices and directed edges. (i) In the matrix representation of an unweighted directed network of n vertices, *aij* = 1 if a directed edge is pointing from vertex *i* to vertex *j,* and a_{ij} = 1 if a directed edge is pointing from vertex *j* to vertex *i*. Vertex *‚in'* degree is the sum of vertices connected to the vertex. Vertex *‚out'* degree is the number of vertices to which the vertex is connected. (ii) A matrix representation of a weighted directed network. Cells of edges directed from vertex *i* to vertex *j* contain the edge weight. Vertex *‚in'* degree is the sum of edges connected to the vertex. Vertex *‚out'* degree is the sum of edges connecting the vertex to other vertices. (adopted with modification from (46)).

Networks are commonly used in phylogenetic research for the reconstruction of evolutionary processes that have a reticulate character in nature including species hybridization, gene recombination, genome fusions, and LGT (47), (48). Network applications can be also used for tree-like gene phylogenies (i.e., genes that did not evolve by LGT) in order to analyse conflicting phylogenetic signals originated from biases in the data or model misspecification. Similarly to phylogenetic trees, phylogenetic networks can be reconstructed from various data types including molecular sequences, evolutionary distances, presence/absence data, and trees (49).

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 (46).

Phylogenomic LGT networks have been reconstructed from LGT events detected in gene phylogenies as well (50, 51). The network reconstruction process requires the distinction between vertical inheritance and lateral gene transfer events. Vertical inheritance is mostly inferred from marker genes, for example 16S rRNA and 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 (50, 51). Other network reconstruction methods relies on the LGT information that is naturally incorporated in the data, like for example bacterial genes within a prophage are very likely to have been acquired through a transduction process (Chapter III).

LGT inference methods that include the identification of the donor and recipient in the gene transfer event enable the reconstruction of directed phylogenomic networks. The first publication in this thesis (Chapter I) shows the reconstruction of a directed phylogenomic network. Recent lateral gene transfer events were therefore inferred from genes having an aberrant nucleotide pattern and a protein family tree with disagreement to the reference tree. Genomic data from 657 fully sequenced prokaryotic genomes was used to reconstruct 32,027 recent LGT events for which a donor gene could be specified (51) (Chapter I). This information was summarized into a directed network of recent lateral gene transfer events (dLGT). In comparison to an undirected phylogenomic network, the dLGT network allows studying the impact of lateral gene transfer to microbial genome evolution regarding the information about the donor. The vertices in this network are contemporary and ancestral microbial species. Edges correspond to LGT

events between the species giving the direction from donor to recipient. Edge weights correspond to the number of genes that were transferred from the donor to the recipient. The structure of the dLGT network can be coupled with several cellular characteristics like species ecology or cell pathogenicity as well as genomic attributes like GC content or coding sequence similarity (51, 52). The dLGT network property revealed, members of highly connected clusters in the network are species with high genomic similarity sharing often the same ecological niche (52) (chapter II). The results further suggests, that DNA repair mechanisms can assist the integration of acquired genes from more distant related donors and therefore moderate the otherwise strict similarity barrier (51).

A more complex version of the dLGT network reconstructed from a total of 3,982 finished and draft prokaryotic genomes in order to study the characteristics of LGT by transduction, combines donor – recipient and phage information into one structure (chapter III). The type of this gene acquisition mechanism requires the definition of two sorts of connected vertices in the network: bacteria and phage nodes. Directed edges connecting the bacteria nodes with the phage nodes models the gene uptake process from the donor by the phage followed by a lytic event. Directed edges connecting the phage node with the bacteria node represent the gene acquisition by the recipient as part of a lysogenic interaction with the phage. This network approach enabled the reconstruction of 17,158 transduction events between 2,573 bacteria nodes and 4,650 phage nodes. Structural properties of the transduction dLGT network revealed LGT by transduction to be mostly restricted to closely related donors and recipients. Further analyses depict a substantial number of gene duplications mediated by transduction in a process, which was proposed to be termed '*autology*'. A comparison of donor recipient genome similarity and ecological properties reveals a strong genetic barrier for transduction while ecological factors are more secondary. The results demonstrate that LGT by transduction occurs within a context of phage-host specificity along with

genetic barriers, which sometimes are more relaxed and therefore facilitating long-range LGT events.

Directed phylogenomic networks encompass the potential to study the reticulate characteristics of microbial genome evolution in a more precise way.

.....

Parts of this introductory chapter have been recently published in a book chapter: Dagan T, **Popa O**, Klösges T, Landan G. 2014. Phylogenomic networks of microbial genome evolution in Manual of Environmental Microbiology, 4th Ed. ASM Press, in press.

References

1. Lawrence JG, Ochman H (1998) Molecular archaeology of the Escherichia coli genome. *Proc Natl Acad Sci USA* 95(16):9413–9417.

2. Martin W (1999) Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays* 21(2):99–104.

3. 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(3):870–875.

4. Sorek R, et al. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318(5855):1449–1452.

5. Griffith F (2009) The Significance of Pneumococcal Types. *J Hyg* 27(02):113.

6. Avery OT (1944) Studies on the chemical nature of the substance inducing transformation of Pneumococcal types: Induction of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus type III. *Journal of Experimental Medicine* 79(2):137–158.

7. Thomas CM, Nielsen KM (2005) Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nat Rev Micro* 3(9):711–721.

8. Stingl K, Müller S, Scheidgen-Kleyboldt G, Clausen M, Maier B (2010) Composite system mediates two-step DNA uptake into Helicobacter pylori. *Proceedings of the National Academy of Sciences* 107(3):1184–1189.

9. Lederberg J, Tatum EL (1946) Gene Recombination in Escherichia-Coli. *Nature* 158(4016):558–558.

10. Tatum EL, Lederberg J (1947) Gene Recombination in the Bacterium Escherichia-Coli. *Journal of Bacteriology* 53(6):673–684.

11. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, la Cruz de F (2010) Mobility of Plasmids. *Microbiology and Molecular Biology Reviews* 74(3):434–452.

12. Salyers AA, Shoemaker NB, Stevens AM, Li LY (1995) Conjugative transposons: an unusual and diverse set of integrated gene transfer elements. *Microbiol Rev* 59(4):579–590.

13. Wozniak RAF, Waldor MK (2010) Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat Rev Micro* 8(8):552–563.

14. Babic A, Lindner AB, Vulic M, Stewart EJ, Radman M (2008) Direct Visualization of Horizontal Gene Transfer. *Science* 319(5869):1533–1536.

15. Babic A, Berkmen MB, Lee CA, Grossman AD (2011) Efficient Gene Transfer in Bacterial Cell Chains. *mBio* 2(2):e00027–11–e00027–11.

16. Garneau JE, Dupuis MÈ, Villion M, Romero DA (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*. doi:10.1038/nature09523.

17. Rosenshine I, Tchelet R, Mevarech M (1989) The mechanism of DNA transfer in the mating system of an archaebacterium. *Science*. doi:10.1126/science.2818746.

18. Dubey GP, Ben-Yehuda S (2011) Intercellular Nanotubes Mediate Bacterial Communication. *Cell* 144(4):590–600.

19. Pande S, et al. (2015) Metabolic cross-feeding via intercellular nanotubes among bacteria. *Nature Communications* 6:6238.

20. Hagemann S, et al. (2014) DNA-bearing membrane vesicles produced by Ahrensia kielensis and Pseudoalteromonas marina. *J Basic Microbiol* 54(10):1062–1072.

21. Zinder ND, Lederberg J (1952) Genetic exchange in Salmonella. *Journal of Bacteriology*.

22. Zinder ND (1992) Anecdotal, Historical and Critical Commentaries on Genetics. *Genetics* 134:291–294.

23. Weinbauer MG (2004) Ecology of prokaryotic viruses. *FEMS Microbiology Reviews* 28(2):127–181.

24. Burke J, Schneider D, Westpheling J (2001) Generalized transduction in Streptomyces coelicolor. *Proc Natl Acad Sci USA* 98(11):6289–6294.

25. Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann M-L, Brüssow H (2003) Phage as agents of lateral gene transfer. *Current Opinion in Microbiology* 6(4):417–424.

26. Jiang SC, Paul JH (1998) Gene transfer by transduction in the marine environment. *Applied and Environmental Microbiology* 64(8):2780–2787.

27. Kenzaka T, Tani K, Nasu M (2010) High-frequency phage-mediated gene transfer in freshwater environments determined at single-cell level. *The ISME Journal* 4(5):648–659.

28. Marrs B (1974) Genetic recombination in Rhodopseudomonas capsulata. *Proc Natl Acad Sci USA* 71(3):971–973.

29. Solioz M, Yen HC, Marris B (1975) Release and uptake of gene transfer agent by Rhodopseudomonas capsulata. *Journal of Bacteriology* 123(2):651–657.

30. Lang AS, Beatty JT (2000) Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of Rhodobacter capsulatus. *Proc Natl Acad Sci USA* 97(2):859–864.

31. Lang AS, Zhaxybayeva O, Beatty JT (2012) Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Micro* 10(7):472–482.

32. Genthner FJ, Wall JD (1984) Isolation of a recombination-deficient mutant of Rhodopseudomonas capsulata. *Journal of Bacteriology* 160(3):971–975.

33. Rapp BJ, Wall JD (1987) Genetic transfer in Desulfovibrio desulfuricans. *Proc Natl Acad Sci USA* 84(24):9128–9130.

34. Bertani G (1999) Transduction-Like Gene Transfer in the Methanogen

Methanococcus voltae. *Journal of Bacteriology* 181(10):2992-3002.

35. Berleman J, Auer M (2013) The role of bacterial outer membrane vesicles for intra- and interspecies delivery. *Environmental Microbiology* 15(2):347–354.

36. Kahn ME, Maul G, Goodgal SH (1982) Possible mechanism for donor DNA binding and transport in Haemophilus. *Proc Natl Acad Sci USA* 79(20):6370–6374.

37. Kahn ME, Barany F, Smith HO (1983) Transformasomes: specialized membranous structures that protect DNA during Haemophilus transformation. *Proc Natl Acad Sci USA* 80(22):6927–6931.

38. Herriott RM, Meyer EM, Vogt M (1970) Defined Nongrowth Media for Stage-Ii Development of Competence in Haemophilus-Influenzae. *Journal of Bacteriology* 101(2):517–&.

39. Rumbo C, et al. (2011) Horizontal Transfer of the OXA-24 Carbapenemase Gene via Outer Membrane Vesicles: a New Mechanism of Dissemination of Carbapenem Resistance Genes in Acinetobacter baumannii. *Antimicrobial Agents and Chemotherapy* 55(7):3084–3090.

40. Pérez-Cruz C, et al. (2013) New Type of Outer Membrane Vesicle Produced by the Gram-Negative Bacterium Shewanella vesiculosa M7T: Implications for DNA Content. *Applied and Environmental Microbiology* 79(6):1874–1881.

41. Gaudin M, et al. (2013) Hyperthermophilic archaea produce membrane vesicles that can transfer DNA. *Environmental Microbiology Reports* 5(1):109–116.

42. Eisen JA (1998) Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. *Genome Research* 8(3):163–167.

43. Eisen JA (2003) Phylogenomics: Intersection of Evolution and Genomics. *Science* 300(5626):1706–1707.

44. Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution* 19(12):2226–2238.

45. Palla G, Barabási A-L, Vicsek T (2007) Quantifying social group evolution. *Nature* 446(7136):664–667.

46. Dagan T (2011) Phylogenomic networks. *Trends in Microbiology* 19(10):483–491.

47. Huson DH, Scornavacca C (2011) A Survey of Combinatorial Methods for Phylogenetic Networks. *Genome Biology and Evolution* 3(0):23–35.

48. Huson DH, Kloepper TH (2005) Computing recombination networks from binary sequences. *Bioinformatics* 21 Suppl 2:ii159–65.

49. Huson DH, Rupp R, Scornavacca C (2010) *Phylogenetic Networks: Concepts, Algorithms and Applications - Daniel H. Huson, Regula Rupp, Celine Scornavacca - Google Books*.

50. Beiko RG, Harlow TJ, Ragan MA (2005) Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA* 102(40):14332–14337.

51. Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Research* 21(4):599–609.

52. Popa O, Dagan T (2011) Trends and barriers to lateral gene transfer in prokaryotes. *Current Opinion in Microbiology*:1–9.

Overview of publications / manuscripts

Chapter I

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

(Own contribution**:** designed the research experiments, performed the data analysis and wrote the manuscript)

Chapter II

Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Current Opinion in Microbiology* 14:615–623.

(Own contribution**:** performed the analysis of barriers to LGT and wrote the review)

Chapter III

Popa O, Landan G, Dagan T. (2015) Phylogenomic transduction networks reveal genetic barriers to phage-mediated lateral gene transfer during microbial evolution*, Submitted* (see confirmation in the next page).

(Own contribution**:** designed the research experiments, performed the data analysis and wrote the manuscript)

Submission confirmation:

Von: journalstaff@pnascentral.org
Betreff: Receipt of New PNAS MS#2015-19973
Datum: 8. Oktober 2015 23:01
An: opopa@ifam.uni-kiel.de

October 8, 2015

Title: "Phylogenomic transduction networks reveal genetic barriers to phage-mediated lateral gene transfer during microbial evolution"
Tracking #: 2015-19973
Author(s):
Ovidiu Popa (Christian-Albrechts-Universität Kiel)
Giddy Landan (Heinrich-Heine Universität Düsseldorf)
Tal Dagan (Christian-Albrechts-Universität Kiel)

Dear Dr. Popa,

"Phylogenomic transduction networks reveal genetic barriers to phage-mediated lateral gene transfer during microbial evolution," for which you participated as an author, was submitted by Prof. Dagan and received in our office on October 8, 2015. The manuscript has been assigned tracking number 2015-19973.

You may check on the status of your manuscript at any time by clicking the link below and selecting the "Check Status" link. Please also check your name and institutional affiliation in the list at the beginning of this email. If the manuscript is accepted for publication, **this is how your name will appear on the published article**. To make any corrections to your contact and affiliation information, please click Manuscript Home to get to your desktop and then click the Modify Profile link to get to your profile page. We encourage you to make these corrections as soon as possible to prevent any possible publication errors. If you have any questions or need help, please contact our office.

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: 2015-10-08

Thank you for submitting to PNAS.

Sincerely yours,

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

Chapter I

Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes

Research

Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes

Ovidiu Popa,[1] Einat Hazkani-Covo,[2] Giddy Landan,[3] William Martin,[1] and Tal Dagan[1,4]

[1]*Institute of Botany III, Heinrich-Heine University Düsseldorf, Düsseldorf 40225, Germany;* [2]*Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, North Carolina 27705, USA;* [3]*Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204-5001, USA*

Lateral gene transfer (LGT) plays a major role in prokaryote evolution with only a few genes that are resistant to it; yet the nature and magnitude of barriers to lateral transfer are still debated. Here, we implement directed networks to investigate donor–recipient events of recent lateral gene transfer among 657 sequenced prokaryote genomes. For 2,129,548 genes investigated, we detected 446,854 recent lateral gene transfer events through nucleotide pattern analysis. Among these, donor–recipient relationships could be specified through phylogenetic reconstruction for 7% of the pairs, yielding 32,028 polarized recent gene acquisition events, which constitute the edges of our directed networks. We find that the frequency of recent LGT is linearly correlated both with genome sequence similarity and with proteome similarity of donor–recipient pairs. Genome sequence similarity accounts for 25% of the variation in gene-transfer frequency, with proteome similarity adding only 1% to the variability explained. The range of donor–recipient GC content similarity within the network is extremely narrow, with 86% of the LGTs occurring between donor–recipient pairs having ≤5% difference in GC content. Hence, genome sequence similarity and GC content similarity are strong barriers to LGT in prokaryotes. But they are not insurmountable, as we detected 1530 recent transfers between distantly related genomes. The directed network revealed that recipient genomes of distant transfers encode proteins of nonhomologous end-joining (NHEJ; a DNA repair mechanism) far more frequently than the recipient lacking that mechanism. This implicates NHEJ in genes spread across distantly related prokaryotes through bypassing the donor–recipient sequence similarity barrier.

[Supplemental material is available for this article.]

In prokaryote genomes, genes come to reside in the DNA via clonal replication, lateral gene transfer (LGT), and combinations thereof (Milkman and Bridges 1990). Genomic studies leave no doubt that LGT plays a qualitatively and quantitatively substantial role in prokaryote genome evolution (Doolittle 1999; Ochman et al. 2000), with virtually all genes affected by it and only a few genes, if any, that are genuinely resistant to it (Sorek et al. 2007). The impact of LGT on our understanding of the network-like—as opposed to the tree-like—nature of microbial evolution is far-reaching, as is its impact on human health via pathogenicity islands (Groisman and Ochman 1996).

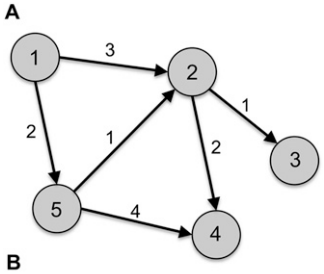
The temporal process of lateral gene acquisition can be divided into three stages (Ochman et al. 2000; Thomas and Nielsen 2005): DNA import into the cytoplasm, integration of the acquired DNA into the genome, and adaptive/selective processes acting within the genome that influence clonal inheritance to subsequent generations (Perez and Groisman 2009). Prokaryotes rapidly delete nonfunctional or otherwise unneeded DNA from their genomes (Moran 2002), such that the fixation or loss of acquired DNA within the genome is highly dependent on its utility to the recipient under selectable environmental conditions. The nature of the enzymatic mechanisms of DNA integration into the genome following the import into the cytoplasm usually depends on the mechanism of DNA transfer, of which four main types are distinguished: transformation (Chen and Dubnau 2004), transduction (Thomas and Nielsen 2005), conjugation (Chen et al. 2005), and gene transfer agents (Lang and Beatty 2007).

In order to be expressed, acquired genes either have to be inserted near, or acquired with a recognized promotor. Genes that are inserted within existing operons (Davids and Zhang 2008) or have a promotor of similar GC content as the recipient genomes (Sorek et al. 2007) have a higher probability to become fixed within the recipient, notwithstanding codon bias and amelioration (Ochman et al. 2000; Ragan et al. 2006). LGT generates genealogies among genomes with unidirectional donor–recipient relationships, corresponding to directed networks (Barabási et al. 2000).

A directed network is a graphical representation of a set of entities, or vertices, linked by edges that represent the connections or interactions between these entities. A directed network of N vertices can be fully defined by a matrix, $A = [a_{ij}]_{N*N}$, with $a_{ij} \neq 0$ if a directed edge is pointing from node i to node j, and $a_{ji} \neq 0$ if a directed edge is pointing from node j to node i. 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 (Fig. 1; Palla et al. 2005, 2007; Leicht and Newman 2008; Foster et al. 2010). In the case of LGT and genomes, the edge weight a_{ij} counts the number of genes transferred from genome i to genome j, and the OUT and IN degrees correspond to the number of connecting donors and recipients per genome.

Directed networks are still quite rare in the literature because they demand specific information about the polarized nature of connections (edges) between entities (vertices), for example, who-to-whom telephone calls (Palla et al. 2007), internet browsing

[4]**Corresponding author.**
E-mail tal.dagan@uni-duesseldorf.de; fax 49-211-811-3554.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.115592.110. Freely available online through the *Genome Research* Open Access option.



B

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	

C

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	

Figure 1. (*A*) A directed network. The circles represent nodes in the network. Arrows represent directed edges connecting between nodes. Edge weights are denoted by Arabic numerals attached to the edge. (*B*) A binary matrix representation of the directed network. If there exists a directed edge from node *i* to node *j* in the matrix, then cell $a_{ij} = 1$. Otherwise, $a_{ij} = 0$. The number of ingoing edges (IN degree) of each node is defined as the sum of the corresponding column. The number of outgoing edges (OUT degree) of each node is the sum of the corresponding row. (*C*) A weighted matrix representation of the directed network. Cells in the matrix correspond to the edge weight of edges connecting between nodes *i* and *j*. The column sums are the total edge weight of ingoing edges. The row sums are the total edge weight of outgoing edges.

paths (Barabási et al. 2000), metabolic pathways (Jeong et al. 2000), or microRNA targeting schemes (Tsang et al. 2010). In the case of prokaryote genome sequence data, the LGT donor–recipient relationships are not known a priori, but they can be estimated for recently acquired DNA sequences through analyses of codon bias, GC content, and nucleotide pattern frequencies (Garcia-Vallve et al. 2000; Nakamura et al. 2004).

Here we report the use of directed networks of recent acquired genes to study LGT-mediated prokaryote genome evolution. The directed networks allow us to formulate and test a wide range of hypotheses regarding LGT patterns and mechanisms operating in nature.

Results

A directed network of recent LGT

To obtain a matrix of recent LGTs, we first scanned the completely sequenced genomes of 657 prokaryote species encoding 2,129,548 proteins for recently acquired genes. We used the criterion of genic GC content that deviated from the genome as a whole (Ochman et al. 2000). This identified 446,854 protein-coding genes (21% of the total) as recently acquired, corresponding to 20 ± 9% recent gene acquisitions per genome, whereby the number of acquired genes per genome correlates positively with genome size ($r = 0.93$, $P \ll 0.01$, using Spearman test). This estimate for the fraction of foreign genes per genome is consistent with other studies using similar methods (Garcia-Vallve et al. 2000; Nakamura et al. 2004).

Within this set of 446,854 acquisitions, we then sought potential donors. While identifying recent acquisitions is relatively straightforward, determining possible donors is far more difficult. Our process of donor identification involves a serial application of GC content, sequence similarity, and individual gene tree comparisons with the goal of finding the genome within our sample or ancestral node within the respective gene tree that would correspond to the most likely donor within the genome sample (for details, see the Methods section). This does not, of course, identify the exact biological donor, which is unlikely to be included in our small sample, but identifies the most likely donor among the genomes available. The method is conservative and specifies a donor for 32,028 (7%) of the recently acquired genes. In those cases, we have good information about the nature of the recipient and some information about the nature of the donor. We call these cases directed recent LGT events, or dLGTs.

For most of the acquired genes (85%) we found no homologs that satisfy the sequence similarity and nucleotide content variation threshold criteria that we set for calling a dLGT. For the remaining 8% acquired genes we could not infer the LGT reliably. The number of completely sequenced genomes per genus explains 29% ($P \ll 0.01$, using Spearman test) of the variation in the proportion of dLGT to gene acquisitions per species; hence, the genome sample is a limiting factor for donor identification. With increasing sample size, larger proportions of dLGTs among the recent LGTs will ensue.

All 32,028 polarized lateral recipient–donor protein-coding gene transfer events were summarized into a directed LGT network (Fig. 2A). The total data comprises 657 contemporary species and 656 ancestral species (internal nodes in the reference tree). Discarding all genomes and ancestors for which no donor–recipient relations were inferred results in a smaller network comprising 715 vertices that are either contemporary genomes (545) or ancestors (170). The vertices are connected by 3021 directed edges that are the actual inferred gene transfer events, pointing from the donor vertex to the recipient vertex. Edge weights (a_{ij}) in this network are the number of genes that were transferred from donor *i* to recipient *j*. The total of all edge weights is the number of protein-coding gene transfers in the network.

Biological examples within the directed network

The dLGT network contains a main connected component of 430 vertices and 63 additional connected components including between two and 44 vertices, with 285 vertices in total. The small components are species that are connected by recent LGT events among themselves, but no dLGT was identified between them and species included in the main connected component, on the basis of the present sample. These small groups typically comprise intracellular pathogens or endosymbionts, such as *Legionella pneumophila, Leptospira interrogans*, and the like, whose host-associated life style is a barrier to LGT, although they are sometimes able to exchange genes among themselves (Russell and Moran 2005). The endosymbiont-specific connected components are an important internal positive control for this directed network approach to LGT, because from the underlying biology of these organisms we know that they should be rarely connected via recent LGT to other species.

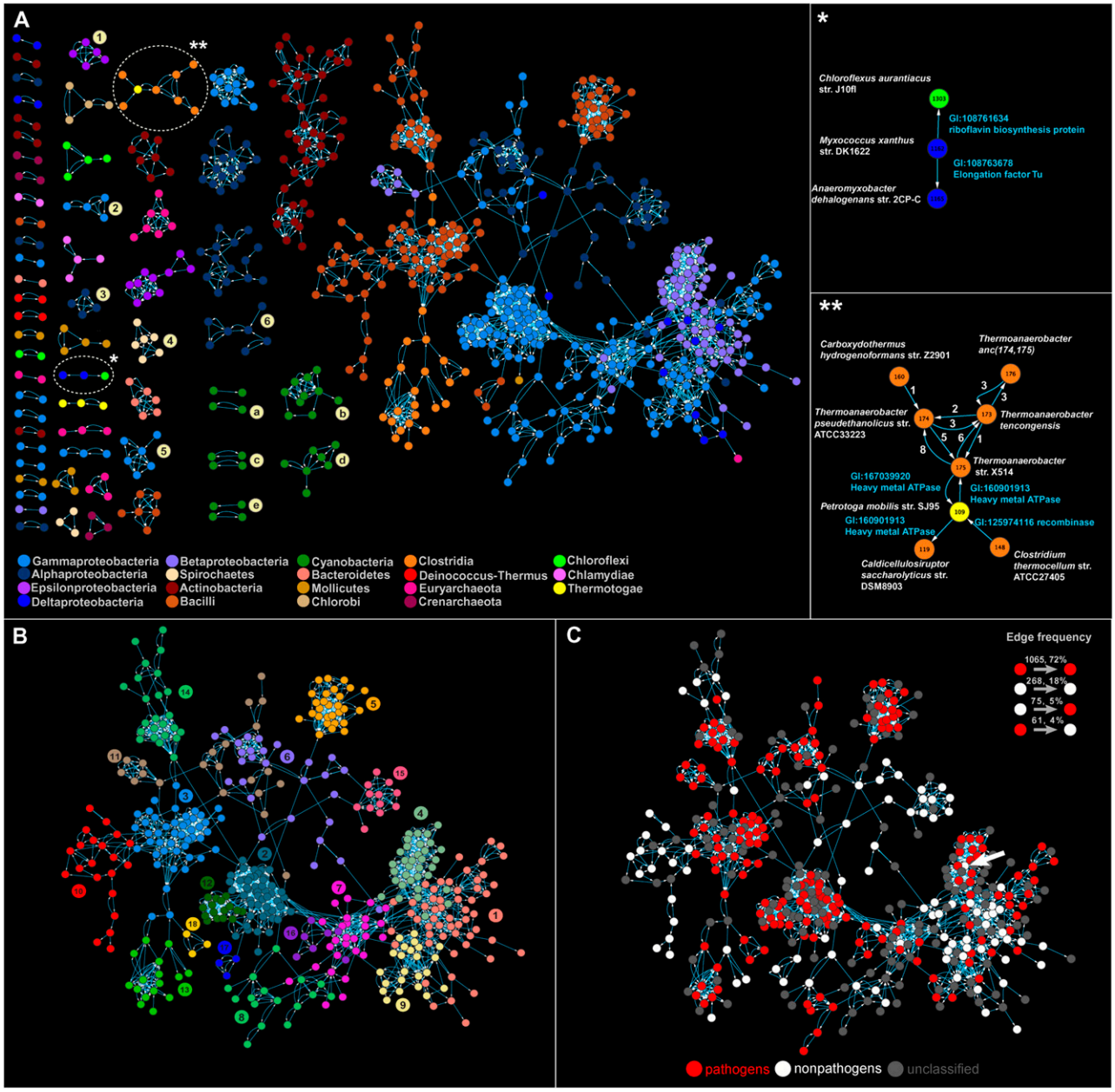


Figure 2. (*A*) The directed network of recent lateral gene transfers. Node color corresponds to the taxonomic group of donors and recipients listed at the *bottom*. Connected components of endosymbionts are marked with numbers: (1) Helicobacter, (2) Coxiella, (3) Bartonella, (4) Leptospira, (5) Legionella, (6) Ehrlichia. Clusters of cyanobacteria are marked with letters: (a) high-light adapted Prochlorococcus, (b) low-light adapted Prochlorococcus, (c) marine Synechococcus, (d) other Synechococcus, (e) Nostocales and Chroococcales. Enlarged images of clusters (*right*) are marked with asterisks. Species names are written by the vertices. Annotations of transferred genes appear next to the edges. (*B*) Community structure within the largest connected component of the dLGT network (for the entire network, see Supplemental Fig. S2). Vertices that are grouped into the same module are colored the same. (*C*) Pathogens in the largest connected component of the dLGT network (for the entire network, see Supplemental Fig. S6). The arrow marks a nonpathogen (*Bukholderia thailandensis*) within a pathogenic community.

The dLGT network method recovers that result. Cyanobacteria form seven distinct connected components within the network. These include high-light adapted *Prochlorococcus* (10 nodes), two connected components of low-light adapted *Prochlorococcus* (three and two nodes), three connected components of *Synecococcus* (eight, two, and two nodes), and Chroococcales with Nostocales (four nodes). In other words, the cyanobacteria in our network are assorted into dLGT donor–recipient-connected components both by genus and by habitat.

The network comprises 662 acquiring genomes and 658 donating genomes, with 598 genomes that are specified as both. Most of the species within the network are connected with only a few other vertices. The number of donors per acquiring species (IN degree) ranges between one and 34, with 25% (164) of the vertices connected to a single donor (Fig. 3A). A total of 25 (4%) species are connected to more than 15 different donors; these are mainly found within Enterobacteriales (γ-proteobacteria), Burkholderiales (β-proteobacteria), and staphylococci (Bacilli). The species harboring

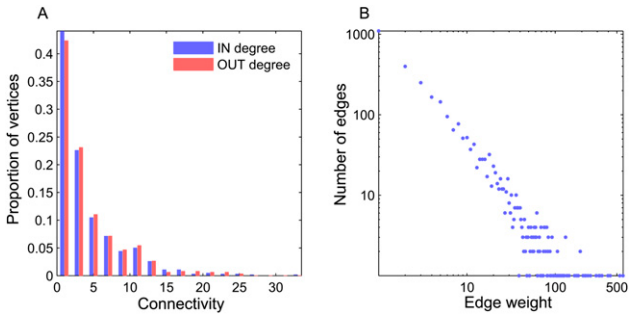


Figure 3. Distribution of connectivity and edge weight in the dLGT network.

the highest frequency of recent acquisitions is *Citrobacter koseri* str. ATCC-BAA-895 (γ-proteobacteria), with 146 IN degree proteins. *C. koseri* is a bacterium that can reside either as a free-living species in soil and water or as a human commensal; it is notable that all of the donors connected to it are Enterobacteriales.

The number of recipients per donating species (OUT degree) ranges between one and 25 recipients, with a majority of a single recipient per donor (159; 25%) (Fig. 3A). The most frequently donating species is *E. coli* str. HS, and all of its 25 recipients are Enterobacteriales. Vertex IN and OUT degrees are positively correlated ($r_s = 0.78$, $P \ll 0.01$); hence, species in the dLGT network are similarly connected as recipients and donors. Both species IN and OUT degrees are weakly correlated with genome size ($r_s = 0.38$ and $r_s = 0.39$, respectively, $P \ll 0.01$).

The distribution of edge weight within the dLGT network is linear in log–log scale; hence, most of the donor–recipient connections only entail a few genes (Fig. 3B). Edges of a single transferred gene are frequent within the dLGT network (1098; 36% of the total), while edges of >20 genes are rare (354; 12%). Most of the heavy edges are concentrated within the heavily connected clusters, which are in turn connected by weaker edges (Supplemental Fig. S1). Edges connecting vertices from the same higher taxonomic group have significantly higher weights than those connecting vertices from different groups ($P \ll 0.01$, using the Kolmogorov-Smirnov test).

Community structure in the directed network of recent LGT

Communities within a network are groups of vertices that are more densely connected among each other than with vertices outside of the group. We examined community structure within the dLGT network using a modularity optimization method that makes an explicit use of the information contained in edge directions (Leicht and Newman 2008). That procedure reveals 85 communities containing between two and 55 vertices, with a median of three vertices per community (Fig. 1B). The main cluster in the dLGT network includes 18 connected communities. Only eight communities include species from different higher taxonomic groups, while the rest of the communities are taxonomically homogeneous. The largest taxonomically heterogeneous community is within the main cluster (community 1 in Fig. 2B). It includes 55 vertices from β-proteobacteria (33), γ-proteobacteria (15), δ-proteobacteria (four), α-proteobacteria (two), and Euryarchaeota (one). The vertices within the communities are connected by 2383 edegs, of which 2341 (98%) are within the same taxonomic group, and 42 (2%) are between species from different groups. The top recipient in this module is *Herminiimonas arsenicoxydans*, a heterotrophic β-proteobacterium that was isolated from heavy metal contaminated sludge from an industrial water-treatment plant (Muller et al. 2006). The donors connected to *H. arsenicoxydans* are *Parvibaculum lavamentivorans*, an α-proteobacterium isolated from urban sewage treatment plants (Schleheck et al. 2000), and *Xanthomonas campestris* str. vesicatoria 85-10, a plant parasitic γ-proteobacterium that can live in both soil and water. All of the seven genes transferred from *P. lavamentivorans* to *H. arsenicoxydans* are hypothetical proteins. One of the two genes that *H. arsenicoxydans* acquired from *X. campestris* is an integrase that also has homologs in other soil bacteria such as Burkhoderiales and Pseudomonadales (Muller et al. 2007), suggesting that a soil phage might be common to and link these genomes.

The most heavily connected higher taxa are β-proteobacteria and γ-proteobacteria, with 64 (42%) edges out of the 150 edges that link higher taxa in the network, and including 13 conjugation/transferase genes and three integrases. Most of the transfers occur among soil bacteria such as Burkholderiales, Xanthomonadales, and Pseudomonadales. Another common order in this subset is the Alteromonadales, represented by the *Shewanella* sp. str. ANA-3. This γ-proteobacterium was isolated from an arsenate-treated wooden pier located in a brackish estuary (Saltikov et al. 2003). The genus *Shewanella* usually resides in marine habitats, so that their link to this subset is probably due to gene exchange with aquatic Burkholderiales.

The second most frequent subset of recent intergroup edges is that of *Clostridium* and *Bacillus* species (32 edges). Most of the edges connect soil-dwelling bacteria such as *Bacillus cereus* str. ATCC 10987 and *Clostridium kluyveri* str. DSM 555. Three edges in the dLGT network connect between human pathogens from these groups, the *Finegoldia magna* str. ATCC 29328 and *Streptococcus pyogenes* str. MGAS10750. The 32 genes transferred between these groups comprise both conjugative transposons and phage proteins, implicating both conjugation and transduction in transfer mechanisms.

Although ancient LGT between eubacteria and archaebacteria is very common and well documented among genomes within our sample, for example, *Thermotoga maritima* (Nelson et al. 1999) or *Methanosarcina mazei* strain Gö1 (Deppenmeier et al. 2002), only one recent LGT edge connects eubacteria to archaebacteria in Figure 2A, with a recent transfer of a methyltransferase from *Geobacter uraniumreducens* str. Rf4 to the uncultured methanogenic archaeon RC-I. The recipient was isolated from the rice rhizosphere (Erkel et al. 2006), while the donor belongs to the Geobacteraceae that resides both in soil and water and is probably capable of nitrogen fixation (Holmes et al. 2004). We note, however, that the genome sample of archaebacterial species in the public databases is very limited.

A striking observation from the dLGT network is that most dLGT occurs between donors and recipients within the same taxonomic group (these are nodes having the same color in Fig. 2A). Closely related species from the same taxonomic group usually have similar genomes. The high frequency of edges among closely related genomes implies that the majority of recent LGT occurs among similar species having similar genomes, as has often been suggested from individual case studies (Mau et al. 2006). The present network analysis provides the means to specifically test this idea for many genomes simultaneously with regard to recent LGT events.

Recent LGT frequency correlates to pairwise genome similarity

Early genetic studies in the *E. coli* and *B. subtilis* systems showed that the frequency of gene acquisition via recombination is dependent

upon the similarity of donor and recipient genes (Majewski and Cohan 1998). We asked whether this same tendency could be observed at the whole-genome level for 657 sequenced genomes. Here we used three different donor–recipient genome similarity measures calculated directly from the genome sequences of the donor and recipient. The first is similarity of genome sequence (S_{gs}), calculated as the proportion of ≥20-bp subsequences in the recipient genome that are found in a perfect match with the donor genome, providing a proxy for the likelihood of gene acquisition mediated by homologous recombination. The S_{gs} is similar to the recently suggested average nucleotide identity (ANI) measure that positively correlates with DNA–DNA hybridization in prokaryotes (Richter and Rosselló-Móra 2009) ($r_s = 0.85$, $P = 5.16 \times 10^{-16}$, $n = 54$). Hence, the S_{gs} is also equivalent with phylogenetic proximity. The second is similarity of proteomes (S_{pr}), calculated as the proportion of recipient genome proteins that share an orthologous protein family (orthogroup) with the donor proteome; it is a proxy for similar ecological lifestyles based on gene content (Chaffron et al. 2010). The third is similarity of GC content (S_{GC}), which is calculated as the similarity between the genomic GC content of the donor and recipient.

The S_{gs} and S_{pr} measures are nonsymmetric; hence, in the comparison of a species pair, the designation of donor and recipient may yield slightly different results. These genome similarity measures correlate, but not strictly so, with phylogenetic classification (Supplemental Fig. S3). For example, in a comparison between Donor: *Escherichia coli* str. CFT073 and Recipient: *E. coli* APEC 01, the following similarity measures are calculated: S_{gs} = 86.2%, S_{pr} = 81.3%, and S_{gc} = 99.3%. With the same donor and a recipient from different species, but still within the *E. coli* complex, *Shigella flexneri* str. 2a, the similarity measures are: S_{gs} = 60%, S_{pr} = 72.3%, and S_{gc} = 96.6%. For a recipient from within the enterobacteriales (same order), *Salmonella typhi*, the values are S_{gs} = 8.6%, S_{pr} = 64.7%, and S_{gc} = 98.4%. These values may change across taxonomic groups and ecological niches.

Particularly low S_{gs} values are observed among cyanobacteria. A comparison between Donor: *Prochlorococcus marinus* str. MED4 and Recipient: *Prochlorococcus marinus* str. MIT 9313 results in: S_{gs} = 0.92%, S_{pr} = 73.3%, and S_{gc} = 92.8%. A comparison of the same donor with Recipient: *Synechococcus sp.* str. WH8102 results in S_{gs} = 0.034%, S_{pr} = 54.2%, and S_{gc} = 71.4%. The low S_{pr} in cyanobacterial species is attributable to the different ecological niches they inhabit (Rocap et al. 2003), while the low S_{gs} is due to their different genomic GC content, meaning different codon usage.

All possible species pairs in our genome set can be readily divided into two groups—those that are connected by a dLGT edge (connected genomes) and those that are not (disconnected genomes). The median in all three genome-similarity measures is significantly higher for connected genomes (Fig. 4; $P \ll 0.01$ using the Wilcoxon test) than for unconnected genomes. Thus, dLGT recipients are more likely to acquire genes from donors of similar genome sequence, similar proteome, and/or similar genomic GC content than from genomes that are more distant by those criteria.

All three genome-similarity measures correlate significantly with the number of transferred genes from the donor to the recipient. Directed gene acquisition frequency is positively correlated with genome sequence similarity S_{gs} ($r_s = 0.41$, $P \ll 0.01$), proteome similarity S_{pr} ($r_s = 0.42$, $P \ll 0.01$), and S_{GC} ($r_s = 0.4$, $P \ll 0.01$). However, several species pairs having high genome similarity by all measures stood out by having very low frequencies of dLGT. Upon closer inspection, we find that many of those are pairs that include one or two host-associated species. Closely related endosymbionts (e.g., two *Legionella* strains) are highly similar by all similarity measures, yet they rarely donate or acquire genes because their symbiotic relation with the host is a barrier to LGT in many cases. Excluding symbiotic species from the correlation tests increases correlation between the number of transferred genes and genome sequence similarity ($r_s = 0.55$, $P \ll 0.01$), proteome similarity ($r_s = 0.53$, $P \ll 0.01$), and GC content similarity ($r_s = 0.47$ $P \ll 0.01$). A multiple correlation analysis using all three similarity measures as predictors of the frequency of transferred genes yielded a model of total $R^2 = 26\%$ explained variability in the number of transferred genes. The variation in S_{gs} contributes 25% to the total explained variability, while variation in S_{pr} contributed only 1%. GC similarity measure (S_{gc}) did not increase the variability explained by the model and was therefore omitted. We note, however, that the range of S_{gc} is highly limited within the network ranging between 75% and 99% GC content similarity (Fig. 4), with 86% of the donor–recipient having $S_{gc} > 95\%$, and 53% of the pairs having $S_{gc} > 99\%$. Accordingly, for the hundreds of genomes contained within this directed network, prokaryotes preferentially assimilate genes from donors with similar genome attributes in terms of sequence identity, GC content, and gene content.

The distribution of both S_{gs} and S_{pr} show that the frequency of recently transferred genes in the dLGT network has a peak around 50% donor–recipient similarity, with a tail toward 100% similarity (Fig. 4). This occurs because the majority of recombination events between almost identical genomes cannot be detected by sequence comparison due to insufficient sequence divergence. Genomes having close to 100% similarity are always from the same species (Supplemental Fig. 3A). Hence, the resolution achieved using our LGT detection method yields a minimum of intraspecific recombination events within the dLGT network. The LGT events that are detected at high genome similarity levels are attributable to genes polymorphic for presence or absence within the population.

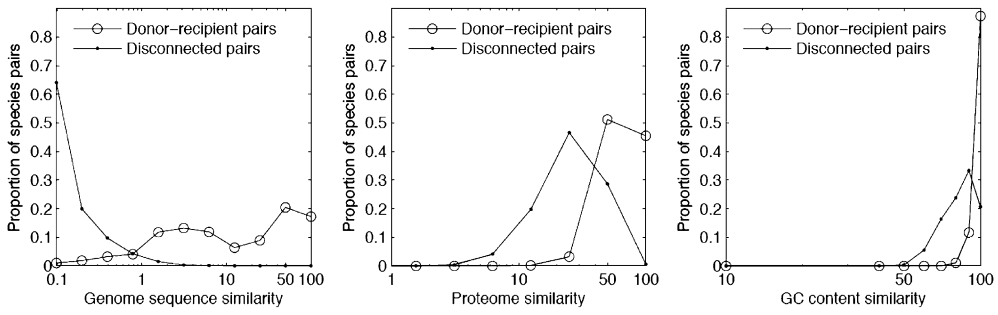


Figure 4. Comparison of genome similarity measures for donor–recipient pairs and disconnected pairs.

Recent LGT between distantly related species

Despite the prevalence of recent LGT from closely related donor genomes in the dLGT network, there remains a substantial fraction of transfers donated by species that are only distantly related to the recipient. For example, if we collapse the network so as to only depict dLGTs at the intergeneric level or higher, 157 vertices remain that are linked by 376 edges carrying 1530 proteins (Supplemental Fig. S4). Most of the small clusters in the complete dLGT network are condensed to vertices in the intergeneric network because they comprise intrageneric donors and recipients only. The edges that remain consist of intergeneric recent lateral gene transfers (irLGTs). The irLGT genus-level network includes one main connected component of 109 nodes with 145 genera, two smaller connected components of Actinobacteria and α-proteobacteria, and 12 additional tiny connected components of two or three genera each. Most of the irLGTs occur among Proteobacteria, again specifically within γ-proteobacteria and β-proteobacteria, and most events involve only one donor and recipient (Supplemental Fig. S4B). The median edge weight is one gene per edge (Supplemental Fig. S4C), similar to the dLGT network.

The establishment of DNA acquired by transduction is mediated by phage enzymes (Ochman et al. 2000; Thomas and Nielsen 2005) and LGT via conjugation and transformation typically involves homologous recombination (HR). But DNA acquired from a more distantly related donor is expected to be less similar to that of the recipient than DNA acquired during an intrageneric LGT, and the minimal requirements for homologous recombination—two anchors of 20–30 bp bearing nearly 100% similarity to the recipient chromosome in *Bacillus subtilis* (Majewski and Cohan 1999) or one anchor of identical 25 bp in *E. coli* (Lovett et al. 2002)—will often not be met. In such cases, other information-processing pathways must be involved in the incorporation of the acquired DNA within the recipient chromosomes. We turned our attention to nonhomologous end-joining.

LGT is mediated by nonhomologous end-joining

Nonhomologous end-joining (NHEJ) is a major DNA double-strand break repair (DSBR) mechanism that was first described in mammalian cells (Bassing and Alt 2004; Lieber et al. 2004). NHEJ involves the religation of two broken ends of a chromosome in the absence of long sequence homology. NHEJ can function either with only a few bases homology between the repaired chromosome ends, known as microhomology, or without microhomology in a blunt-directed repair. During religation, exogenous DNA may be captured in the chromosome, leading to insertion of DNA into the genome. In eukaryotes, DNA inserted into the genome by NHEJ during evolution may include foreign DNA fragments such as mitochondrial DNA, transposable elements, and viral DNA (Moore and Haber 1996; Ricchetti et al. 1999; Lin and Waldman 2001a; Lin and Waldman 2001b; Nakai et al. 2003; Hazkani-Covo and Covo 2008). The classical eukaryotic NHEJ machinery includes the KU70/80 heterodimer (KU), XRCC4, Ligase IV, and DNA-PKcs proteins (Bassing and Alt 2004; Lieber et al. 2004). A prokaryotic NHEJ pathway was predicted from whole-genome analyses, and later shown to be functional in *B. subtillis* (Aravind and Koonin 2001; Weller et al. 2002). The prokaryotic NHEJ is similar to the eukaryotic system in its reliance on a DNA end-binding Ku protein and a dedicated ATP-dependent DNA ligase (Lig4 in eukaryotes and LigD in prokaryotes). Contrary to the eukaryotic system that includes various factors promoting the end processing and ligation stages, in the prokaryotic system the ATP-dependent ligase includes an additional nuclease domain that enables interaction between the Ku and the LigD proteins, thus forming a two-component NHEJ system (Shuman and Glickman 2007).

There are 141 genomes in our sample that encode both Ku and LigD, 116 of which are inferred recipients in the dLGT network. If NHEJ is indeed involved in gene acquisition by LGT, then those genomes harboring Ku and LigD proteins should have a higher frequency of intergeneric dLGT than genomes that lack the nonhomologous end-joining proteins.

To test this, we divided the genomes in our sample according to the presence of both Ku and LigD proteins ($NHEJ^+$), or the absence of one or both proteins ($NHEJ^-$), and examined the distribution of S_{gs}, S_{pr}, and S_{GC} for all donor–recipient pairs, comparing $NHEJ^+$ and $NHEJ^-$ recipients. The average recipient genome similarity to the donor, using S_{gs} and S_{pr}, is significantly lower in the $NHEJ^+$ than the $NHEJ^-$ group ($P = 0.029$ and $P = 1.4 \times 10^{-7}$, respectively, using the Wilcoxon test) (Fig. 5). No significant difference in genomic GC content similarity was found between the two groups ($P = 0.26$, using the Wilcoxon test). To test for a possible bias in this result due to our genome sample, we repeated the test using all 657 sampled genomes regardless of their inclusion in the dLGT network, but found no significant difference in the genome similarity measures

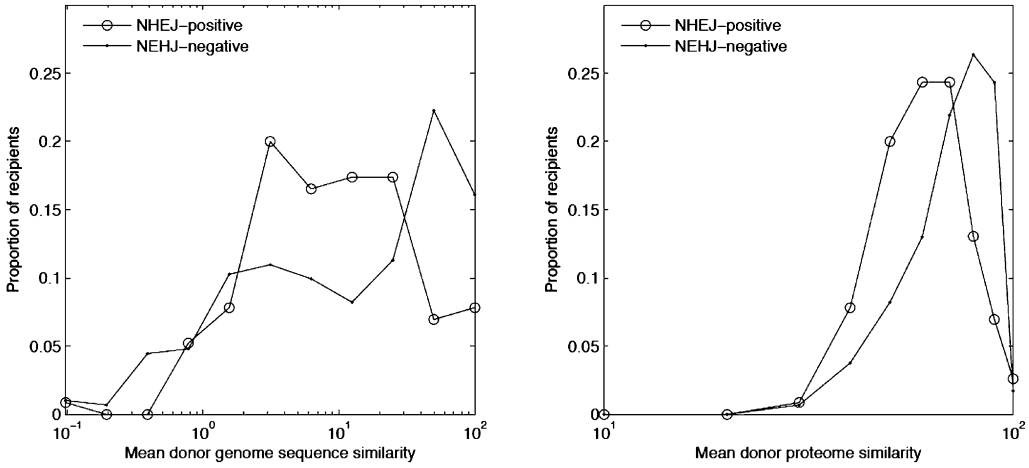


Figure 5. Comparison of genome similarity measures between NHEJ-positive and NHEJ-negative recipients.

between the two groups. This shows that more frequent acquisition from distant donors in NHEJ$^+$ genomes is not biased by the genome sample and that S_{gs} is the more sensitive measure among the three. The microhomologies typical of insertion via NHEJ (Hazkani-Covo and Covo 2008) could not be detected in the present data, probably due to the insufficiently dense genome sample.

Types of genes and types of genomes

Sorting all genes within the dLGT into functional categories using the COG scheme (Tatusov et al. 2003) revealed that the functional distribution of transferred genes is not random ($P < 1 \times 10^{-16}$, using the χ^2 test) with most of the classified genes performing metabolism functions (7306; 29%). The most frequently transferred classes are amino acid transport and metabolism, energy production and conversion, and carbohydrate transport and metabolism. Genes involved in cellular processes and signaling comprise 3623 (14.4%), while information storage and processing genes are transferred less often than the other categories (2802; 11.1%) (Fig. 6). The distribution of donor–recipient genome similarity using all three measures is significantly different among the four main functional categories (Kruskal-Wallis test, S_{gs}: $P < 1 \times 10^{-15}$; S_{gc}: $P < 1 \times 10^{-16}$; S_{pr}: $P < 1 \times 10^{-16}$). Donor–recipient similarity for transferred genes in the information storage and processing category is significantly lower than all other functional categories by all genome similarity measures ($\alpha = 0.05$, using Tukey post hoc comparisons).

Most of the transferred genes are either unclassified in the COG database or are classified in COG as unknown (11,470; 45.5%). The distribution of recipient taxa within the unknown genes shows that β-proteobacteria and Clostridia recipients include disproportionately higher numbers of unknown genes in the dLGT network. In contrast, Bacilli and γ-proteobacteria recipients (Supplemental Fig. S5) contain more classified genes than their proportion in the dLGT network.

What kinds of organisms are involved in recent LGT? Using NCBI's organism information table (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) we classified 433 sequenced species in the dLGT network into 261 pathogens and 172 nonpathogens. Most of the edges within the network connect pathogenic recipients and donors (Fig. 2C). To test whether this result is biased by our genome sample, which contains a majority of pathogens (299 vs. 254 nonpathogens), we compared these frequencies with the expected number of edges among all 657 genomes regardless of their inclusion in the dLGT network. The observed edge frequency within the pathogens/nonpathogens groups is independent of the genome sample alone ($P = 0.06$, using the χ^2 test), with edges from a pathogenic donor to a pathogenic recipient over-represented in the network. Pathogenic species have a significantly higher IN degree and OUT degree in comparison to nonpathogenic species ($P < 1 \times 10^{-16}$ in both cases using the one-tailed Kolmogorov-Smirnov test; Supplemental Fig. S7). However, donor–recipient pairs having S_{gs} <10% reveal similar IN and OUT degrees for edges connecting to pathogens and nonpathogens, respectively ($P < 1 \times 10^{-16}$ using the Kolmogorov-Smirnov test). Moreover, pairs of pathogenic donor and recipient connected by a LGT event have a significantly higher S_{gs} and S_{pr} than other pathogenic and nonpathogenic donor and recipient combinations ($P < 1 \times 10^{-16}$ using the Kruskal-Wallis and Tukey post-hoc comparisons). Hence, for closely related donors and recipients, pathogens receive and donate genes by LGT more frequently than nonpathogenic species. The modules in the dLGT can be classified with regard to pathogenicity of the connected species. A total of 39 modules comprise only nonpathogens, 27 modules comprise only pathogens, 17 modules are mixed pathogens and nonpathogens, and the remaining two are of an unclassified species (Fig. 2C). Module no. 4 (Fig. 2B,C, arrow) is an example of a mixed community that includes five pathogens, four nonpathogens, and 18 unclassified species. In this module we detected abundant recent LGT between the nonpathogenic *Burkholderia thailandensis* str. E246 and pathogenic Burkholderia, including *B. pseudomallei* strains K96243, 1710b, and 1106a, and *B. mallei* strains ATCC 23344, NCTC 10229, SAVP1, and NCTC 10247. *B. thailandensis* and *B. pseudomallei* are considered as two distinct species (Gevers et al. 2005); however, their genomes are highly similar in sequence and content (Yu et al. 2006). The abundant lateral gene transfer among these genomes is thought to be mediated mainly by transduction (Summer et al. 2007). We find

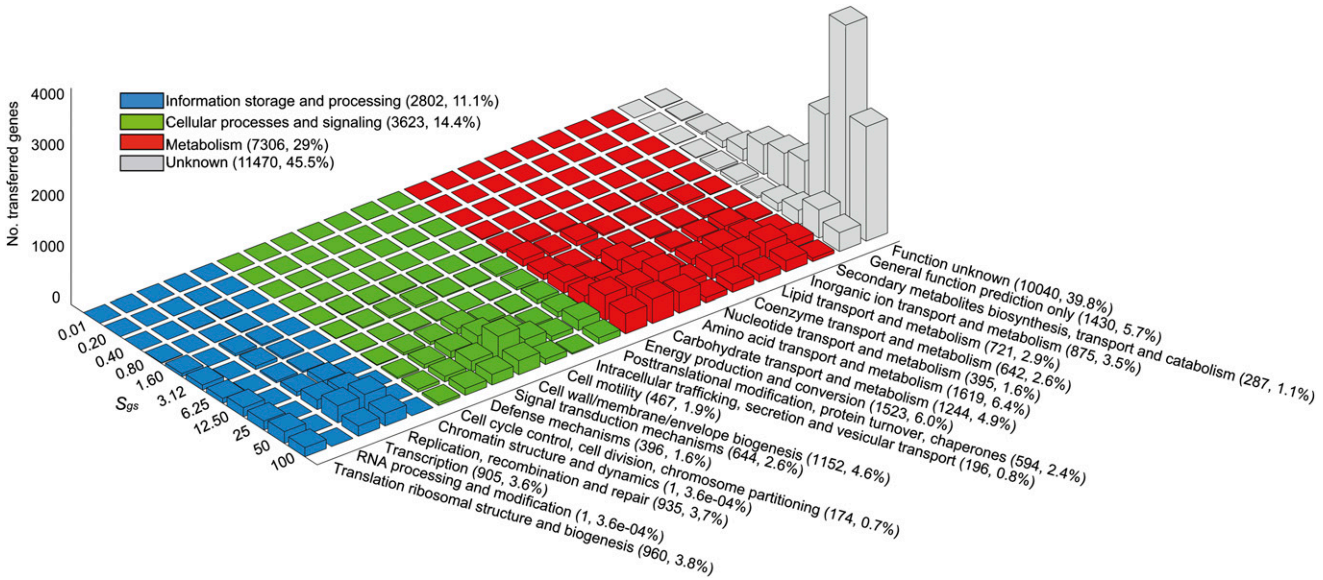


Figure 6. Frequency of transferred genes by functional category and donor–recipient genome similarity.

evidence for LGT by transduction in the transferred phage genes such as phage minor tail protein (Donor: *B. pseudomallei* str. K96243 to Recipient: *B. thailandensis* str. E246) and phage major tail tube protein (Donor: *B. pseudomallei* str. 1710b to Recipient: *B. thailandensis* str. E246). The dLGT network reveals that nonpathogens can sometime mediate gene transfer between pathogenic populations.

Discussion

Directed networks in which donor–recipient relations are coded as polarized vectors, as they occur in nature, open up fundamentally new avenues of pursuit in the investigation of microbial genome dynamics. Among 2,129,548 proteins in 657 prokaryotic genomes, we identified 446,854 as having been recently acquired on the basis of their aberrant nucleotide pattern properties relative to the rest of their genome. For 32,028 of those genes we inferred the identity of the donor among the present sample based on sequence identity, GC content, and phylogenetic reconstruction. With improved genome sampling or in metagenomic data of finite complexity, such as intestinal flora (Warnecke et al. 2007), the proportion of specifiable donors in the data, hence, the density of the directed network for recent transfers should improve.

The dLGT network reveals a high correlation between donor–recipient genome similarity and lateral gene-transfer frequency. Hence, the majority of recent LGT events in the dLGT network occur among closely related species. This finding is in agreement with earlier suggestions that there exists a gradient of LGT frequency that is higher within taxonomic groups and lower between taxonomic groups (Gogarten et al. 2002; Puigbò et al. 2010). The high LGT frequency between similar genomes can be largely explained by the mechanisms for LGT in prokaryotes. The incorporation of acquired DNA into the recipient genome in both transformation and conjugation is commonly mediated by homologous recombination (Thomas and Nielsen 2005). Thus, gene acquisition by these two LGT mechanisms has an inherent donor–recipient sequence similarity threshold. In contrast, during gene acquisition by transduction the DNA is incorporated into the recipient chromosome by the phage enzymes (Ochman et al. 2000); hence, the donor-recipient genome similarity barrier is less apparent. The reconstructed gene transfers of integrons and phage parts in the dLGT network are evidence that some of the reconstructed LGTs in the network were mediated by transduction. Consequently, our results suggest that the genome similarity barrier applies also to phage-mediated gene transfer. This implies that most of the phages are transferring DNA between similar species. Indeed, a similar scenario is described for bacteriophages of the oceanic cyanobacterium *Prochlorococcus* (Sullivan et al. 2003).

Examples of gene acquisition from distantly related donors are documented in the literature (Nelson et al. 1999; Mongodin et al. 2005) and are also apparent in our dLGT network. It follows that donor–recipient genome similarity is not always a barrier to LGT. We demonstrated that genomes encoding the nonhomologous end-joining (NHEJ) proteins Ku and LigD are significantly more likely to acquire DNA from a distantly related donor genome than genomes lacking NHEJ. While we cannot exclude the possibility that our result is biased by a hidden genomic variable related to LGT and covariates with NHEJ presence/absence, in the lack of evidence to that effect we conclude that NHEJ has a role in LGT within prokaryotes. NHEJ is not the only mechanism to bypass the genome similarity barrier to LGT. For example, transformation frequencies at different genomic loci of *Acinetobacter baylyi*, which lacks the NHEJ proteins, were shown to vary up to 10,000-fold (Ray et al. 2009). Moreover, Chayot et al. (2010) recently showed that DNA acquisition in *E. coli* can be mediated in vitro by a mechanism that is independent of homologous recombination. *E. coli*, which lacks the NHEJ pathway, possesses an alternative end-joining mechanism (A-EJ) for DNA double-strand break repair (Chayot et al. 2010). The A-EJ mechanism recruits the RecBCD complex for end-restriction and Ligase A for DNA ligation. Chayot et al. (2010) showed that an acquisition of antibiotic resistance gene in *E. coli* can be mediated by the A-EJ proteins, demonstrating the possible role of DSBR end-joining mechanisms in LGT. This suggests the existence of yet unexplored roles of DNA repair mechanisms for integrating acquired DNA into prokaryotic genomes.

Our results show that the functional distribution of transferred genes is not random, as suggested earlier (Choi and Kim 2007). The abundance of metabolic genes and scarceness of informational genes within the dLGT network are in agreement with the complexity hypothesis (Jain et al. 1999), according to which informational genes are transferred less frequently than those in the operational class. The overall similarity of donor–recipient genomes is lesser for transferred genes in the information storage and processing category in comparison to the other functional categories. This finding seems at first counterintuitive with regard to the complexity hypothesis. However, the low similarity between the donor and recipient might actually explain how these genes are still transferred. Sorek et al. (2007) showed that information genes can be readily acquired as long as they are not expressed. Hence, it is possible, if not likely, that many of the informational genes identified here are not expressed in the recipient genomes.

For recent LGT, it appears that the lateral component of prokaryotic genome evolution can be accurately modeled with directed networks and that the accuracy should increase with increasing sample density. For more ancient acquisitions it should, in principle, be possible to approximate donors using gene phylogeny-dependent methods, and thereby further expand the application spectrum of directed networks in the study of microbial genome evolution.

Methods

Data

Fully sequenced genomes of 657 prokaryotes were downloaded from the NCBI website (www.ncbi.nlm.nih.gov April 2008 version). Our recent LGT inference operates within the framework of orthologous protein families and is assisted by a reference species tree as described in Dagan and Martin (2007). First, we classified all 2,129,548 proteins encoded within chromosomes in our genome sample into orthologous protein families. The common protein families reconstruction methods COGs (Tatusov et al. 2003) and MCL (Enright et al. 2002) are inappropriate for our purpose since they sometimes yield protein families that include paralogs in addition to orthologs, and a reciprocal best BLAST hit (rBBH) procedure outperforms many more complicated clustering algorithms (Altenhoff and Dessimoz 2009). We therefore used a greedy algorithm similar to the bits-score algorithm used in COG database (Tatusov et al. 2003), which groups all rBBHs into one orthogroup. Only BLAST hits having an E-value $\leq 1 \times 10^{-10}$, amino acid identity $\geq$25%, and query/hit length ratio $\geq$80% were considered in the rBBH procedure. A new orthogroup begins with a previously unclustered seed gene and all of its rBBHs. Next, all genes included in the orthogroup are used to iteratively search for rBBHs within the genomes not yet represented in the orthogroup. Only genes

identified as recently acquired by LGT (see below) were used as seeds. A comparison of the orthogroups to MCL protein families (Enright et al. 2002) shows that the orthogroups are much more exclusive, yet in the genus scale they overlap completely in 92% of the cases (Supplemental Fig. S8).

Identification of recently acquired genes

Recently acquired genes are expected to have unusual codon usage and GC content when compared with the whole proteome. Therefore, GC content may be used to detect the foreign origin of a gene (Garcia-Vallve et al. 2000; Nakamura et al. 2004). The statistical analysis of GC content is favored over codon usage because it is more statistically robust due to the smaller relative standard errors of the estimates resulting from a larger sample size (nucleotides vs. codons) and smaller number of states (two vs. 61). Genes with atypical GC content are detected by comparing their GC content with the genomic GC using the χ^2 test with a false discovery rate (FDR) of 5% (Benjamini and Hochberg 1995).

Gene acquisitions within each orthogroup are superimposed upon the reference tree. If a certain group of recipients is monophyletic, then the acquisition event is reconstructed to their common ancestor (an internal node in the reference species tree). Otherwise, the acquisition event is reconstructed to the species in which it was detected. These are designated as putative recipients.

Identification of gene donor

The candidate gene-donor of each acquired gene is the genome bearing an ortholog with the highest sequence similarity to the acquired gene, excluding all orthologs that share a common acquisition event with the acquired gene. In case of equally similar candidates, all orthologs are stored as candidate donors.

In the next stage, we filtered out candidate donors whose GC content does not fit the expected content by the sequence divergence of the donor and recipient genes. For this purpose we developed an empirical model that describes the difference in GC content between donor and recipient sequences as a function of the evolutionary distance between them. The underlying data for the model are 68,923 pairwise alignments of non-LGT orthologs (genes that are not identified as recent acquisition in the previous stage) from our data set for 74 pairs of sibling species having significantly different genomic GC content ($P < 0.05$ using the χ^2 test). From each pairwise alignment (280,836 alignments in total) we extracted the frequency and type of nucleotide substitutions (A↔T, C↔G, and A/T↔C/G). The data was binned by the frequency of nucleotide substitutions (sequence divergence) per alignment (Supplemental Fig. 9A). The 95% percentile within each bin signifies the confidence interval for the expected difference in G and C nucleotides in that sequence divergence range in $\alpha = 0.05$ significance level. Because sequence divergence and the 95% percentile of A/T↔C/G substitutions frequency are linearly correlates in log–log scale, we could fit a logarithmic equation for the relation between the two variables. The result model is $hbGC = e^{0.8638 \cdot \log n}$, where *hbGC* is the higher bound for the difference in G and C nucleotides between donor and recipient genes, and n is the total number of different nucleotides between the two sequences (Supplemental Fig. S9B). Candidate donor sequences that differ from the recipient in more G and C nucleotides than expected under the model are excluded, those that remain are called putative donors.

At this stage, we filtered for nonfunctional genes by testing for relaxation of purifying selection on the recipient gene. The recipient and donor proteins were aligned using ClustalW (Thompson et al. 1994), and were converted to codons alignment using PAL2NAL (Suyama et al. 2006). The ratio of $\omega = {}^{d_n}/_{d_s}$ (Nei and Gojobori 1986) was calculated by PAML (Yang 2007). A total of 140 genes having $\omega > 0.95$ were considered as pseudogenes and were excluded from the analysis.

Aberrant nucleotide pattern (or codon usage) alone is not sufficiently reliable to predict a gene as laterally transferred (Medrano-Soto et al. 2004). In the next stage of the analysis we reconstructed a phylogenetic tree for each of the putative laterally transferred genes. For each gene acquisition event, all of the putative recipients and putative donors are aligned together with two outgroup sequences. One outgroup is an ortholog from a species that branches between the putative recipients and putative donors in the reference tree. That is, assuming vertical inheritance only, this outgroup is more closely related to the putative recipients than the putative donors. The second outgroup (root outgroup) is an ortholog from a species that branches outside of the clade, including the putative recipients and donors in the reference tree (Supplemental Fig. S10). DNA sequences of the putative recipients, putative donors, and both outgroup sequences are aligned using ClustalW (Thompson et al. 1994). A phylogenetic tree is reconstructed employing the neighbor joining (Saitou and Nei 1987) approach using NEIGHBOR (Felsenstein 2005) with F84 substitution matrix. The phylogenetic trees were rooted with the root outgroup and scanned for sister clades containing only donors in one clade and only recipients in the other. Such sister clades define the source and target of the gene transfer event, and when mapped upon the reference tree, define a directed edge in the dLGT network. We repeated the analysis with phylogenetic trees reconstructed by the maximum likelihood (ML) approach using PhyML (Guindon and Gascuel 2003) with HKY substitution model and empirical base frequency estimates. The ML-dLGT network includes an additional 407 transferred genes and overlaps with the dLGT in 2886 (96%) of the edges. Trends of genome similarity measures in the comparison of NHEJ-positive and NHEJ-negative genomes (see below) are identical to those that resulted from the dLGT network.

dLGT network analysis

Community structure and modules within the dLGT network were inferred by an application of the modularity function to directed networks (Leicht and Newman 2008) using MatLab. The input for the inference script is a binary form of the dLGT network where all edges weights are set to one.

Network views were produces by Cytoscape freeware (Cline et al. 2007) using the force-directed layout (unweighted) option with default parameters. The force-directed layout is a new layout based on the "force-directed" paradigm and implemented by J. Heer as part of the *prefuse* toolkit (http://prefuse.org/). Input files for Cytoscape including the customized vertices and edge coloring were produced using an in-house Perl script.

Genome similarity measures

Genome sequence similarity (S_{gs}) between a recipient and a donor was calculated as the number of identical 20-bp segments between the two genomes, divided by the genome size (total chromosomes length) of the recipient. Identical segments 20-bp long were located using Mummer (Kurtz et al. 2004) and their total length was calculated taking into account possible overlaps, using an in-house Perl script. Proteome similarity (S_{pr}) between a recipient and a donor was calculated as the number of orthogroups that are common to both genomes, divided by the number of orthogroups in which the recipient is represented. GC content similarity (S_{gc}) was calculated by: $100 - |\Delta(GC_{recipient}, GC_{donor})|$. Statistical analysis was performed using MatLab. For the multiple correlation analysis, the *log* value of the predictors and variable was used. The correlation

coefficient of S_{gs} with the ANI measure (Richter and Rosselló-Móra 2009) was calculated from ANIm estimates of 54 species included in the dLGT network using Spearman correlation.

NHEJ-positive genomes

Homologs to YkoU and YkoV proteins were identified by a reciprocal best BLAST hit procedure using the YkoU (gi:16078405) and YkoV (gi:16078406) proteins from *Bacillus subtilis* as the query. Only BLAST hits having an E-value $\leq 1 \times 1^{-10}$ and $\geq$25% amino acids identity were considered. Genomes bearing both NHEJ proteins are designated as NHEJ positive.

Functional classification

Functional classification of recipient genes was extracted from the COG database (Tatusov et al. 2003); http://www.ncbi.nlm.nih.gov/COG/). When the COG annotation of a recipient gene was missing, the donor COG annotation was used instead.

Acknowledgments

This study was supported by the National Evolutionary Synthesis Center (NESCent) grant NSF #EF-0423641 (E.H.-C.), NESCent short-term Sabbatical (T.D.), German Federal Ministry of Education and Research (O.P., T.D., W.M.), European Research Council grant NETWORKORIGINS (W.M.), and the US National Library of Medicine grant LM010009-01 (G.L.).

References

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* **5:** e1000262. doi: 10.1371/journal.pcbi.1000262.

Aravind L, Koonin EV. 2001. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Res* **11:** 1365–1374.

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

Bassing CH, Alt FW. 2004. The cellular response to general and programmed DNA double strand breaks. *DNA Repair* **3:** 781–796.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met* **57:** 289–300.

Chaffron S, Rehrauer H, Pernthaler J, von Mering C. 2010. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res* **20:** 947–959.

Chayot R, Montagne B, Mazel D, Ricchetti M. 2010. An end-joining repair mechanism in *Escherichia coli*. *Proc Natl Acad Sci* **107:** 2141–2146.

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

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

Choi IG, Kim SH. 2007. Global extent of horizontal gene transfer. *Proc Natl Acad Sci* **104:** 4489–4494.

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, et al. 2007. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2:** 2366–2382.

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

Davids W, Zhang Z. 2008. The impact of horizontal gene transfer in shaping operons and protein interaction networks—direct evidence of preferential attachment. *BMC Evol Biol* **8:** 23. doi: 10.1186/1471-2148-8-23.

Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, Henne A, Wiezer A, Baumer S, Jacobi C, et al. 2002. The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. *J Mol Microbiol Biotechnol* **4:** 453–461.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* **284:** 2124–2129.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30:** 1575–1584.

Erkel C, Kube M, Reinhardt R, Liesack W. 2006. Genome of Rice Cluster I archaea–the key methane producers in the rice rhizosphere. *Science* **313:** 370–372.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package). Department of Genome Sciences, University of Washington, Seattle, WA.

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

Garcia-Vallve S, Romeu A, Palau J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10:** 1719–1725.

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL, et al. 2005. Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol* **3:** 733–739.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19:** 2226–2238.

Groisman EA, Ochman H. 1996. Pathogenicity islands: Bacterial evolution in quantum leaps. *Cell* **87:** 791–794.

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

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.

Holmes DE, Nevin KP, Lovley DR. 2004. Comparison of 16S rRNA, nifD, recA, gyrB, rpoB and fusA genes within the family Geobacteraceae fam. nov. *Int J Syst Evol Microbiol* **54:** 1591–1599.

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

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

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5:** R12. doi: 10.1186/gb-2004-5-2-r12.

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

Leicht EA, Newman ME. 2008. Community structure in directed networks. *Phys Rev Lett* **100:** 118703. doi: 10.1103/PhysRevLett.100.118703.

Lieber MR, Ma Y, Pannicke U, Schwarz K. 2004. The mechanism of vertebrate nonhomologous DNA end joining and its role in V(D)J recombination. *DNA Repair* **3:** 817–826.

Lin Y, Waldman AS. 2001a. Capture of DNA sequences at double-strand breaks in mammalian chromosomes. *Genetics* **158:** 1665–1674.

Lin Y, Waldman AS. 2001b. Promiscuous patching of broken chromosomes in mammalian cells with extrachromosomal DNA. *Nucleic Acids Res* **29:** 3975–3981.

Lovett ST, Hurley RL, Sutera VA Jr, Aubuchon RH, Lebedeva MA. 2002. Crossing over between regions of limited homology in *Escherichia coli*. RecA-dependent and RecA-independent pathways. *Genetics* **160:** 851–859.

Majewski J, Cohan FM. 1998. The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*. *Genetics* **148:** 13–18.

Majewski J, Cohan FM. 1999. DNA sequence similarity requirements for interspecific recombination in Bacillus. *Genetics* **153:** 1525–1533.

Mau B, Glasner JD, Darling AE, Perna NT. 2006. Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*. *Genome Biol* **7:** R44. doi: 10.1186/fb-2006-7-5-r44.

Medrano-Soto A, Moreno-Hagelsieb G, Vinuesa P, Christen JA, Collado-Vides J. 2004. Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. *Mol Biol Evol* **21:** 1884–1894.

Milkman R, Bridges MM. 1990. Molecular evolution of the *Escherichia Coli* Chromosome. 3. Clonal frames. *Genetics* **126:** 505–517.

Mongodin EF, Nelson KE, Daugherty S, DeBoy RT, Wister J, Khouri H, Weidman J, Walsh DA, Papke RT, Perez GS, et al. 2005. The genome of *Salinibacter ruber*: Convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proc Natl Acad Sci* **102:** 18147–18152.

Moore JK, Haber JE. 1996. Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. *Nature* **383:** 644–646.

Moran NA. 2002. Microbial minimalism: Genome reduction in bacterial pathogens. *Cell* **108:** 583–586.

Muller D, Simeonova DD, Riegel P, Mangenot S, Koechler S, Lievremont D, Bertin PN, Lett MC. 2006. *Herminiimonas arsenicoxydans* sp. nov., a metalloresistant bacterium. *Int J Syst Evol Microbiol* **56:** 1765–1769.

Muller D, Medigue C, Koechler S, Barbe V, Barakat M, Talla E, Bonnefoy V, Krin E, Arsene-Ploetze F, Carapito C, et al. 2007. A tale of two oxidation states: Bacterial colonization of arsenic-rich environments. *PLoS Genet* **3:** e53. doi: 10.1371/journal.pgen.0030053.

Nakai H, Montini E, Fuess S, Storm TA, Grompe M, Kay MA. 2003. AAV serotype 2 vectors preferentially integrate into active genes in mice. *Nat Genet* **34:** 297–302.
Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36:** 760–766.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3:** 418–426.
Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson LD, Nelson WC, Ketchum KA, et al. 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399:** 323–329.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405:** 299–304.
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.
Palla G, Barabási AL, Vicsek T. 2007. Quantifying social group evolution. *Nature* **446:** 664–667.
Perez JC, Groisman EA. 2009. Evolution of transcriptional regulatory circuits in bacteria. *Cell* **138:** 233–244.
Puigbò P, Wolf YI, Koonin EV. 2010. The tree and net components of prokaryote evolution. *Genome Biol Evol* **2:** 745–756.
Ragan MA, Harlow TJ, Beiko RG. 2006. Do different surrogate methods detect lateral genetic transfer events of different relative ages? *Trends Microbiol* **14:** 4–8.
Ray JL, Harms K, Wikmark OG, Starikova I, Johnsen PJ, Nielsen KM. 2009. Sexual isolation in *Acinetobacter baylyi* is locus-specific and varies 10,000-fold over the genome. *Genetics* **182:** 1165–1181.
Ricchetti M, Fairhead C, Dujon B. 1999. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* **402:** 96–100.
Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci* **106:** 19126–19131.
Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, et al. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424:** 1042–1047.
Russell JA, Moran NA. 2005. Horizontal transfer of bacterial symbionts: Heritability and fitness effects in a novel aphid host. *Appl Environ Microbiol* **71:** 7987–7994.
Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406–425.
Saltikov CW, Cifuentes A, Venkateswaran K, Newman DK. 2003. The ars detoxification system is advantageous but not required for As(V) respiration by the genetically tractable Shewanella species strain ANA-3. *Appl Environ Microbiol* **69:** 2800–2809.
Schleheck D, Dong WB, Denger K, Heinzle E, Cook AM. 2000. An alpha-proteobacterium converts linear alkylbenzenesulfonate surfactants into sulfophenylcarboxylates and linear alkyldiphenyletherdisulfonate surfactants into sulfodiphenylethercarboxylates. *Appl Environ Microbiol* **66:** 1911–1916.
Shuman S, Glickman MS. 2007. Bacterial DNA repair by non-homologous end joining. *Nat Rev Microbiol* **5:** 852–861.
Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318:** 1449–1452.
Sullivan MB, Waterbury JB, Chisholm SW. 2003. Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* **424:** 1047–1051.
Summer EJ, Gill JJ, Upton C, Gonzalez CF, Young R. 2007. Role of phages in the pathogenesis of Burkholderia, or 'Where are the toxin genes in Burkholderia phages?' *Curr Opin Microbiol* **10:** 410–417.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34:** W609–W612.
Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al. 2003. The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41. doi: 10.1186/1471-2105-4-41.
Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* **3:** 711–721.
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.
Tsang JS, Ebert MS, van Oudenaarden A. 2010. Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. *Mol Cell* **38:** 140–153.
Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, et al. 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450:** 560–565.
Weller GR, Kysela B, Roy R, Tonkin LM, Scanlan E, Della M, Devine SK, Day JP, Wilkinson A, d'Adda di Fagagna F, et al. 2002. Identification of a DNA nonhomologous end-joining complex in bacteria. *Science* **297:** 1686–1689.
Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586–1591.
Yu Y, Kim HS, Chua HH, Lin CH, Sim SH, Lin D, Derr A, Engels R, DeShazer D, Birren B, et al. 2006. Genomic patterns of pathogen evolution revealed by comparison of *Burkholderia pseudomallei*, the causative agent of melioidosis, to avirulent *Burkholderia thailandensis*. *BMC Microbiol* **6:** 46. doi: 10.1186/1471-2180-6-46.

Received September 21, 2010; accepted in revised form January 13, 2011.

Chapter II

Trends and barriers to lateral gene transfer in prokaryotes

Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Microbiology

Trends and barriers to lateral gene transfer in prokaryotes

Ovidiu Popa and Tal Dagan

Gene acquisition by lateral gene transfer (LGT) is an important mechanism for natural variation among prokaryotes. Laboratory experiments show that protein-coding genes can be laterally transferred extremely fast among microbial cells, inherited to most of their descendants, and adapt to a new regulatory regime within a short time. Recent advance in the phylogenetic analysis of microbial genomes using networks approach reveals a substantial impact of LGT during microbial genome evolution. Phylogenomic networks of LGT among prokaryotes reconstructed from completely sequenced genomes uncover barriers to LGT in multiple levels. Here we discuss the kinds of barriers to gene acquisition in nature including physical barriers for gene transfer between cells, genomic barriers for the integration of acquired DNA, and functional barriers for the acquisition of new genes.

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

Corresponding author: Dagan, Tal (tal.dagan@hhu.de)

Current Opinion in Microbiology 2011, **14**:615–623

This review comes from a themed issue on
Genomics
Edited by Luis Serrano and Rotem Sorek

Available online 17th August 2011

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2011.07.027

Introduction

Prokaryotes possess the unique ability to acquire DNA from the environment, or their neighbors, and incorporate it into their genome in a process called lateral gene transfer (LGT) [1]. Accumulating evidence shows that LGT plays a major role in prokaryote genome evolution [2–4], affecting virtually all genes [5–7], with only few genes that are resistant to it [8]. Lateral gene transfer is crucial to our understanding of microbial evolution; furthermore, as a source of natural variation it facilitates the emergence of novel infectious diseases through the spread of virulence mechanisms (e.g. [9,10]).

The known mechanisms for LGT include transformation, conjugation, transduction, and gene transfer agents. Transformation involves the uptake of naked DNA from the environment [11,12]. Conjugation is the transfer of DNA via plasmids, a process that is mediated by a proteinaceous cell-to-cell junction, forming a tunnel through which the DNA is transferred [13,14]. Transduction is DNA acquisition following a phage infection [12], and gene transfer agents (GTA) are phage-like DNA-vehicles that are produced by a donor cell and released to the environment [15,16] (Figure 1). An additional transfer mechanism – nanotubes – was discovered recently [26••]. These are tubular protrusions composed of membrane components that can bridge between neighboring cells and conduct the transfer of DNA and proteins (Figure 2).

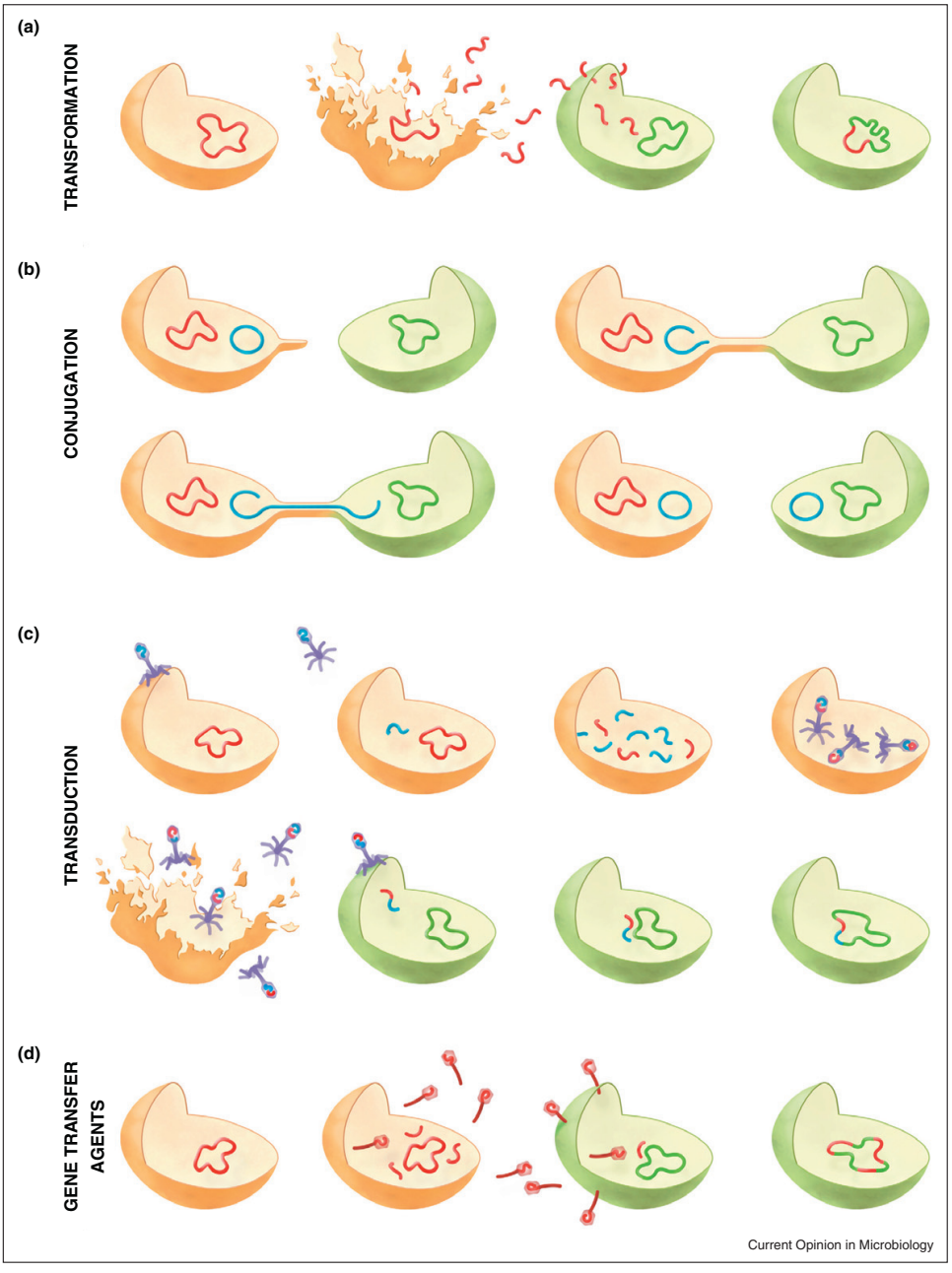
Lateral gene transfer frequency

Several experiments have been conducted in order to quantify the frequency of LGT in nature. For example, Babić *et al.* [27] tested the success rate of gene acquisition by conjugation in *Escherichia coli*. Using a plasmid encoding a gene for fluorescence protein (YFP) they quantified the odds for a successful integration of plasmid genes into the recipient genome. They found that in 96% of the population the YFP gene was integrated into the chromosome and inherited to the next generation. The percolation of an acquired DNA within the population can be extremely fast in *Bacillus subtilis* where the cells are arranged in chains. Tracking the spread of an integrative and conjugative element (ICE) encoding a gene for green fluorescence protein (GFP) under the microscope showed that in 43 (81%) out of 53 cases a recipient cell turned into a donor and transconjugated the ICE to the next cell in line, often within 30 min [28••].

Lateral gene transfer via transduction takes place during a phage infection. Hence gene acquisition by this transfer mechanism depends on the survival of the recipient. In a recent study Kenzaka *et al.* [29] quantified the survival rate of phage infected enteric bacteria as 20% of the population. These surviving bacteria may acquire DNA from previous hosts of the attacking phage. Recent measurements of LGT by gene transfer agents (GTAs) in marine α-proteobacteria revealed that this transfer mechanism is probably the most efficient one. McDaniel *et al.* [30••] measured the frequency of LGT by the acquisition rate of Kanamycin resistance gene. Their results show that gene transfer by GTA is more efficient than transformation or transduction by orders of magnitude.

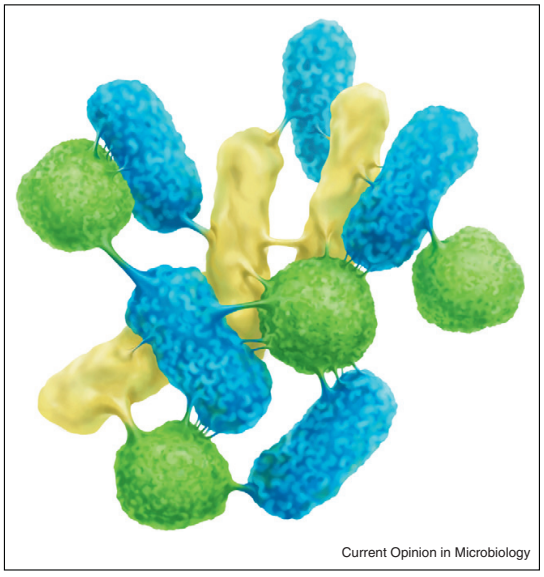
The exact mechanism of DNA transfer via nanotubes is yet unknown. In addition to intracellular molecules, nanotubes conduct also nonconjugative plasmids and even viral particles [26••]. The promiscuity of the nanotubes attachment and their architecture dimensions suggest that they play an important role in all transfer

Figure 1



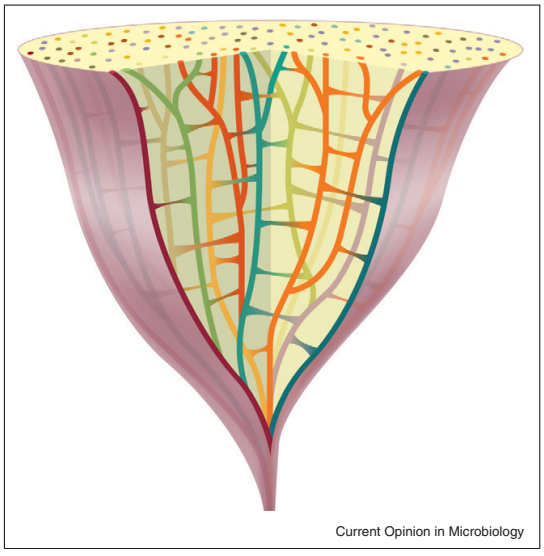
LGT mechanisms. **(a)** The uptake of raw DNA in transformation is enabled during a competence state that involves 20–50 proteins, including the type IV pillus and type II secretion system proteins [11,12]. In some species, an effective transformation requires the presence of uptake signal sequences (USSs; called also DUS: DNA uptake signal). These are specific DNA motifs, about 10 bp long, that are encoded within the recipient genome in a frequency that is much above that expected by chance [17]. Environmental DNA molecules bearing the USS motif are recognized by specific receptors at the cell surface, imported into the cytoplasm, and can then be readily integrated into the recipient chromosomes, usually via homologous recombination [11,12,18,19]. **(b)** During conjugation, plasmids can integrate into the recipient chromosomes by homologous

Figure 2



A schematic representation of cells interconnected by nanotubes. The nanotubes are between 30 and 130 nm wide and up to 1 μm long. The tube dimension is correlated with the distance between the connected cells. Proximal cells are commonly connected by several small nanotubes, while thicker tubes connect distant cells. The rate of transfer via the nanotubes correlates negatively with the size of the transferred substance. Cellular interconnection mediated by nanotubes is not species specific. However, morphology and diameter of the tubes seem to depend on characteristics of the connected cells [26••].

Figure 3



A phylogenetic tree of microbial genomes. Tree branches correspond to genomes and branch colors represent different lineages. Horizontal connections between the branches correspond to LGT events. Bifurcating phylogenetic tree models cannot account for the lateral component of microbial genome evolution [3].

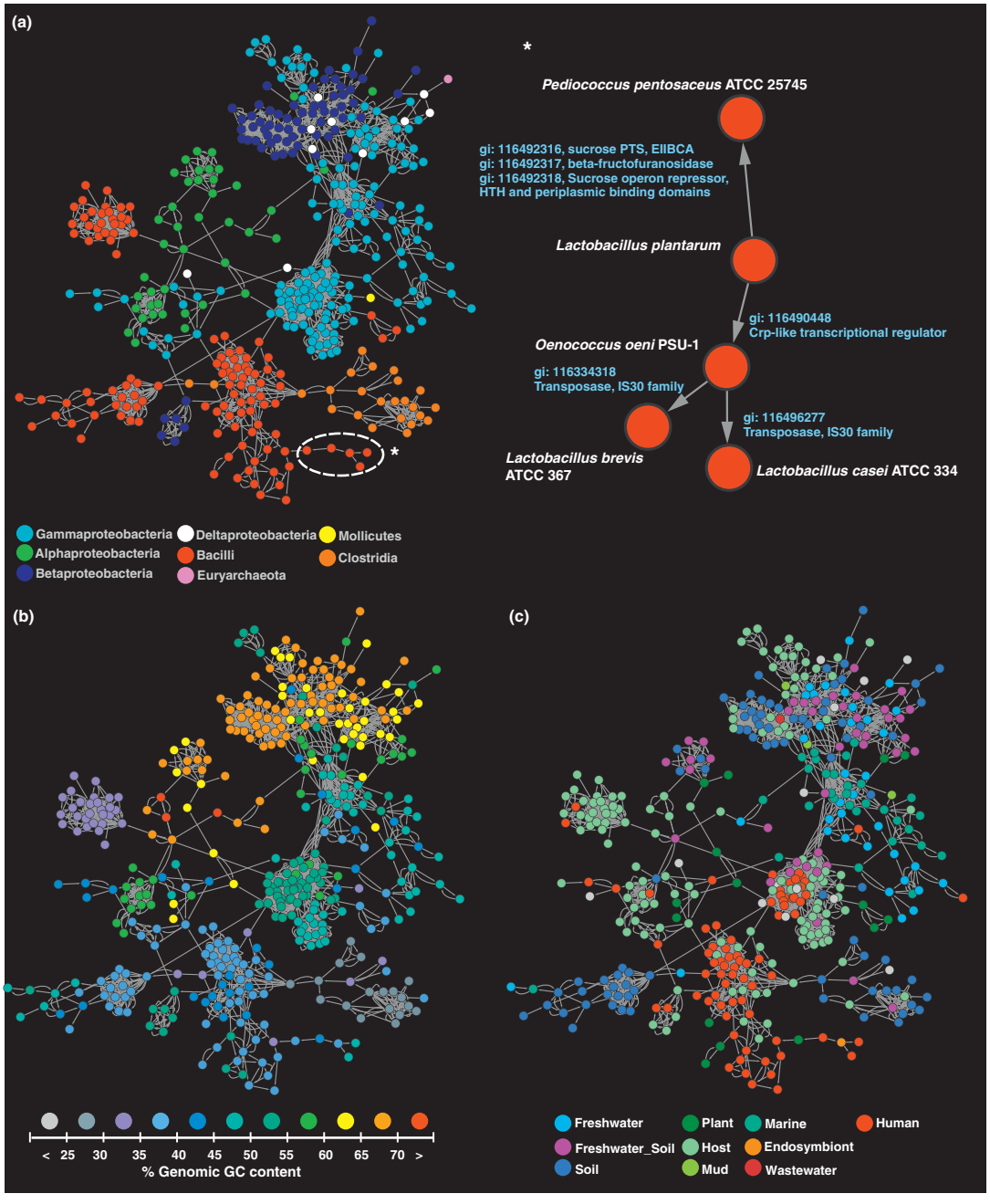
mechanisms, by enabling the propagation of acquired DNA within the population.

We know that LGT occurs in the laboratory, the issue is how often it occurs in the wild and how important it is during evolution. Phylogenetic reconstruction of microbial genes reveals that LGT plays a major role in shaping prokaryotic genomes [5–7,31•,32]. In a pioneering study, Lawrence and Ochman [33] identified all *E. coli* genes that were acquired since its divergence from the *Salmonella* lineage by their aberrant codon usage. They estimated that 755 (18%) of the 4288 genes in *E. coli* strain MG1655 were laterally acquired over a time period of about 14 million years (Myr) and estimated the LGT rate as 16 kb/1 Myr per lineage [33]. Using gene distribution patterns across 329 proteobacterial genomes, Kloesges *et al.* [32] recently estimated that at least 75% of the protein families have been affected by LGT during evolution. Gene transfer rate in those families is on average 1.9 events per protein family per lifespan [32]. Similar estimates were found in phylogenetic analyses of broader taxonomic samples [5–7].

The impact of LGT during genome evolution can be estimated either by the proportion of recently transferred genes whose unusual base composition and codon usage still bears the marks of acquired DNA [33–35] or by phylogenetic analysis of individual genes including recent and ancient LGTs alike (e.g. [36–39]). A survey of genes having aberrant nucleotide composition within proteobacterial genomes revealed that $21 \pm 9\%$ of the genes in those genomes comprises recent acquisitions

(Figure 1 Legend Continued) recombination that may entail insertion sequences (ISs) or other sequences conserved between plasmid and recipient chromosomes that carry the minimal sequence similarity required for homologous recombination [14,20,21]. **(c)** Phages recognize possible hosts by specific receptors found on the cell surface. Many phages include in their genomes chunks of DNA taken coincidentally from previous hosts. These are transferred to the new host during the integration of the phage genome into the host chromosomes. DNA integration into the host chromosome is generally mediated by the phage-encoded enzymes that specifically integrate the phage into the chromosome of the infected recipient [12]. **(d)** DNA stored in GTAs is imported into the recipient in a generalized transduction process mediated by a cellular *RecA* recombination system [22]. GTAs, unlike phages, are linked to transfer of genomic DNA only and GTA-induced cell lysis was not observed [23]. The mechanism of DNA packing and capsule release from the cell is still unknown. GTA systems have been documented not only in oceanic α-proteobacteria, but also in few archaebacteria and some spirochaetes [16,24,25].

Figure 4



A directed network of LGT [49]. The nodes correspond to contemporary or ancestral species that are connected by directed edges of LGT. The edges point from the LGT donor to the recipient. **(a)** Node color corresponds to species taxonomic classification (see legend at the bottom). A cluster of

[32]. Gene distribution patterns across the same species sample suggest that, on average, 74 ± 11% of the genes in each genome have been laterally transferred at least once during evolution [32].

Phylogenetic reconstruction of microbial genome evolution

Lateral gene transfer during microbial genome evolution poses an acute problem to standard phylogenetic reconstruction methodology. Species phylogeny is customarily represented by using bifurcating phylogenetic trees. However, the tree model applies only to the reconstruction of vertical inheritance where genetic material is transferred from ancestral species to their descendants. More realistic models of prokaryotic genome evolution have to embrace lateral gene transfer in addition to vertical inheritance (Figure 3) [2,3,40,41•,42], yet such methods are still scarce. Network models are an alternative to bifurcating trees because they permit the reconstruction and depiction of reticulated evolutionary events such as recombination, gene fusions, and lateral gene transfer [43]. A network is composed of nodes (or vertices) connected by edges corresponding to entities connected by pairwise relations [44,45]. In a phylogenomic network the nodes are completely sequenced genomes and the edges correspond to phylogenetic relations between the genomes that they connect. Phylogenomic networks can be reconstructed from shared gene content (e.g. [6,31•,46]), shared sequence similarity (e.g. [47,48]), or phylogenetic trees [36,49].

The directed network of lateral gene transfer (dLGT) is a phylogenomic network recently developed in order to study the lateral component of recent microbial genome evolution [49]. The nodes in this network correspond to species or their ancestors. The edges represent recent lateral gene transfer events between the species that they connect and they are directed from the donor to the recipient in the LGT event (Figure 4A). Reconstructing a network of recent LGTs has two main advantages: first, the phylogenetic reconstruction of recent transfers is based on a comparison of relatively conserved gene sequences, which is less susceptible to phylogenetic artifacts [53]. Second, bacterial genomes are highly dynamic and may change considerably over time [54,55]; focusing on recent LGTs allows the coupling of the information regarding LGT and current cellular characteristics of donors and recipients. Using the dLGT network one can study trends in – and barriers to – LGT during microbial evolution. In what follows we present some of the insights that this approach permits.

Donor–recipient similarity barrier

Most of the detected LGT in the dLGT network occur between closely related species from the same taxonomic group (Figure 4A) [49]. A graphical representation of the network with species colored by their genomic GC content reveals that clusters of densely connected donors and recipients are very similar in their genomic GC content (Figure 4B). Furthermore, the difference in genomic GC content between donors and recipients is <5% for most (86%) of connected pairs [49]. This suggests that there exists a biological barrier for gene acquisition from donors of dissimilar genomic GC content. Indeed, one such mechanism was discovered in *Salmonella typhimurium* where a histone-like protein (*H-NS*) functions as a transcriptional repressor of GC-poor ORFs [56]. A comparison between the GC content of genes silenced by the *H-NS* repressor (46.8%) and the overall genomic GC content of *S. typhimurium* LT2 (52.2%) reveals that this mechanism is highly sensitive to foreign DNA with lower GC content than that of the genome [56]. However, it is apparent from the dLGT network that some LGT does occur between donors and recipients having difference GC content (Figure 4B). A possible bypass for *H-NS* silencing is provided by the plasmid encoded protein *sfh*, which has been shown to suppress the activity of *H-NS*, enabling the expression of GC-poor ORFs within the genome of *S. typhimurium* [57]. The *sfh* bearing plasmid was isolated from several enteric species, and its DNA sequence is GC-poor. Hence the *sfh* gene allows this plasmid to be transferred among enteric bacteria by escaping from the transcription suppression of the *H-NS* protein [57,58].

To further study the effect of donor–recipient genome sequence similarity on LGT frequency we calculated the similarity between the genomes of connected donors and recipients as the total length of all identical ≥20 bp genomic segments between the two genomes, divided by the recipient genome size. Using this measure we found that the dLGT network is enriched for connected donors and recipients having similar genome sequences [49]. Furthermore, donor–recipient genome sequence similarity and LGT frequency are positively correlated ($r_s = 0.55$, $P \ll 0.01$) [49]. This suggests that LGT is more frequent among closely related species, having similar

(Figure 4 Legend Continued) connected Bacilli (marked with a star) is enlarged to exemplify the network underlying data. Species names are shown next to the nodes. Gene identifier and protein annotation of detected recent gene transfers are noted next to the corresponding edge. The lateral acquisition of genes for sucrose utilization in *Pediococcus pentosaceus* from *Lactobacillus plantarum* has been suggested before [50]. *Oenococcus oeni* (strain PSU-1) – that is associated with malolactic fermentation (MLF) in wine – is connected with three different *Lactobacillus* species as donor and recipient. Bon *et al.* [51] showed recently that gene acquisition from various donors, especially lactic acid Bacilli, contributes to genome plasticity in this species and suggested LGT as a mechanism to enhance *O. oeni* tolerance for harsh wine conditions. **(b)** Node color corresponds to the genomic GC content calculated as the proportion of Guanine and Cytosine (GC) nucleotides within the genome (see scale at the bottom). **(c)** Nodes in the network are colored by habitat (the ten main habitats are listed at the bottom). Habitat annotation was extracted from the GOLD database (ver. 12/2010) [52]. Ancestral nodes are colored by the habitat of their descendants if it is homogeneous or gray otherwise.

genomes, while LGT between distantly related species is more rare [49]. This observation has been recently supported by a study of LGT using simulated genome evolution [59] and through the study of tyrosil-tRNA synthetase phylogeny [60]. High donor–recipient genome similarity in Gram-negative bacteria from the Neisseriales or Pasteurellales orders could be due to frequent gene acquisition by transformation (Figure 1) leading to a high frequency of uptake signal sequences (USSs) in the genomes of both donor and recipient.

The distribution of shared genes across genomes has a strong phylogenetic signal [7] hence there must be some restrictions to DNA acquisition among prokaryotes. Barriers to LGT between distantly related species (having dissimilar genomes) are still poorly understood but are thought to depend on the transfer mechanism. During transformation and conjugation the acquired DNA is commonly integrated into the genome by homologous recombination [12], which requires high similarity between recombining sequences [61,62]. Genes encoded within plasmids that are transferred by conjugation may also be integrated into the recipient chromosome by transposases [13,14] whose function is independent of donor–recipient sequence similarity. However, recent advances in studying the function of the microbial anti-phage CRISPR system (for review see [63•]) revealed that this system also identifies and degrades foreign plasmids in addition to phages [64]. Hence the CRISPR system may function as a barrier for conjugative gene transfer by blocking non-self plasmids [63•]. During transduction, the acquired DNA is integrated into the recipient genome by the phage enzymes whose function is independent of donor–recipient genome similarity [12]. Barriers to phage-mediated gene transfer between distantly related species could be related to the frequency of phages whose host range is species-specific. Such phages indeed exist in marine environments; a test of host specificity for 44 clonal cyanophages revealed that 25 phages were *Prochlorococcus*-specific and 7 were *Synechococcus*-specific, while the remaining 12 (27%) could infect both species [65]. A recent report of antagonistic coevolution between *Pseudomonas fluorescens* and its parasitic phage [66] reveals that species-specific phages exist in terrestrial environments as well.

Ecological barrier

The physical distance between the donor and recipient in the LGT event depends upon the LGT mechanism. In transformation the distance between the donor and recipient depends upon the raw DNA stability within the environment [67]. Conjugation requires that the donor and recipient will be close enough for the formation of the conjugation tunnel. Transduction is considered as the longest range LGT mechanism because it entails phage mobility [67]. This suggests that most transfers should occur within habitats. A graphical representation of the dLGT network with species colored according to their habitat (Figure 4C) reveals that several clusters of the highly connected donors and recipients are variable in their habitat classification, yet, most (74%) of the detected LGT in the network occur between donors and recipients residing in the same habitat. A network of shared transposases among 774 microbial genomes supplies further support for the rarity of inter-habitat gene transfers [68•].

Phylogenomic analyses of microbial genomes support this notion [31•,32]. For example, Halary *et al.* [31•] reconstructed a network of shared protein families among various genetic entities including microbial chromosomes, plasmids, and phage genomes. A comparison of network properties between plasmids and phage genomes revealed that plasmids are more frequently connected within the network in comparison to phages. From this they concluded that conjugation is more frequent than transduction in nature [31•].

Functional barrier

Once imported and integrated into the genome, acquired genes still have to adapt within the genome in order to be retained during evolution. Microbes tend to delete nonfunctional or otherwise unneeded DNA from their genomes [69,70]. Therefore, the fixation of the acquired DNA within the genome is highly dependent on its functionality or utility to the recipient under selectable environmental conditions [54,71,72]. In order to be expressed, the gene has to be either inserted near a recognized promotor, bring one with it, or be acquired together with the corresponding regulator. Hence, acquired genes that are inserted within existing regulatory circuits [73,74] or have a promotor of similar GC content as the recipient genomes [8], have a higher probability to be retained by the recipient.

Gene encoding by suboptimal codons that do not fit the tRNA pool of the recipient has been considered a barrier to LGT [75,76]. However, two recent experimental studies show that the impact of codon usage on the expression and retention of acquired genes might have been overrated. Kudla *et al.* [77] compared the expression level of 154 synthetic gene copies encoding for GFP that varied randomly in synonymous sites (i.e. in their codon usage). The synthetic genes were cloned (i.e. transferred) into *E. coli* and the expression level of GFP was measured by the fluorescence level of the cultures. The result showed that codon usage and fluorescence level of genes are not correlated within the recipient ($r = 0.02$) [77]. Hence the expression level of acquired genes within a recipient cell immediately after the acquisition is independent of their codon usage. In another study, Amorós-Moya *et al.* [78•] compared the fitness of three *E. coli* cultures into which they cloned a chloramphenicol resistance gene encoded by three different codon usage regimes: optimal, GC-rich, and AT-rich. Their results showed that cultures encoding for

the suboptimal gene variants were 10–20 times more sensitive to the antibiotics than those that encoded for the optimal variant. However, within 358 generations of experimental evolution (roughly 54 days) under antibiotic selection, these differences vanished. Interestingly, the compensating mutations were restricted to *in cis* substitutions within the gene promotor or *in trans* substitutions in the host genome, with no substitutions in the gene coding sequence [78•]. In this experiment the new gene acquisition, even of one that is encoded by suboptimal codons, is highly advantageous because the selection regime acts on bacteria that have either low or no antibiotic resistance (i.e. its a matter of life and death). Notably, the evolution of elevated protein expression level in this experiment took place without nucleotide substitutions leading to optimal codons.

Most laterally transferred genes perform metabolic functions while the transfer of genes performing information processing (including replication, transcription, and translation) is rare [49,79,80]. According to the complexity hypothesis [79], the scarcity of lateral transfer of information processing genes is attributed to their role in complex structures. Proteins that function in a complex structure, for example ribosomal proteins, are adapted to their common function. An LGT event that leads to replacement of such a gene with a less adapted homolog will result in a 'squeaking wheel' within the complex and reduced fitness of the recipient [79]. In a recent study, Cohen *et al.* [81] tested the relative impact of functional category and the number of interacting partners on LGT frequency. Their result showed that the complexity hypothesis still passes a reality test in the genomic era. However, LGT barriers owing to multiple interacting partners are not restricted to information processing genes only, but may be observed in all functional categories [81]. Acquisition frequency of metabolic genes depends on their role within the cellular metabolic network [71]. A study of the laterally acquired genes within the *E. coli* metabolic network showed that LGT is more frequent among enzymes involved in peripheral reactions (uptake and metabolism of nutrients) in comparison to those involved in central reactions (biomass production) [71].

Another functional barrier to LGT is protein dosage [8]. A genome sequencing project that includes the preparation of Fosmid libraries may be considered as a large-scale experiment in LGT into *E. coli* [8]. Genomic fragments whose cloning into *E. coli* is lethal are suspects for encoding proteins whose acquisition in *E. coli* is extremely disadvantageous [8]. An extensive dataset of lethal fragments collected during genome sequencing projects of 79 diverse species showed that these fragments typically encode for single copy genes. The integration of an additional gene copy into the *E. coli* genome resulted in an elevated protein production that was lethal to the cell [8].

Conclusions

Experimental work shows that gene acquisition by LGT among prokaryotes is frequent and that the percolation of acquired DNA among populations and across generations is rapid. Phylogenomic analyses reveal that LGT has a substantial impact on long-term genome evolution, supplying a mechanism for natural variation that is specific for the prokaryotic domains and allows their adaptation in dynamic environments. Prokaryote genome evolution comprises thus vertical (tree-like) and lateral (network-like) components. At the same time, different types of barriers to LGT on the genomic, species, and habitat levels are becoming increasingly apparent.

The recent discovery of nanotubes [26••] and the highlight of GTA transfer efficiency [30••] along with their high frequency in natural habitats [25] show that there is still much to discover about LGT. Some of the open questions are: what is the role of nanotubes in LGT? What are the pathways for DNA transfer via the tubes? How are GTAs produced in donor cells? How is the GTA cargo-DNA received in recipient cells? Understanding the mechanisms for gene acquisition in prokaryotes will enrich our understanding of prokaryote genetics in the wild. Moreover, many lab protocols are inspired from what we see in nature. Examples for utilizing LGT in basic research include cell transformation and genome sequencing. More research in this field might contribute to novel developments in synthetic biology. Understanding the barriers to LGT and the way they are bypassed in nature, will improve our ability to manipulate microbial cells in the laboratory for research and industrial purposes.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research and a European Research Council grant NETWORKORIGINS to W. Martin. We are thankful to W. Martin, M. Lercher, K. Stucken, and G. Landan for their help in refining the manuscript.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Dickerson RE: **Evolution and gene transfer in purple photosynthetic bacteria**. *Nature* 1980, **283**:210-212.
2. Doolittle WF: **Phylogenetic classification and the universal tree**. *Science* 1999, **284**:2124-2129.
3. Martin W: **Mosaic bacterial chromosomes: a challenge en route to a tree of genomes**. *Bioessays* 1999, **21**:99-104.
4. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation**. *Nature* 2000, **405**:299-304.
5. 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.

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

7. 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.

8. 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.

9. Weigel LM, Clewell DB, Gill SR, Clark NC, McDougal LK, Flannagan SE, Kolonay JF, Shetty I, Killgore GE, Tenover FC: **Genetic analysis of a high-level vancomycin-resistant isolate of *Staphylococcus aureus***. *Science* 2003, **302**:1569-1571.

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

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

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

13. Norman A, Hansen LH, Sørensen SJ: **Conjugative plasmids: vessels of the communal gene pool**. *Philos Trans R Soc B* 2009, **364**:2275-2289.

14. Wozniak RAF, Waldor MK: **Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow**. *Nat Rev Microbiol* 2010, **8**:552-563.

15. Solioz M, Yen HC, Marris B: **Release and uptake of gene transfer agent by *Rhodopseudomonas capsulata***. *J Bacteriol* 1975, **123**:651-657.

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

17. Smith HO, Tomb JF, Dougherty BA, Fleischmann RD, Venter JC: **Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome**. *Science* 1995, **269**:538-540.

18. Wang Y, Goodman SD, Redfield RJ, Chen C: **Natural transformation and DNA uptake signal sequences in *Actinobacillus actinomycetemcomitans***. *J Bacteriol* 2002, **184**:3442-3449.

19. Snyder LA, McGowan S, Rogers M, Duro E, O'Farrell E, Saunders NJ: **The repertoire of minimal mobile elements in the *Neisseria* species and evidence that these are involved in horizontal gene transfer in other bacteria**. *Mol Biol Evol* 2007, **24**:2802-2815.

20. Weinert LA, Welch JJ, Jiggins FM: **Conjugation genes are common throughout the genus *Rickettsia* and are transmitted horizontally**. *Proc Biol Sci* 2009, **276**:3619-3627.

21. Bahl MI, Hansen LH, Sørensen SJ: **Persistence mechanisms of conjugative plasmids**. *Methods Mol Biol* 2009, **532**:73-102.

22. Genthner FJ, Wall JD: **Isolation of a recombination-deficient mutant of *Rhodopseudomonas capsulata***. *J Bacteriol* 1984, **160**:971-975.

23. Marrs B: **Genetic recombination in *Rhodopseudomonas capsulata***. *Proc Natl Acad Sci USA* 1974, **71**:971-973.

24. Berglund EC, Frank AC, Calteau A, Vinnere Pettersson O, Granberg F, Eriksson AS, Näslund K, Holmberg M, Lindroos H, Andersson SG: **Run-off replication of host-adaptability genes is associated with gene transfer agents in the genome of mouse-infecting *Bartonella grahamii***. *PLoS Genet* 2009, **5**:e1000546.

25. Zhao Y, Wang K, Budinoff C, Buchan A, Lang A, Jiao N, Chen F: **Gene transfer agent (GTA) genes reveal diverse and dynamic *Roseobacter* and *Rhodobacter* populations in the Chesapeake Bay**. *ISME J* 2009, **3**:364-373.

26. •• Dubey GP, Ben-Yehuda S: **Intercellular nanotubes mediate bacterial communication**. *Cell* 2011, **144**:590-600.
Using electron microscopy the authors revealed the existence of tubular extensions bridging between adjacent cells, serving as a route for exchange of cytoplasmic molecules. This study revealed a previously unrecognized type of bacterial communication.

27. Babić A, Lindner AB, Vulić M, Stewart EJ, Radman M: **Direct visualization of horizontal gene transfer**. *Science* 2008, **319**:1533-1536.

28. •• Babić A, Berkmen MB, Lee CA, Grossman AD: **Efficient gene transfer in bacterial cell chains**. *MBio* 2011, **2**:e00027–11.
Using fluorescence microscopy the authors visualized in real time sequential conjugation events of an integrative and conjugative element (ICE) that encodes for a GFP. This study shows that ICEs can propagate rapidly among bacteria that grow in chains.

29. Kenzaka T, Tani K, Nasu M: **High-frequency phage-mediated gene transfer in freshwater environments determined at single-cell level**. *ISME J* 2010, **4**:648-659.

30. •• 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.
The authors quantified the frequency of GTA-mediated LGT by cloning a streptomycin kinase gene into donor bacteria encoding for a GTA cassette and measuring the spread of Kanamycin resistance in the population. This study highlights GTAs as the most efficient LGT mechanism currently known in marine environment.

31. • 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.
Using a network of shared gene content among cellular genomes, plasmids and phages, the authors chart the highways and byways for DNA flow in nature. This study highlights the central role of plasmids as mediators of LGT.

32. 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.

33. Lawrence JG, Ochman H: **Molecular archaeology of the Escherichia coli genome**. *Proc Natl Acad Sci USA* 1998, **95**:9413-9417.

34. Garcia-Vallve S, Romeu A, Palau J: **Horizontal gene transfer in bacterial and archaeal complete genomes**. *Genome Res* 2000, **10**:1719-1725.

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

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

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

38. Puigbò P, Wolf YI, Koonin EV: **The tree and net components of prokaryote evolution**. *Genome Biol Evol* 2010, **2**:745-756.

39. Chan Cheong Xin, Beiko Robert G, Ragan Mark A: **Lateral transfer of genes and gene fragments in *Staphylococcus* extends beyond mobile elements**. *J Bacteriol* 2011 doi: 10.1128/JB.01524-10.

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

41. • Ragan MA: **Trees and networks before and after Darwin**. *Biol Direct* 2009, **4**:43.
In this review the author presents an extensive historical survey of metaphors and models to describe natural systems and genealogical relations over the past 500 years.

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

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

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

45. Newman MEJ: *Networks: An Introduction*. Oxford University Press; 2010.

46. 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.

47. 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.

48. 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.

49. 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.

50. Naumov DG, Livshits VA: **Molecular structure of the locus for sucrose utilization by *Lactobacillus plantarum*: comparison with *Pediococcus pentosaceus***. *Mol Biol* 2001, **35**:19-27.

51. Bon E, Delaherche A, Bilhère E, De Daruvar A, Lonvaud-Funel A, Le Marrec C: ***Oenococcus oeni* genome plasticity is associated with fitness**. *Environ Microbiol* 2009, **75**:2079-2090.

52. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC: **The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata**. *Nucleic Acids Res* 2010, **38D**:D346-D354.

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

54. Hao W, Golding GB: **The fate of laterally transferred genes: life in the fast lane to adaptation or death**. *Genome Res* 2006, **16**:636-643.

55. van Passel MW, Marri PR, Ochman H: **The emergence and fate of horizontally acquired genes in *Escherichia coli***. *PLoS Comput Biol* 2008, **4**:e1000059;
Kuo C-H, Ochman H: **The extinction dynamics of bacterial pseudogenes**. *PLoS Genet* 2010, **6**:e1001050.

56. Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ, Fang FC: **Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella***. *Science* 2006, **313**:236-238.

57. Doyle M, Fookes M, Ivens A, Mangan MW, Wain J, Dorman CJ: **An H-NS-like stealth protein aids horizontal DNA transmission in bacteria**. *Science* 2007, **315**:251-252.

58. Dillon SC, Cameron AD, Hokamp K, Lucchini S, Hinton JC, Dorman CJ: **Genome-wide analysis of the H-NS and Sfh regulatory networks in *Salmonella typhimurium* identifies a plasmid-encoded transcription silencing mechanism**. *Mol Microbiol* 2010, **76**:1250-1265.

59. Puigbò P, Wolf YI, Koonin EV: **Search for a 'Tree of Life' in the thicket of the phylogenetic forest**. *J Biol* 2009, **8**:59.

60. Andam CP, Williams D, Gogarten JP: **Biased gene transfer mimics patterns created through shared ancestry**. *Proc Natl Acad Sci USA* 2010, **107**:10679-10684.

61. Majewski J, Cohan FM: **DNA sequence similarity requirements for interspecific recombination in Bacillus**. *Genetics* 1999, **153**:1525-1533.

62. Lovett ST, Hurley RL, Sutera VA Jr, Aubuchon RH, Lebedeva MA: **Crossing over between regions of limited homology in *Escherichia coli*: RecA-dependent and RecA independent pathways**. *Genetics* 2002, **160**:851-859.

63. • Marraffini LA, Sontheimer EJ: **CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea**. *Nat Rev Genet* 2010, **11**:181-190.
CRISPRs (clustered regularly interspaced short palindromic repeats) are a more recently discovered inherited prokaryotic immune system that protects cells from bacteriophages and conjugative plasmids. In this review the authors elucidate the mechanisms of CRISPR interference and its role in microbial physiology and evolution.

64. Marraffini LA, Sontheimer EJ: **CRISPR interference limits horizontal gene transfer in *Staphylococci* by targeting DNA**. *Science* 2008, **322**:1843-1845.

65. Sullivan MB, Waterbury JB, Chisholm SW: **Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus***. *Nature* 2003, **424**:1047-1051.

66. Gómez P, Buckling A: **Bacteria-phage antagonistic coevolution in soil**. *Science* 2011, **332**:106-109.

67. Majewski J: **Sexual isolation in bacteria**. *FEMS Microbiol Lett* 2001, **199**:161-169.

68. • Hooper SD, Mavromatis K, Kyrpides NC: **Microbial co-habitation and lateral gene transfer: what transposases can tell us**. *Genome Biol* 2009, **10**:R45.
Using a network of shared transposases the authors reveal that the majority of LGTs occur between closely related species residing in the same habitat. This study demonstrates the utility of genomic data for research in microbial ecology.

69. Moran NA, McLaughlin HJ, Sorek R: **The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria**. *Science* 2009, **323**:379-382.

70. Burke GR, Moran NA: **Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids**. *Genome Biol Evol* 2011, **3**:195-208.

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

72. Zhaxybayeva O, Doolittle WF: **Lateral gene transfer**. *Curr Biol* 2011, **21**:R242-R246.

73. Davids W, Zhang Z: **The impact of horizontal gene transfer in shaping operons and protein interaction networks – direct evidence of preferential attachment**. *BMC Evol Biol* 2008, **8**:23.

74. Dorman CJ: **Regulatory integration of horizontally-transferred genes in bacteria**. *Front Biosci* 2009, **14**:4103-4112.

75. Medrano-Soto A, Moreno-Hagelsieb G, Vinuesa P, Christen JA, Collado-Vides J: **Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes**. *Mol Biol Evol* 2004, **21**:1884-1894.

76. Tuller T, Girshovich Y, Sella Y, Kreimer A, Freilich S, Kupiec M, Gophna U, Ruppin E: **Association between translation efficiency and horizontal gene transfer within microbial communities**. *Nucleic Acids Res* 2011, **39**:4743-4755.

77. Kudla G, Murray AW, Tollervey D, Plotkin JB: **Coding-sequence determinants of gene expression in *Escherichia coli***. *Science* 2009, **324**:255-258.

78. • Amorós-Moya D, BedhommeS, Hermannand M, Bravo IG: **Evolution in regulatory regions rapidly compensates the cost of nonoptimal codon usage**. *Mol Biol Evol* 2010, **27**:2141-2151.
Using experimental evolution approach the authors demonstrate that non-optimal codon usage is a minor obstacle for the fixation of laterally acquired genes. The regulation of gene expression level in this experiment was independent of codon adaptation.

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

80. Coscollá M, Comas I, González-Candelas F: **Quantifying nonvertical inheritance in the evolution of *Legionella pneumophila***. *Mol Biol Evol* 2011, **28**:985-1001.

81. 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.

Chapter III

Phylogenomic transduction networks reveal genetic barriers to phage-mediated lateral gene transfer during microbial evolution

Chapter III

Biological sciences: Evolution

Phylogenomic transduction networks reveal genetic barriers to phage-mediated lateral gene transfer during microbial evolution

Ovidiu Popa, Giddy Landan, Tal Dagan

Genomic Microbiology Group, Institute of General Microbiology, Christian-Albrechts Universtiy of Kiel, Am Botanischen Garten 11, 24118 Kiel, Germany.

Corresponding author: Tal Dagan, Genomic Microbiology Group, Institute of General Microbiology, Christian-Albrechts Universtiy of Kiel, Am Botanischen Garten 11, 24118 Kiel, Germany. Tel: +49 431 8805712. Email: tdagan@ifam.uni-kiel.de

Keywords: horizontal gene transfer, phage-bacteria coevolution, gene duplication

Abstract (250 words)

Bacteriophages are recognized DNA vectors and transduction is considered as a common mechanism of lateral gene transfer (LGT) during microbial evolution. Anecdotal events of phage-mediated gene transfer were studied extensively, however, a coherent evolutionary viewpoint of LGT by transduction, its extent and characteristics, is still lacking. Here we report a large-scale evolutionary reconstruction of transduction events in 3,982 genomes. We inferred 17,158 recent transduction events linking donors, phages and recipients into a phylogenomic transduction network view. We find that LGT by transduction is mostly restricted to closely related donors and recipients. Furthermore, a substantial number of the transduction events (9%) are best described as gene duplications that are mediated by mobile DNA vectors. We propose to distinguish this type of paralogy by the term *autology*. A comparison of donor and recipient genomes revealed that genetic similarity is a superior predictor of species connectivity in the network in comparison to common habitat. This indicates that barriers for transduction during microbial evolution are largely genetic while ecological factors are secondary. A striking difference in the connectivity pattern of donors and recipients shows that while lysogenic interactions are highly species-specific, the host range for lytic phage infections can be much wider, serving to connect dense clusters of closely related species. Our results thus demonstrate that DNA transfer via transduction occurs within the context of phage-host specificity, but that this tight constraint can be breached, on rare occasions, to produce long range LGTs of profound evolutionary consequences

Significance statement (120 words)

Prokaryotes can acquire genes from unrelated species via lateral gene transfer (LGT). Viruses that infect bacteria – phages – are known vehicles of DNA transfer by transduction, and are considered as a common LGT mechanism. A coherent evolutionary viewpoint of phage-mediated LGT, its extent and characteristics, is however still lacking. Here we report a large-scale evolutionary reconstruction of 17,158 transduction events linking donors, phages and recipients into a phylogenomic transduction network view. We find that transduction occasionally leads to LGT between distantly related bacteria. The vast majority of transductions, however, are restricted to genetically similar bacteria that interact with similar phages. Moreover, transduction often operates within the species, leading to gene duplication. Transferred genes include antibiotic resistance genes and other defence mechanisms.

\body

Introduction

DNA transfer is an important mechanism for natural variation in the prokaryotic domains (1, 2). Recombination at the species level plays a role in selective sweeps through the population (3) while lateral gene transfer across species boundaries has important implications to microbial adaptation and evolutionary transitions (e.g. (4)). Viruses that infect bacteria – termed phages – are known vectors of DNA transfer between microbial cells (5, 6). Temperate (or lysogenic) phages multiply via the lysogenic cycle, which is established by an integration of the phage genome into the host chromosomes, creating a prophage within the host genome. The phage typically remains dormant within the host and is replicated with the host genome until the lytic cycle is induced. In the lytic cycle new phages are produced using the host metabolism and are released during the host cell lysis (7). The excision of phage DNA from the host genome and the production of phages may be accompanied by packing of host DNA into the phages, which can then transfer it to the next host in a process that has

been termed transduction (5). Specialized transduction occurs when the phage integrases cleave, in addition to the prophage, bacterial genes that are encoded at the prophage flanking regions. These are packed with the phage DNA into the phages. Generalized transduction occurs when random bacterial DNA is packed into the phages (8). A recent analysis of enterobacterial genomes revealed an extensive domestication of genes of viral origin. The gene adaptation process is accompanied by a rapid prophage inactivation followed by a gradual genetic degradation that is marked by a strong purifying selection on the acquired gene sequence followed by their vertical inheritance within the lineage (9).

The frequency of transduction in nature has been estimated to range between $1.33x10^{-7}$ and $5.33x10^{-9}$ transductants/PFU in marine environments (10). A higher transduction frequency ranging between $0.3x10^{-3}$ and $8x10^{-3}$ transductants/PFU was observed in a freshwater environment, where the DNA transfer frequency was estimated to be up to $2x10^{-3}$ transductants/PFU with 20% of the gene-recipients retaining their viability (11). Phage lethality, as measured by the ratio of phage infection to adsorption, and host specificity may however differ between various phage taxa. For example, cyanophages of diverse taxa are highly host-specific and their interaction is characterized by 100% lethality while heterosiphoviruses have been shown to be adsorbed by a wide range of Pseudoalteromonas strains and their lethality ranges between 10 and 40% (12). The realized host range in the wild is however determined not only by the host permissibility but also largely by phage-bacteria co-occurrence in the same geographic habitat (13).

Bacteria and their parasitic phages are co-evolving in a constant arms race, yet their interaction may include also mutualistic aspects. The beneficial contribution of phage-mediated gene transfer to the host fitness has been documented in diverse environments (14). For example, genomes of phages that infect marine cyanobacteria have been found to encode components of both photosystem I (15) and photosystem II (16). The elevated dose effect of these gene products within the host is assumed to

increase the photosystems recycling efficiency and by that compensate for the energetic cost of phage proliferation (16). Recently sequenced metagenomic samples from hydrothermal vents revealed a high abundance of phages that encode components of the dissimilatory sulfite reductase gene (*rdsr*) (17). This gene is essential for sulfur-oxidation and may confer an energetic advantage to chemolythoautotrophic bacteria that typically inhabit such environments. In addition to the transfer of metabolic functions between closely related hosts, phages have been found to mediate intergeneric gene transfer across species boundaries as exemplified in the transfer of toxin genes between *Staphylococcus aureus* and *Listeria monocytogenes* in raw milk (18).

Bacterial genomes that include a prophage may be considered as recipients in gene transfer events. Bacterial genes in prophages are the result of gene acquisition by transduction and their origin can be identified by homology and phylogenetic analysis. Here we study the extent of phage-mediated gene transfer during microbial evolution using networks approach. The networks are composed of donors, phages, and recipients that are connected by recent transduction events reconstructed from genomic data. Structural properties of the network supply a large-scale view of barriers for transduction and gene transferability by phages in nature.

Results

The transduction network. To study the general properties of LGT by transduction we combine individual donor-recipient inferences into a network representation. Transduction events are characterized by two distinct phases: the uptake of a gene from a donor into a phage and the acquisition of a gene as part of a prophage by the recipient. We constructed a directed lateral gene transfer (dLGT) network that includes two types of entities: bacteria and phages. A directed edge from a phage node to a bacteria node designates a gene acquisition following transduction as inferred from the prophage annotation where the bacteria node is the recipient. A directed edge from a

bacteria node to a phage node specifies the acquired gene origin as inferred from the phylogenetic analysis where the bacteria node is the donor. We analysed a total of 2,103 finished and 1,879 draft microbial genomes including a total of 9,468 annotated complete and partial prophages (19). Applying conservative sequence similarity thresholds in the different inference stages, our approach identified a total of 17,158 transduction events where donor and recipient are specified. Constructing the network from those events where a single most likely donor was identified yielded a directed LGT network (dLGT) (20), comprising 2,573 bacteria and 4,650 phage nodes that are connected by 15,298 edges summarizing all 17,158 transduction events (Fig. 1A; Table S1). Edge weight in the network is calculated as the total number of genes that where transferred between the bacteria and phage nodes.

The dLGT network comprises a large component of 4,982 nodes, including 1,538 bacteria and 3,444 phages. The remaining nodes in the network fall into 326 smaller clusters including, on average, 3 bacteria and 4 phages. For example, the *Natrialba magadii* ϕCH1 virus has a temperate interaction with the chemoorganotrophic euryarchaeon *N. magadii* isolated from Magadi lake in Kenya (21). The virus encodes a total of 24 bacterial genes. Our inference algorithm yielded putative donors for two of those genes. One gene, annotated as a hypothetical protein, was putatively acquired from *Halobiforma lacisalsi*, an extreme halophylic archaeon. The second gene, annotated as a gas vesicle protein, was putatively acquired from *Natronobacterium gregoryi*, a haloalkaliphilic Euryarchaeon (Fig. 1B). In another cluster we identified orthologous prophages that are encoded within the genomes of two chloroflexi strains: *Chloroflexus aurantiacus* J-10-fl and *Chloroflexus* sp. Y400-fl. The two prophages have a 100% match of their protein content when applying a sequence similarity threshold of 95% identical amino acids, thus they are considered as orthologous prophages. One of the eleven bacterial genes encoded in this prophage is annotated as a threonine synthase and was putatively acquired by the phage from *Chloroflexus aggregans* (Fig.

1C). This small cluster exemplifies how phage-mediated laterally transferred genes can enter the lineage.

Phage-mediated gene duplication – autology. While most of the genes acquired via prophages are xenologs, the network reveals a substantial number of genes where the recipient genome is also the donor (e.g., Fig. 1D). Thus, per definition these genes are paralogous rather than xenologous genes. We suggest terming such genes *autologs*. According to our definition, an autologous gene is the result of gene duplication that is mediated by a mobile DNA vector where the donor is also the recipient. Our analysis revealed 1,550 (9%) autologous genes that are distributed over 543 (21%) microbial genomes. About half of the gene duplications in the network are of a single gene and up to a maximum of 48 genes in *Magnetococcus* sp. MC-1. Of the self-donor recipients, 72% are connected to a single phage. A maximum of 9 phages are connected by a self-edge to *Methylobacterium nodulans* ORS_2060 (Fig. S1).

Of the 1,550 autologs, 697 have no nucleotide substitutions at all, whereas the remaining 54% autologs show the hallmarks of gene duplications. They contain significantly more synonymous substitutions than non-synonymous ones ($p<10^{-15}$, using paired-Wilcoxon test). The median d_N/d_S ratio (ω) is 0.12, which is significantly larger than the observed for the *bona fide* gene acquisitions (ω=0.09, $p=3\times10^{-13}$, using Wilcoxon test). Moreover the codon adaptation index (CAI) is significantly smaller for the prophage gene than the genomic copy (p=0.035, using paired-Wilcoxon test). These observations are consistent with the observed relaxation of purifying selection in *entrobacterialles* prophages (9). The high frequency of autologous genes is best understood in the context of phage host-specificity and recurrent infection of the same lineage.

Donor and recipient components. Structural properties of the dLGT network are the result of two different phage-bacteria interaction modes. Phages connected to

recipients represent a stable interaction that involves temperate phages and their hosts. Links between donors and phages are evidence for a transient interaction that is typical to lytic phages, where donors connected to the same phage designate the putative hosts of that phage. The network thus combines two components – edges that link donors to phages and edges that connect phages to recipients – corresponding to uptake and acquisition events respectively. Large-scale structural differences among the two components reveal the differential contribution of lytic and temperate phage-bacteria interactions to transduction dynamics during microbial evolution. The donor and recipient parts of the network, termed here D-dLGT and R-dLGT respectively, comprise a similar number of bacteria and phage nodes (Table S1). Yet, the node connectivity degree is significantly larger in the D-dLGT in comparison to the R-dLGT for both phage and bacteria nodes (D-dLGT: $p < 10^{-15}$, R-dLGT: $p < 10^{-15}$, using Kolmogorov-Smirnov test; Fig. 2). Consequently the D-dLGT nodes are more densely interconnected in comparison to nodes in the R-dLGT network. Most (93%) of the phages in the R-dLGT are connected to a single recipient node and at most to eight recipients (Fig. 2). Only 46% (2,131) of phages in the D-dLGT network are connected to a single donor node, while 25% (1,146) phages are connected to two donors and the remaining 29% (1,373) phages are connected to three donors or more (Fig. 2).

It is noteworthy that highly connected phages include genes that have the potential to be beneficial for the recipient. The most connected phage in the D-dLGT network encodes 29 genes of bacterial origin for which we identified 20 *Enterobacterialles* donors (Fig. S2; Table S2; PhageID:10223). The phage encodes the MazE/F toxin-antitoxin (TA) system that can mediates cell growth arrest and was shown to increase the persistence and survival of *Escherichia coli* under antibiotic stress (22). Our analysis further uncovered the transfer of 73 TA genes mediated by 32 phages (Table S3). These transduction events suggest that phages may encode for addiction-mechanisms similarly to plasmids. Another highly connected phage in the D-dLGT network is connected to 19 Enterobacteriales species (Fig. S2; Table S2;

PhageID: 11150). The phage encodes for MdtH, a multidrug resistance gene that confers resistance to norfloxacin and enoxacin (23). Transferred genes in the network include additional 46 genes coding for a broad range of antibiotic resistance (Table S4) demonstrating a putative role of phages in the spread of antibiotics resistance. The most connected phage in the R-dLGT is connected to eight *Bacillus* recipients (for details see Fig. S2; Table S2; PhageID:5008). The phage contains eight genes of bacterial origin. One of those, bclA, encodes for a spore surface glycoprotein in *B. anthracis* (24).

The different bacteria and phage connectivity pattern of the R-dLGT and D-dLGT is evident also in their global structure. The D-dLGT contains significantly less connected components in comparison to the R-dLGT. Furthermore, nodes in the recipient network are clustered into significantly smaller components in comparison to the donor network ($p=7\times10^{-10}$, using Kolmogorov-Smirnov test) and the number of nodes in the D-dLGT largest component is 25-fold larger in comparison to that of the R-dLGT largest component (Table S1). In consequence, edge weights in the D-dLGT are significantly lower in comparison to the R-dLGT ($p<10^{-15}$, using Kolmogorov-Smirnov test) with medians of single gene per donor edge and two genes per recipient edge (Table S1). The R-dLGT comprises a total of 230 (4.6%) edges with an edge weight ≥10. In the D-dLGT network, for comparison, we observe only 79 (0.77%) edges having an edge weight ≥10 (Fig. S3).

The different structural properties of the donor and recipient network components suggest that gene uptake from transient hosts into the phage genome typically include a single gene, while gene transfer into stable hosts usually comprises several genes. Yet, the transient interactions constitute an important contribution to the global network structure by connecting among clusters of stable hosts.

Host range in the transduction network. Phages that are linked to more than one donor or recipient in the network supply an insight into the phage host range. In the D-

dLGT component about half of the phages (2,519; 54.17%) are connected to multiple donors. Most of these phages are connected to donors of the same species (1,329, 53%), or genera (782, 31%) revealing a very narrow taxonomic host range at the donor side (Fig. 2). Only 22 phages in the D-dLGT network are connected to two donors that are members of different phyla, 20 of which are connected to firmicutes strains (Table S5). A single phage is connected to three donors from different phyla including *Bacteroides* sp. 3_1_33FAA and *Clostridium* sp. M62/1 that were isolated from the human gastrointestinal tract and *Cardiobacterium hominis* ATCC 15826 (Gammaproteobacteria) that was isolated from the human cardiovascular system (25) (Table S5; PhageID: 9283).

Phages connected to more than a single recipient in the R-dLGT network (333; 7%) show even stronger species-specificity, with most phages (261, 78%) connected to recipients classified into the same species (Fig. 2). A total of 57 (17%) phages are linked to recipients from different species within the same genus (Fig. S2; Table S2; phageIDs: 5548, 5273). Only 11 (3.3%) phages are found in recipients of different genera within the same taxonomic order. The rare inter-generic transduction events include a phage connected to two Clostridiales recipients: *Blautia hansenii* DSM 20583 and *Ruminococcus gnavus* ATCC29149 (Fig. S2; Table S2; PhageID: 5915). Both strains were isolated from the human digestive system (25), hence they probably share a common habitat.

A single phage links two recipients from different classes within the Firmicutes phylum, *Clostridium M62-1* and the *Lactobacillus ruminis* ATCC-25644, both isolated from the human gastrointestinal tract (25) (Fig. S2; Table S2; PhageID: 6299). Only two phages link to recipients from different phyla (Fig. S2; Table S2; PhageIDs: 5805, 6260). One of those connects the *Bifidobacterium pseudocatenulatum* DSM 20438 (Phylum: Actinobacteria) and *Parvimonas micra* ATCC 33270 (Phylum: Firmicutes), both isolated from the human gastrointestinal tract (25) .

The narrow taxonomic range of multiple donors and recipients observed in the network components is in agreement with experimental observations of phage species-specificity (26) and is expected from the tight phage-host co-evolutionary dynamics. Our results reveal however several genomic footprints of rare cross-species infections. Many of these examples are observed in microbial genomes sequenced as part of the human microbiome project, thus it is possible that the high sampling density of that habitat facilitated that recovery of those rare interactions.

Barriers for gene transfer by transduction. The majority of phages (2664, 64%) connect donors and recipient from different strains of the same species (Fig. 2). These phage-mediates DNA transfers are best viewed as genetic recombination rather than lateral gene transfer events. A Siphoviridae phage connected to multiple *Vibrio cholera* strains illustrated this phenomenon (Fig. S2; Table S2; PhageID: 8390). The phage encodes the *nqr* operon that has an important function in the bioenergetics and homeostasis of *V. cholerae* (27). The frequency of observed LGTs decreases dramatically when the donor-recipient taxonomic separation increases (Fig. 2). At the inter-domain level, only a single phage was observed, connecting *Methanobrevibacter smithii* DSM 2374 as the recipient with *Bacillus cereus* Rock3-28 as the donor (Fig. S2; Table S2; PhageID: 9888). The recipient strain was isolated from human feces (25), whereas the donor was isolated from the soil (28). The prophage includes a gene encoding for tetracyclin resistance that has 100% identical amino acids to the gene encoded in *B. cereus*. To our knowledge, this is the first genomic evidence for transduction of an archaebacterium by a eubacterial bacteriophage; hence, this putative inter-domain transfer represents a very exceptional event.

Barriers for transduction may be related to the genetic requirements for a successful gene acquisition and the ecological co-occurrence of the connected partners. In contrast to transduction, in transformation and conjugation the integration of acquired DNA into the recipient genome is mediated by homologous recombination

and therefore depends on sequence similarity between the donor and recipient (29). During transduction, however, the acquired DNA is integrated into the recipient genome using the phage mechanism (29), hence no such dependency is expected. To test for genetic barriers to DNA transfer by transduction we calculated the genome similarity between donors and recipients using four measures. Genome similarity (S_{GS}) is calculated as the Jaccard index of identical ≤20bp sequences between the donor and recipient genomes. Coding sequence similarity (S_{CDS}) is calculated similarly but is restricted to protein coding sequences. Codon usage distance (D_{CU}) is calculated as the Euclidean distance between the relative codon frequencies within the donor and recipient genomes. GC content similarity (S_{GC}) is calculated from the genomic content of Guanine and Cytosine in the donor and recipient genomes. The distribution of all similarity measures was compared between the dLGT network and a set of 1,000 networks where the edges have been randomly shuffled.

We find that donors and recipients connected in the dLGT network are significantly more similar to each other than expected by chance using all similarity measures (Fig. 3). The four similarity measures are correlated – closely related genomes will score high on each measure, yet it is of interest to grade their importance as barriers to LGT. To this end, we consider each pairwise similarity measure as a predictor of the connectedness state of the pair of species, and conduct a receiver operating characteristics (ROC) analysis (e.g. (30)). We find that genome similarity is the best predictor for dLGT connectedness, with an area under the ROC curve (AUC) of 0.99, and an optimal discrimination of 0.97 true-positive rate (TPR) and 0.03 false-positive rate (FPR). The next best measure is codon usage distance (AUC 0.98; TPR 0.93; FPR 0.03), followed closely by coding-sequence similarity (AUC 0.96; TPR 0.94; FPR 0.04). GC content similarity is an inferior predictor in comparison to the other measures (AUC 0.95; TPR 0.87; FPR 0.08). Our results demonstrate that low donor-recipient genome similarity is an important barrier that constrains the extent of LGT via transduction.

Another possible barrier for transduction is the need for ecological co-occurrence of donors, phages and recipients. This barrier may be partially breached by phage mobility that is thought to enable the transfer of genetic material between donors and recipients across a larger spatial separation compared to other LGT mechanism that are dependent of physical proximity (31). Donor-recipient pairs share the same habitat in 3,330 (44%) cases, of which the majority (1,383, 41%) are members of the "human associated" habitat group. In the remaining 4,187 (56%) donor-recipient pairs classified in different habitat groups (cross-habitat transfer events), we observed the majority (858, 20.49%) of links between the donor group "host" and the recipient group "human associated" bacteria (Fig. S4). To evaluate whether these values are different from the expectation given that habitat sampling is heavily skewed towards certain habitats, we estimated the expected within-and cross-habitat frequencies from 1,000 randomized dLGT networks. Links between donors and recipients from the same habitat are significantly overrepresented in the dLGT network, with the corollary that most cross-habitat links are occurring at a lower frequency than expected. However, some habitats do show a higher than expected cross-habitat LGT frequencies (Fig. S4). For example we found 51 (expected 17) links between "soil and sediment" and the "plant" group and 41 (expected 19) transfers between the habitat "plant" and "soil and sediment" group. 48% of these transfers are intra-specific and 95% are intra-generic. Indeed, habitat sharing is only a weak predictor of species connectedness, with equivalent AUC of only 0.64 (TPR 0.44; FPR 0.17). Our analysis thus reveals that the barriers for gene transfer via transduction are primarily genetic while ecological barriers play a smaller role.

Functional classification and evolutionary constraints. The functional composition of dLGT genes is significantly different than that of the analysed genomes ($p<10^{-15}$, using χ^2 test). Information processing functions are overrepresented in the network, while cellular processes and metabolism functions are depleted (Fig. S5). Of the genes

that could be classified into putative functions (2,274, 13%), 42% perform metabolism functions, whereas 35% were involved in information processing; most of those are annotated as transcription genes. Another 23% of the genes were classified into cellular processes, with a majority of cell wall and membrane biogenesis function (Fig. S5). Interestingly, information genes are transferred between less similar donors and recipients than the other functions, while metabolism and cellular processes genes are transferred between equally similar donors and recipients (α=0.05, using Tukey test). This observation may be attributed to the universality of information processing genes.

The nucleotide substitution pattern of genes in the transduction network indicates that their acquisition was very recent or that they evolve under extremely strong purifying selection. Half (52%) of the donor-recipient pairs have no nucleotide substitutions at all. Comparing the rate between the donor and recipient lineages for the remaining 48%, we observe a very slight and not significant increase in the recipient lineage rates (Table S6). The ω ratio is also not significantly different between the two lineages and in 95% of the genes is below 0.5 in both lineages. Together these observations suggest that the strength of purifying selection in recipient lineages remains similar to that in the donor lineages with no apparent relaxation of selective constraints or nonfunctionalization. The great majority (95%) of bacterial genes that are encoded in prophages are single-copy genes, that is, there is no pre-existing homologous gene in the recipient genome. Taken together with the evidence for gene functionality, this suggests that most transduction events result in an acquisition of a new function. Furthermore it could indicate that the accessibility of the host to the new function is maintained as long as the stable interaction with the phage is maintained.

Discussion

Here we study the contribution of phage-mediate gene transfer to microbial genome evolution. The transduction network reconstruction revealed a substantial frequency of autologs. Our results add support to previous studies showing that protein family

expansion in bacteria is mediated more often by LGT than by gene duplication (32, 33). Indeed, we have to assume that low sampling density may obscure a gene donor among closely related strains, whose genome has not been sequenced yet. Nevertheless, the high sequence similarity and lack of alternative homologs besides the recipient genomic copy indicate that autologs originate from within the pan-genome.

The topological differences between the donor and recipient network components suggest that host-specificity is much more prevalent in temperate interactions and that phages have a broader host range for lytic interactions. Because lysogenic phages are highly dependent on the host cellular processes (e.g. (34)) it is likely that closely related strains having a similar genetic background can have a temperate interaction with the same phage. Moreover, the multiplicity of donors and different edge weight distributions imply that phages sporadically accumulate bacterial genes during lytic interactions, probably one gene (or very few genes) at a time, whereas gene acquisition at the recipient side constitutes a simultaneous acquisition of many genes. Previous studies of LGT dynamics inferred that most LGT events involve very few genes while bulk transfers are relatively rare (20, 35). However, taking the transfer mechanism into account reveals that phages can mediate bulk LGTs. Hence, transduction is characterized by a significant addition of genes into the recipient genome within a single transfer event.

The dLGT network reveals the existence of strong taxonomic and genetic barriers for phage-mediated lateral gene transfer. Previous studies advanced the view that gene transfer during microbial evolution is largely determined by ecological rather than phylogenetic factors (36). While we do find an over-representation of transfers within habitats, habitat sharing is only a weak predictor of species connectedness and is much inferior to all sequence derived similarity measures. The significant high codon usage similarity of donors and recipients is consistent with previous observations of non-random codon usage in phage genomes, leading to the suggestion that phage

codon usage is adapted to that of the host (37, 38). Previous studies have emphasized the importance of codon usage similarity for LGT, suggesting that high codon usage similarity between acquired genes and the recipient genome will increase the xenolog retention prospects (39, 40). Our results reveal that genes acquired via transduction are expected to comprise a similar codon usage to that of the recipient; hence, the translational barrier for their adaptation is expected to be rather low. Several temperate phages have been reported to encode genes that are transcribed independently from the prophage excision mechanism (41). Thus, the transcriptional regulation of genes acquired via transduction is likely to be promoted by the prophage encoded promoters so that genes acquired by transduction are functional upon acquisition.

In summary, our results demonstrate that LGT via transduction occurs within the tight constraints of phage-host specificity. Consequently, transduction is probably more important for genetic recombination, and selfing in the case of autologs, within the species rather than for long-range gene transfers between distinct lineages. LGT is commonly viewed as a source for reticulated events that reduce the tree signal during prokaryotic evolution (1). Our current results show that the reticulated events introduced by transduction affect mostly clades of closely related species and very rarely do they traverse the tree and disrupt its global topology.

Methods

Data. Annotation of 14,920 prophages encoded in 8,540 genomic sequences was downloaded from PHAST database (ver. 10/2012) (19). Genomes of 2,103 complete and 1,879 draft prokaryotic genomes were downloaded from GenBank (NCBI; ver. 10/2012). PHAST entries not found in GenBank were discarded. This resulted in 9,468 annotated prophages encoded in 1,330 complete and 1,281 draft genomes. Coding sequences (CDSs) in PHAST database are classified into viral or bacterial if they have a significant BLAST hit within viral or bacterial genomes respectively (19). Prophages encoding only phage genes were excluded. The remaining 9,201 (97%) prophages

encode a total of 281,616 CDSs, of which 89,234 (32%) are classified as bacterial genes. Prophages were clustered into orthologous prophage clusters (42) using 70% shared gene families as a threshold (see supplemental methods for details). This yielded 2,397 orthologous prophage clusters and 8,426 singleton phages.

Donor inference. The donor inference procedure operates within the framework of orthologous protein families and is assisted by a phylogenetic tree. In the first step we created clusters of orthologous proteins using the bacterial genes encoded within prophages as queries for a BLAST search against the GenBank genomes, and identified 3,908,830 homologous sequences to 75,172 of the query genes, whereas no homologs were detected for the remaining 14,197 (15.89%) queries. The protein sequences were clustered into orthologous protein families to yield 20,904 protein families with at least two proteins. Protein clusters containing at least three protein sequences (12,611) were aligned using MAFFT (43) and maximum likelihood trees were reconstructed using PhyML3 (44) with the LG model. For each gene acquired by transduction, we identified the most likely gene-donor as the genome bearing a homologous gene which is the sister clade of the acquired gene in the phylogenetic tree or the unique homolog in the case of clusters with two members.

Network construction. Donor-recipient relations were coded into the directed lateral gene transfer (dLGT) unipartite network, in which nodes represent bacterial species and edges lateral gene transfers mediated by transduction. Bacteria-phage relations were coded into a bipartite directed network where nodes represent either phages (5064 nodes) or bacterial species (3982 nodes). This enables partitioning of the network into two subsets; the stable subset (R-dLGT) consists of directed edges from prophages to their host bacterial species (i.e., the recipients), and the transient subset (D-dLGT) that consists of directed edges from donor bacterial nodes to phages (i.e. the

transfer vector). As in the dLGT, edge weights correspond to the number of transferred genes.

Additional details of and the description network randomization, characterization of genomes by similarity measures and habitat, and characterization of genes by substitution rates and functional classification, are found in supplemental methods.

Acknowledgements

The authors thank Anne Kupczok, Nils Hülter and Elie Jami for their helpful comments on the manuscript and acknowledge support from the European Research Council (Grant No. 281357).

References

1. Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science* 284(5423):2124–2129.

2. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784):299–304.

3. Shapiro BJ, et al. (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science* 336(6077):48–51.

4. Nelson-Sathi S, et al. (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517(7532):77–80.

5. Zinder ND, Lederberg J (1952) Genetic exchange in Salmonella. *J Bact* 64(5):679–699.

6. Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E (2010) Network analyses structure genetic diversity in independent genetic worlds. *Proc Nat Acad Sci* 107(1):127–132.

7. Campbell A (2003) The future of bacteriophage biology. *Nat Rev Genet* 4(6):471–477.

8. Weinbauer MG (2004) Ecology of prokaryotic viruses. *FEMS Microbiol Rev* 28(2):127–181.

9. Bobay L-M, Touchon M, Rocha EPC (2014) Pervasive domestication of defective prophages by bacteria. *Proc Nat Acad Sci* 111(33):12127–12132.

10. Jiang SC, Paul JH (1998) Gene transfer by transduction in the marine environment. *Appl Env Micro* 64(8):2780–2787.

11. Kenzaka T, Tani K, Nasu M (2010) High-frequency phage-mediated gene

transfer in freshwater environments determined at single-cell level. *ISME J* 4(5):648–659.

12. Deng L, et al. (2012) Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as revealed by viral tagging. *mBio* 3(6).

13. Flores CO, Valverde S, Weitz JS (2013) Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *ISME J* 7(3):520–532.

14. Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann M-L, Brüssow H (2003) Phage as agents of lateral gene transfer. *Current Opin Micro* 6(4):417–424.

15. Sharon I, et al. (2009) Photosystem I gene cassettes are present in marine virus genomes. *Nature* 461(7261):258–262.

16. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW (2005) Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* 438(7064):86–89.

17. Anantharaman K, et al. (2014) Sulfur Oxidation Genes in Diverse Deep-Sea Viruses. *Science* 344(6185):757–760.

18. Chen J, Novick RP (2009) Phage-Mediated Intergeneric Transfer of Toxin Genes. *Science* 323(5910):139–141.

19. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS (2011) PHAST: a fast phage search tool. *Nucleic Acids Res* 39(Web Server issue):W347–52.

20. Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* 21(4):599–609.

21. Klein R, et al. (2002) *Natrialba magadii* virus phiCh1: first complete nucleotide sequence and functional organization of a virus infecting a haloalkaliphilic archaeon. *Mol Micro* 45(3):851–863.

22. Zhang Y, Zhang J, Hara H, Kato I, Inouye M (2005) Insights into the mRNA cleavage mechanism by MazF, an mRNA interferase. *J Biol Chem* 280(5):3143–3150.

23. Nishino K, Yamaguchi A (2001) Analysis of a complete library of putative drug transporter genes in Escherichia coli. *J Bact* 183(20):5803–5812.

24. Sylvestre P, Couture-Tosi E, Mock M (2002) A collagen-like surface glycoprotein is a structural component of the *Bacillus anthracis* exosporium. *Mol Micro* 45(1):169–178.

25. Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402):207–214.

26. Hyman P, Abedon ST (2010) Bacteriophage host range and bacterial resistance. *Adv Appl Microbiol* 70:217–248.

27. Barquera B, et al. (2002) Purification and characterization of the recombinant Na^+-translocating NADH:quinone oxidoreductase from *Vibrio cholerae*. *Biochemistry* 41(11):3781–3789.

28. Zwick ME, et al. (2012) Genomic characterization of the *Bacillus cereus* sensu lato species: backdrop to the evolution of *Bacillus anthracis*. *Genome Res* 22(8):1512–1524.

29. Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Micro* 3(9):711–721.

30. Fawcett T (2006) An introduction to ROC analysis. *Pattern recogn lett*

27(8):861–874.

31. Majewski J (2001) Sexual isolation in bacteria. *FEMS Micro Lett* 199(2):161–169.

32. Hooper SD, Berg OG (2003) Duplication is more common among laterally transferred genes than among indigenous genes. *Genome Biol* 4(8):R48.

33. Treangen TJ, Rocha EPC (2011) Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLoS Genet* 7(1):e1001284.

34. Tal A, Arbel-Goren R, Costantino N, Court DL, Stavans J (2014) Location of the unique integration site on an *Escherichia coli* chromosome by bacteriophage lambda DNA in vivo. *Proc Nat Acad Sci* 111(20):7308–7312.

35. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA (2005) The net of life: reconstructing the microbial phylogenetic network. *Genome Res* 15(7):954–959.

36. Smillie CS, et al. (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241–244.

37. Sharp PM, Rogers MS, McConnell DJ (1985) Selection pressures on codon usage in the complete genome of bacteriophage T7. *J Mol Evol* 21:150–160.

38. Roux S, Hallam SJ, Woyke T, Sullivan MB (2015) Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife* 4.

39. Medrano-Soto A, Moreno-Hagelsieb G, Vinuesa P, Christen JA, Collado Vides J (2004) Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. *Mol Biol and Evol* 21(10):1884–1894.

40. Tuller T, et al. (2011) Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Res* 39(11):4743–4755.

41. Cumby N, Davidson AR, Maxwell KL (2012) The moron comes of age. *Bacteriophage* 2(4):225–228.

42. Bobay LM, Rocha EPC, Touchon M (2013) The Adaptation of Temperate Bacteriophages to Their Host Genomes. *Mol Biol and Evol* 30(4):737–751.

43. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol and Evol* 30(4):772–780.

44. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biol* 59(3):307–321.

Figure Legends

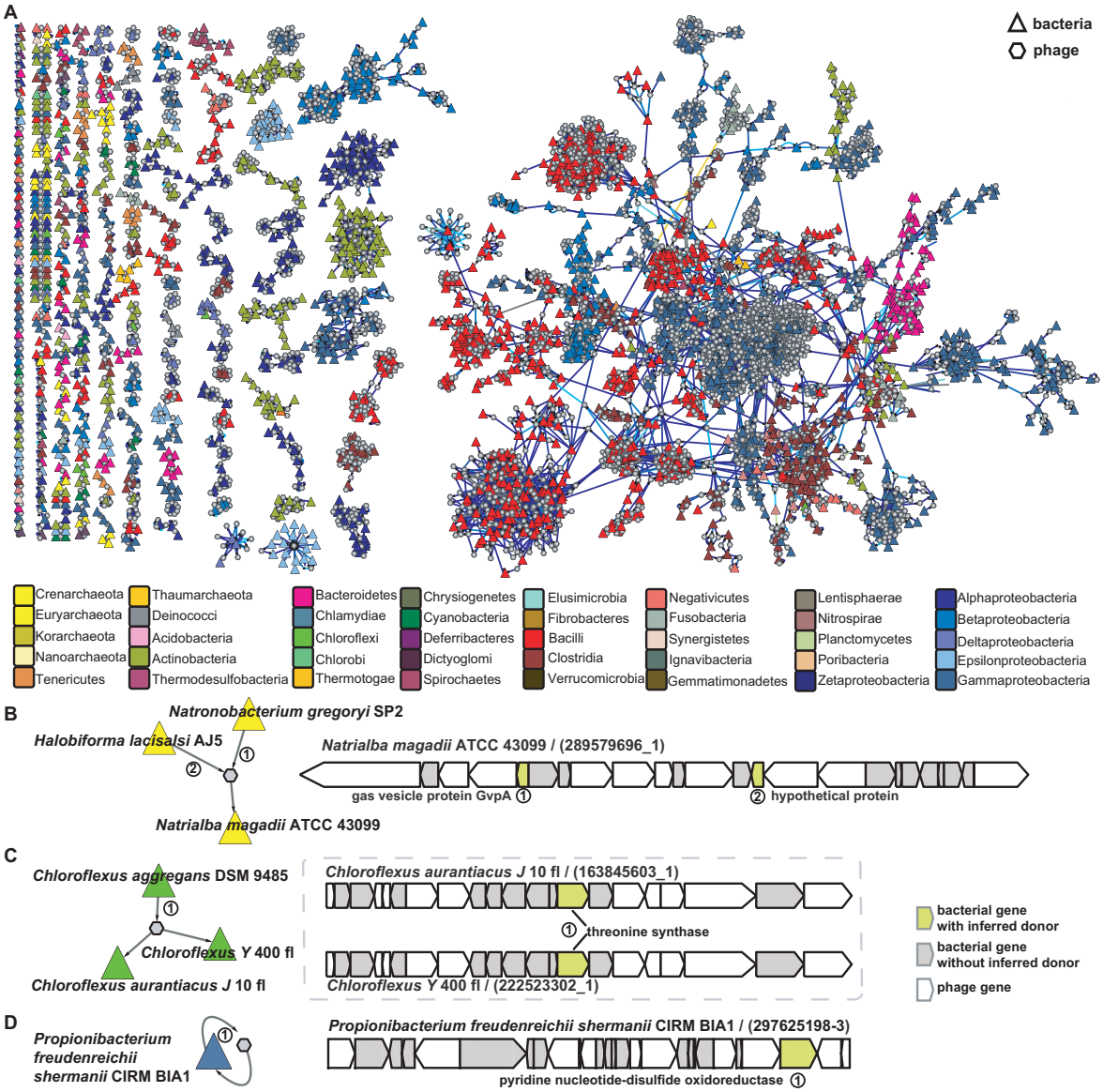


Figure 1. The directed transduction network. (A) Graphic representation of the directed, bi-partite, transduction network (dLGT) reconstructed from transduction events where a single most likely donor could be identified. The nodes correspond to bacterial genomes (triangles) and phages (hexagons). Bacterial nodes are colored by their taxonomic group. Directed edges correspond to genes that were transferred between bacterial genomes via a phage - bacteria to phage edges describe uptake of genes in transient interactions (D-dLGT) while phage to bacteria edges designate the acquisition of genes in prophages (stable interactions, R-dLGT). (B), (C) and (D) Detailed examples showing enlarged parts the dLGT, including bacterial species

names and the prohages genomic maps. Circled numbers identify specific genes in both views.

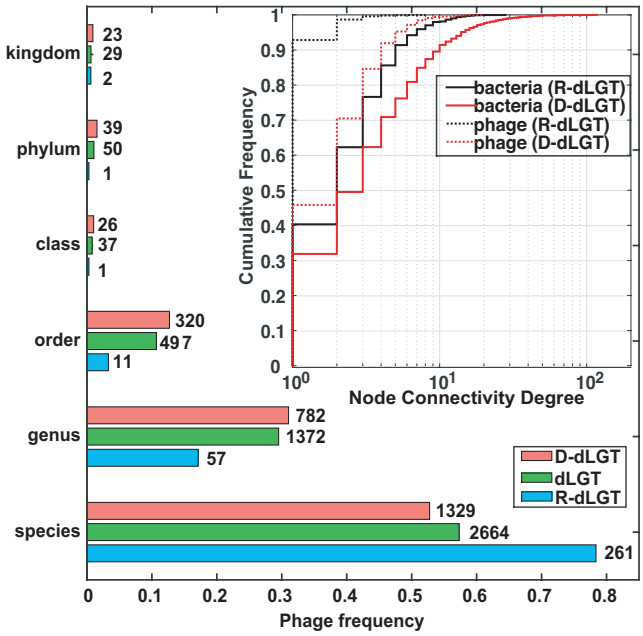


Figure 2. Taxonomic distribution and connectivity. (A) Taxonomic distribution of donors and recipient. (B) Cumulative distributions of node connectivity degree.

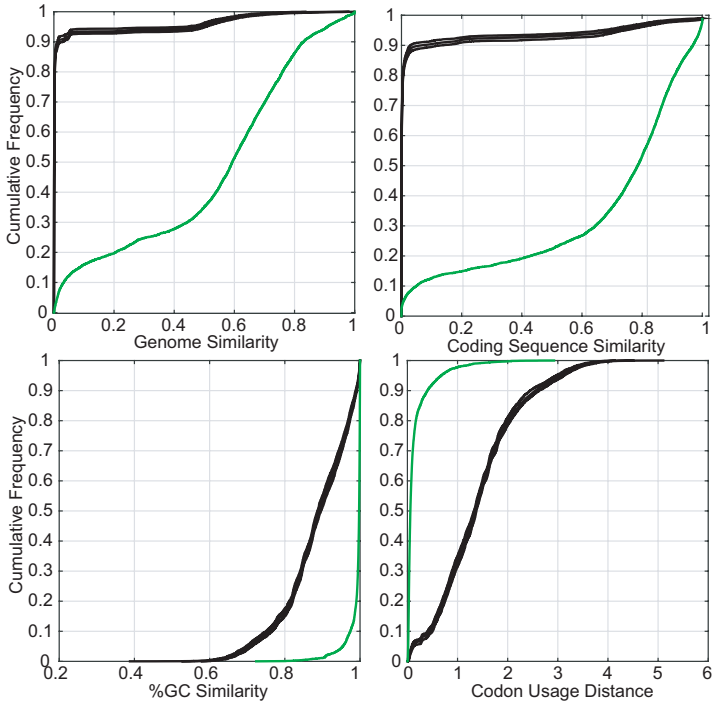


Figure 3. Donor-recipient genome similarity. Cumulative distributions of donor-recipient genome similarity measures in the dLGT network (green) and 1,000 randomized networks (black). (A) genome sequence similarity; (B) coding sequence similarity; (C) GC content similarity; and (D) codon usage distance.

Phylogenomic transduction networks reveal genetic barriers to phage-mediated lateral gene transfer during microbial evolution

Ovidiu Popa, Giddy Landan, Tal Dagan

Genomic Microbiology Group, Institute of General Microbiology, Christian-Albrechts Universtiy of Kiel, Am Botanischen Garten 11, 24118 Kiel, Germany.

SUPPLEMENTARY MATERIAL

Supplemental Methods

Data

Annotation of 14,920 prophages encoded within 8,540 genomics sequences was downloaded from the PHAST database (ver. 10/2012) (19). Genomes of 2,103 complete and 1,879 draft prokaryotic genomes were downloaded from GenBank (NCBI; ver. 10/2012). PHAST entries that could not be linked to GenBank were discarded. The resulting dataset included 9,468 annotated prophages encoded within 1,330 complete and 1,281 draft genomes. Coding sequences (CDSs) in PHAST database are classified into viral or bacterial if they have a significant BLAST hit within viral or bacterial genomes respectively (19). Prophages encoding only phage genes were excluded. The remaining 9,201 (97%) prophages encode a total of 281,616 CDSs, of which 89,234 (32%) are classified as bacterial genes.

Prophages were clustered into orthologous prophage clusters (42) based on their gene content. The first step included an all-against-all BLAST of prophage protein sequences. Reciprocal best-BLAST hits (45) with E-value<10^{-10} were aligned globally using *needle* (46). Pairs having <95% amino-acids identity were excluded. The remaining CDSs were clustered into orthologous protein families using MCL (47) with

default parameters. Pairwise prophage gene content similarity was calculated from frequencies of shared protein families using the Jaccard index. Prophages having >70% shared families are considered orthologous. Additional threshold combinations of sequence and gene content similarity were tested. In agreement with previous reports (9), we found that higher prophage similarity thresholds are too stringent, while lower values lead to clusters of distantly related prophages. Our pipeline clustered 6,494 (43.5%) prophages into 2,397 orthologous prophages. Those are considered as a phage entity in our evolutionary reconstruction. The remaining 8,426 unclustered prophages are designated as singleton phages.

Donor inference

The donor inference procedure operates within the framework of orthologous protein families and is assisted by a phylogenetic tree. In the first step we created clusters of orthologous proteins using the 89,234 bacterial genes encoded within prophages as queries for a BLAST search against the GenBank genomes. Employing an e-value cutoff of 10e-10 we identified 3,908,830 homologous sequences to 75,172 of the query genes, whereas no homologs were detected for the remaining 14,197 (15.89%) queries. Protein pairs were aligned globally using the Needleman-Wunsch algorithm (48) with *needle* from the EMBOSS package (46). Genes were considered as homologs if they had at least 90% global similarity to the query genes, resulting in a dataset of 42,760 (57%) prophage bacterial genes and 252,159 homologs. This dataset was clustered into orthologous protein families using MCL (47) with default parameters to yield 20,904 protein families with at least two proteins. A total of 12,611 protein clusters containing at least three protein sequences were aligned using MAFFT (43) Maximum likelihood trees were reconstructed using PhyML3 (44) with the LG model. The root of each tree was defined using the midpoint criteria. In protein families with multiple recipient genes, we examined the monophyly of the group of recipient genes, and when these were paraphyletic (2,205 trees, 17.48%) we tested for the likelihood of an alternative tree with recipients consolidated into one clade. We used

CONSEL (49) with the approximately unbiased test (au test) and the multi-scale bootstrap technique. Of the reconfigured trees, 829 (37.6%) were not significantly less likely than the original tree (au – test, $p \geq 0.05$) and were retained for downstream analyses. For each gene acquired by transduction, we identified the most likely gene-donor of as the genome bearing a homologous gene which is the sister clade of the acquired gene in the phylogenetic tree or the unique homolog in the case of clusters with two members.

Network construction

Donor-recipient relations were coded into the directed lateral gene transfer (dLGT) unipartite network, in which nodes represent bacterial species and edges lateral gene transfers mediated by transduction. The network is defined by an adjacency matrix $A_{(i,j)}$ of size 3,982x3,982 nodes representing bacterial genomes, in which a directed edge is weighted by the number of genes that were transferred from donor node *i* to recipient node *j* via any phage. In cases where multiple donors form a clade of size *n*, we infer the transfer to have occurred in the ancestral lineage, and assign each member of the clade a weight of *1/n*. This ensures that all transfer events are represented by an equal weight of 1, regardless of any subsequent diversification in the donor lineage. Similarly, when a group of orthologous prophages form a recipient clade, the acquisition is considered as ancestral, and edges are weighted accordingly. When a clade of recipients includes non-orthologous prophages, each subgroup of orthologous prophages is treated as separate acquisition events in that clade and all edges are weighted equally with a weight of 1.

Bacteria-phage relations were coded into a bipartite directed network where nodes represent either phages (5064 nodes) or bacterial species (3982 nodes). This enables partitioning of the network into two subsets; the stable subset (R-dLGT) consists of directed edges from prophages to their host bacterial species (i.e., the recipients), and the transient subset (D-dLGT) that consists of directed edges from donor bacterial nodes to phages (i.e. the transfer vector). As in the dLGT, edge weights

correspond to the number of transferred genes, with appropriate scaling for ancestral transfers.

Network randomization

Randomization of the dLGT network was carried out using the switching methodology (50), which rewires the weighted edges while preserving the In- and Out-degree of each node. The method was implemented in an in-house MatLab script and used to generate 1000 randomly connected networks.

Genome similarity measures

Genome sequence similarity (S_{gs}) between a recipient and a donor was calculated as the Jaccard coefficient based on the proportion of 20 bp segments common to the two genomes (using MUMmer, (51)). Proteome similarity (S_{pr}) between bacterial species was similarly calculated as the Jaccard index of identical segments, but restricted to segments which have an overlap at least of 10% within a protein coding region. GC content similarity (S_{GC}) was calculated as: $100 - |\%GC_{recipient} - \%GC_{donor}|$. The genome codon usage distance (D_{cu}) was calculated as the Euclidean distance $\mathrm{D_{CU}}(don, rec) = \sqrt{\sum_{i=1}^{n}(don_i - rec_i)^2}$ between the vectors of relative codon frequencies per amino acid within the donor and recipient genomes.

Synonymous and Non-synonymous Substitution Rates

The number of non-synonymous (d_N) and synonymous (d_S) substitutions, and their ratio ω, was calculated using the branch model implemented in PAML (52). We used "model 2" of PAML, allowing ω (d_N/d_S) to vary among the donor branches, recipient branches, and the remaining ("background") branches. For the special case of gene duplication (self-donor recipient loops), we estimated d_N, d_S and ω using the software package PAL2NAL (53). PAL2NAL creates a codon alignment from a pair of protein and their corresponding DNA sequences and calculated the d_N and d_S values using the PAML (52). program.

Functional and Habitat classification

Functional classification of each cluster was derived from the Clusters of Orthologous Groups (COG) database (45) by a majority vote of cluster members. The habitat classification of donor and recipient nodes was extracted from the GOLD database version March 2014. We defined 11 main habitat classes using the combination of isolation place and ecology annotation (see Fig. S3). Laterally transferred toxin-antitoxin (TA) operons were surveyed using PanDaTox (54) as a query. An additional survey for laterally transferred genes for antibiotics resistance was performed using the genes in CARD database (55) as queries.

Statistics and Visualization

All statistical calculations were done using the Statistics -Toolbox from the MatLab© platform. The network layout was calculated with Cytoscape (56) using the force-directed graph drawing module.

Supplemental references

45. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278(5338):631–637.

46. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16(6):276–277.

47. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7):1575–1584.

48. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48(3):443–453.

49. Shimodaira H, Hasegawa M (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17(12):1246–1247.

50. Artzy-Randrup Y, Stone L (2005) Generating uniformly distributed random networks. *Phys Rev E* 72(5):056708.

51. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.

52. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591.

53. Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34(Web Server issue):W609–12.

54. Amitai G, Sorek R (2012) PanDaTox: A tool for accelerated metabolic engineering. *Bioengineered* 3(4):1–4.

55. McArthur AG, et al. (2013) The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* 57(7):3348–3357.

56. Shannon P, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13(11):2498–2504.

Supplementary Figures

Figure S1. Cumulative distribution function of (A) the number of autologs per self-donor edge and (B) the number of phage nodes that are connected to self-donor nodes.

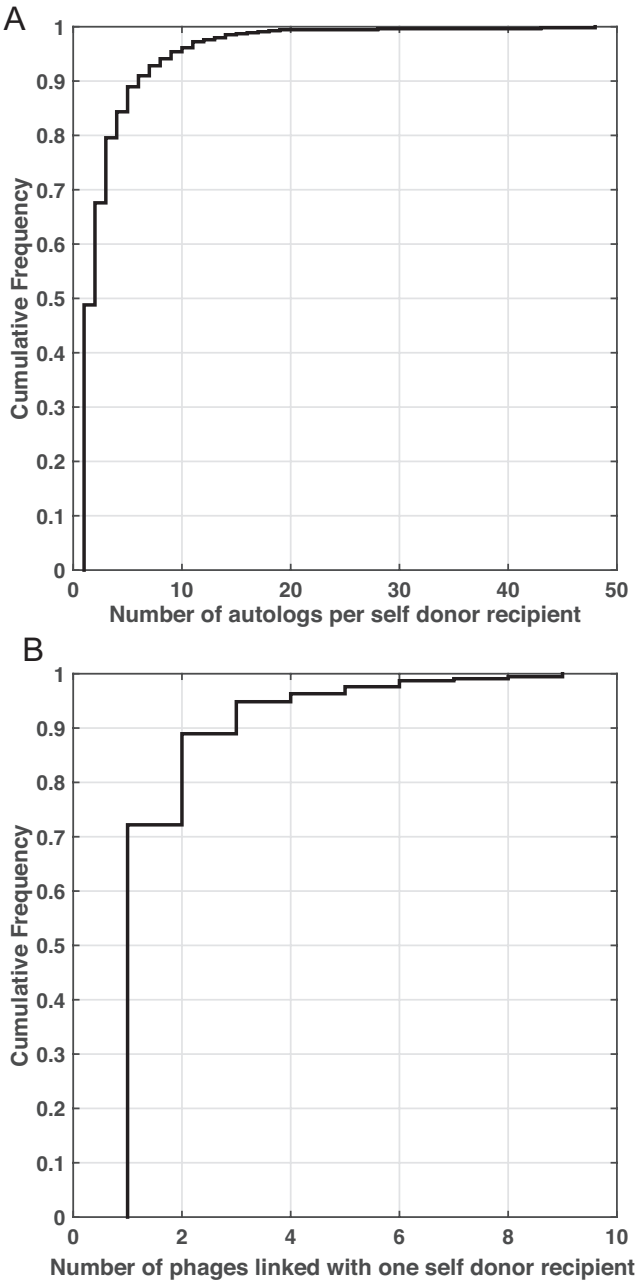


Figure S2. Transduction events. Detailed examples of phage entities from the dLGT network (A) Each phage entity (gray box) encompasses a single prophage or several orthologous prophages. Phage genes are shaded white and bacterial genes are shaded green when a donor could be inferred, or gray when a donor could not be inferred. Green numbers are dLGT node identifiers (see supplementary table 2), black numbers are prophage GI and region number as recorded in the PHAST DB. (B) Corresponding dLGT network views with donor and recipient bacterial genomes. Node color correspond to taxonomical group as described in Figure 1A. Edge color represents edge weight (color bar at bottom).

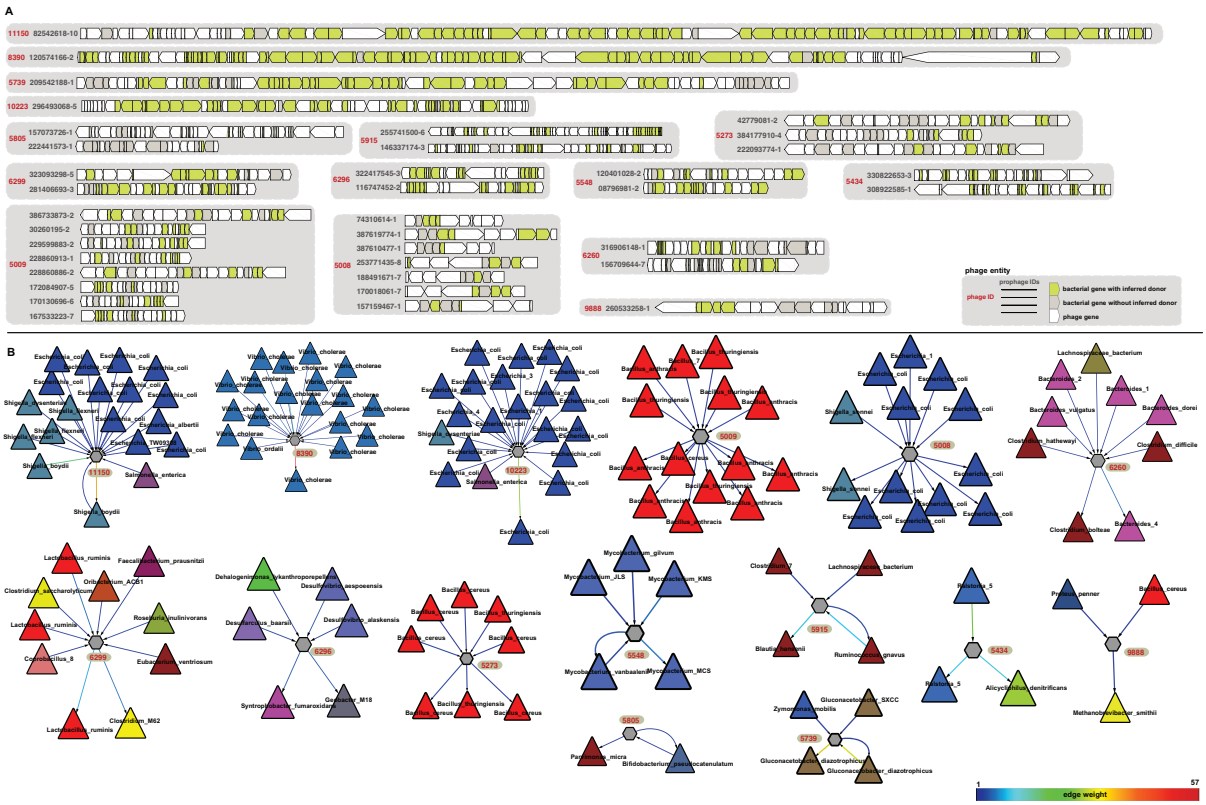


Figure S3. Edge weight distribution. Cumulative distribution function of edge weight in thenetwork.

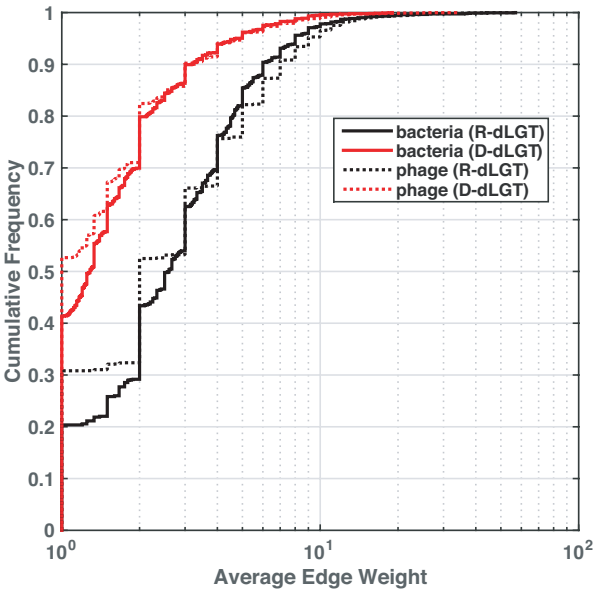
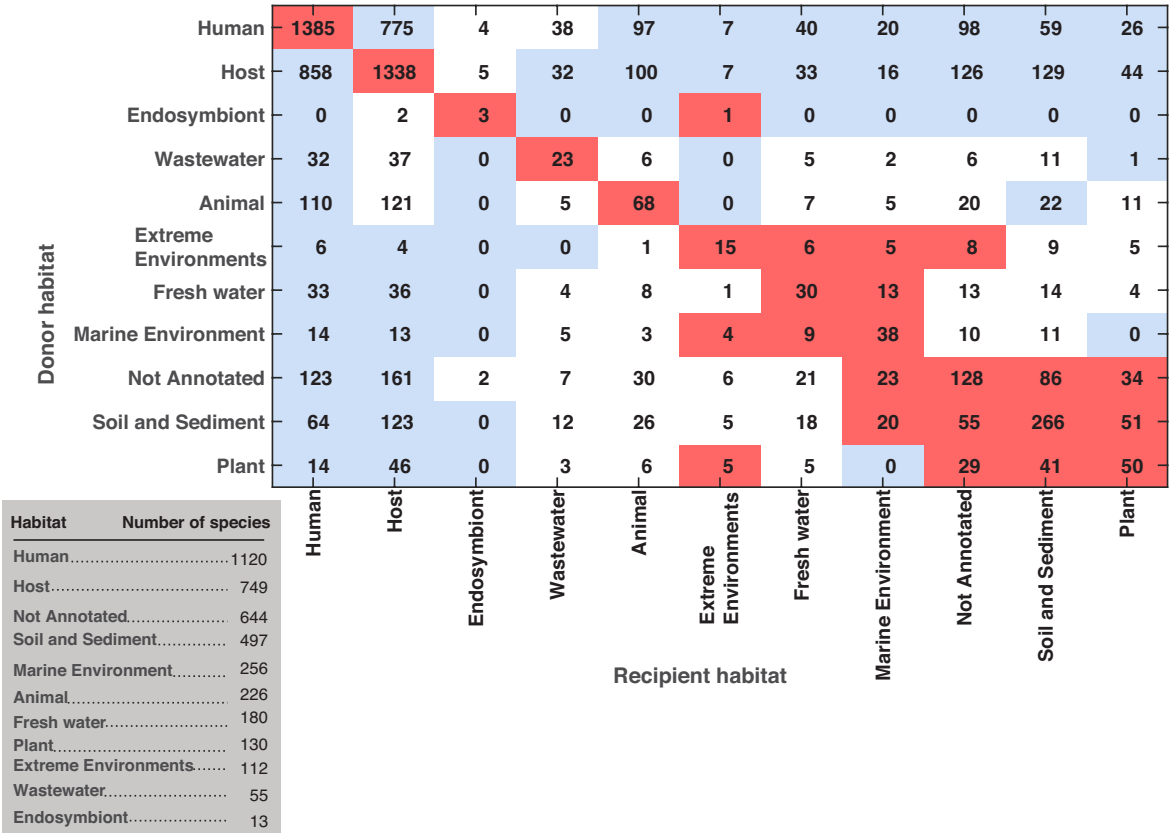
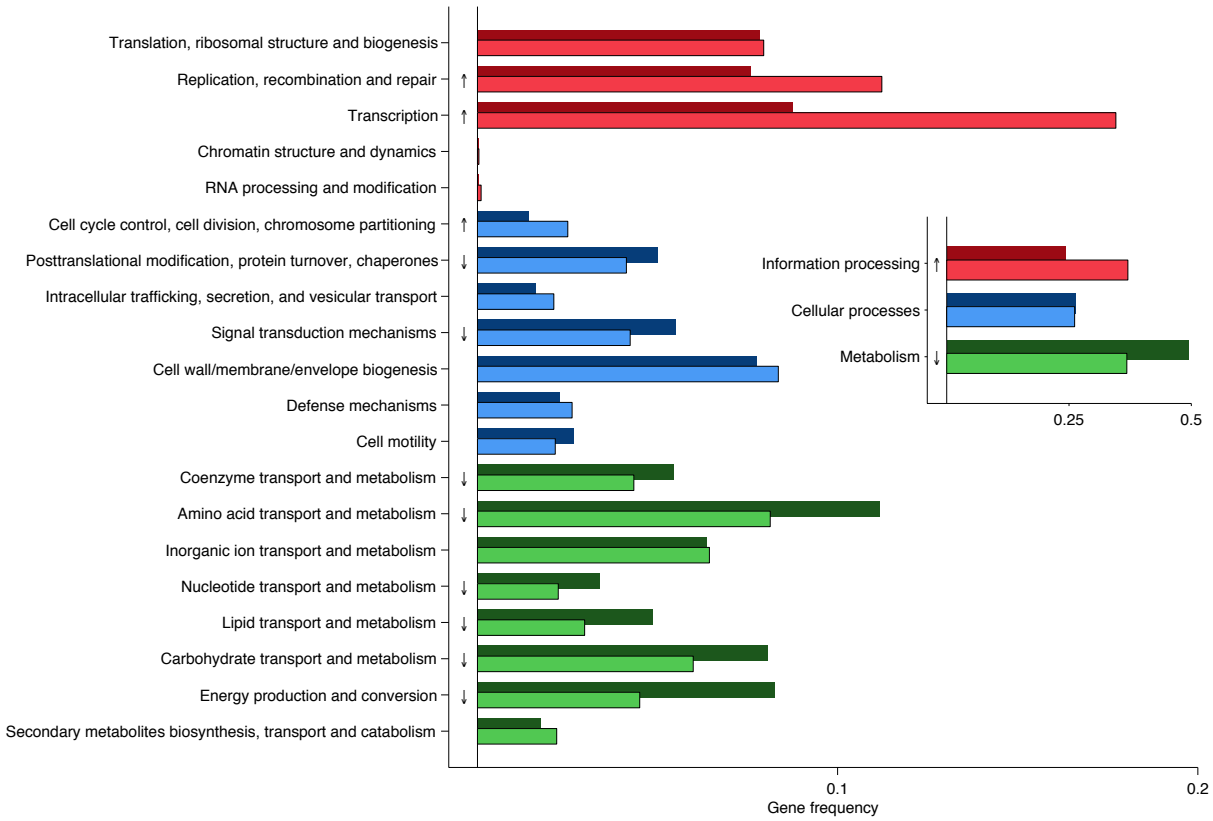


Figure S4. Heat map of connected donor and recipient species within the same or between different habitats. The numbers within each cell, correspond to the number of connected donors and recipients from a particular habitat. Cells coloured in red represent connected habitats in the dLGT network that show a higher frequency than expected, blue coloured cells correspond to lower frequency than expected and white cells are not significantly different than expected. Gray shaded box represents the distribution of the 3,982 bacterial species into a putative habitat. The values are sorted by size from top to down in a descend order.



Habitat	Number of species
Human	1120
Host	749
Not Annotated	644
Soil and Sediment	497
Marine Environment	256
Animal	226
Fresh water	180
Plant	130
Extreme Environments	112
Wastewater	55
Endosymbiont	13

Figure S5. Histogram of the relative proportion of acquired genes (LGTs) that were assigned to a COG category: information storage and processing (red), cellular processes and signalling (blue) and metabolism (green). Arrows on the left side indicates if the category is significantly overrepresented or underrepresented compared to the full data set. Dark colored bars correspond to the prophage distribution and light colored ones to the full data set.



Concluding remarks

In biological systems where reticulated evolutionary events are common, phylogenomic networks offer a general computational approach that is more biologically realistic and evolutionarily more accurate. The prevalence of LGT during microbial and viral evolution makes phylogenomic networks an essential tool in the study of these systems.

Directed networks of LGT among prokaryotes reconstructed from completely sequenced genomes uncover barriers to LGT in multiple levels including physical barriers for gene transfer between cells, genomic barriers for the integration of acquired DNA, and functional barriers for the acquisition of new genes. Furthermore the bipartite structure of a directed, phylogenomic network reconstructed for transduction events, indicates that donor-recipient whole genome similarity is an important factor that shapes the transduction network connectivity pattern demonstrating the implication of phage-bacteria coevolution to phage-mediated gene transfer during microbial evolution. The difference in the connectivity pattern of the transduction dLGT network reveals that DNA transfer via transduction occurs within the context of phage-host specificity, but that this tight restriction can be occasionally violated, enabling long range LGTs of profound evolutionary consequences

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

Acknowledgements

First and foremost I would like to thank my family for all the support they have given me over the years. Their time, patience and courage the have granted me, made all this work possible.

I would like to host a very special thank to Prof. Dr. Tal Dagan for her confidence she put in my work over the years and of course for the opportunity to do this thesis. The time under her supervision was very informative; full of ideas, advices and great support for every issue I had at day and night.

Thanks to Prof. Dr. William Martin for supporting and founding me during my undergraduate time and the beginning of my PHD studies. It is always a pleasure to work in his institute.

Thanks to Dr. Giddy Landan who had taught me the way you need to look into the data you are analyzing and that the best statistics is always the one with the minimal amount of manipulation steps of your data. It was a great time and a pleasure for me to talk and discuss with you.

Thanks to Jun. Prof. Dr. Oliver Ebenhöh who supports me for the last year of my PHD thesis and for his great ability to encourage people.

Thanks to all my colleagues, especially Julia Weißenbach, Dr. David Bogumil, Robin Koch und Judith Ilhan from genomic microbiology in Kiel, we had a wonderful time with a lot of fun and very important with a lot of good food and drinks ☺!

Thanks to my colleagues from the Quantitative and Theoretical Biology (QTB) group in Düsseldorf for great inspiration moments in 2015.