REVIEWS

HORIZONTAL GENE TRANSFER, GENOME INNOVATION AND EVOLUTION

J. Peter Gogarten and Jeffrey P. Townsend

Abstract | To what extent is the tree of life the best representation of the evolutionary history of microorganisms? Recent work has shown that, among sets of prokaryotic genomes in which most homologous genes show extremely low sequence divergence, gene content can vary enormously, implying that those genes that are variably present or absent are frequently horizontally transferred. Traditionally, successful horizontal gene transfer was assumed to provide a selective advantage to either the host or the gene itself, but could horizontally transferred genes be neutral or nearly neutral? We suggest that for many prokaryotes, the boundaries between species are fuzzy, and therefore the principles of population genetics must be broadened so that they can be applied to higher taxonomic categories.

TREE OF LIFE
The tree-like representation of the history of all living and extinct organisms.

MUTUALISM
An association between two organisms, often from different species, that benefits both partners.

RETICULATION
A network that is formed through the fusion of independent branches.

Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut 06269-3125, USA.
e-mails: gogarten@uconn.edu; Jeffrey.Townsend@uconn.edu
doi:10.1038/nrmicro1204
Published online 1 August 2005

Bifurcating trees, in which evolutionary lineages split and evolve independently from each other, have a long history as tools to visualize the evolution of species: Lamarck introduced tree-like binary schemes for taxonomic classification[1] and Charles Darwin described the evolution of species as the TREE OF LIFE[2]. Darwin also noted that the 'coral of life' might be a more appropriate metaphor, because only the outermost layer in the tree of life is actually alive, resting on a base of dead branches[3]. The tree of life became the standard imagery to depict species evolution, implying a common root of all life on Earth and a bifurcating evolutionary process. However, there are several clear exceptions to this standard view. For example, botanists found that many plant species violate a bifurcating model as they are allopolyploid, combining the genomes of different parental species. This process of a new line of descent originating from the hybridization of two parent species has been termed reticulate evolution[4–8]. The fungal–algal symbiosis of lichens illustrates that symbiosis can lead to long-term partnerships with different properties from those of either parent species, and some of the most dramatic breakthroughs in cellular evolution, that is, the mitochondria and plastids, are the result of endosymbiosis[9]. Throughout the decades, MUTUALISM and RETICULATION have often been considered the most important processes in species evolution[10,11]. However, for most branches of biology these processes were only exceptions, albeit important ones, in an otherwise steadily furcating process of species evolution.

By introducing ribosomal RNA (rRNA) as a taxonomic marker molecule, Woese and Fox extended the tree paradigm to the realm of microorganisms[12,13]. However, the large-scale availability of sequence data provided information that effectively sundered the cambium of the tree of life metaphor. Different molecules were shown to have different histories[14], and members of the same species were found to differ dramatically in gene content. For example, of the genes revealed by the sequencing of three *Escherichia coli* genomes, fewer than 40% were common to all three[15]. Furthermore, it has been suggested that extinct species have contributed genes to the extant layer of life, even though these contributors might not have been in the direct line of ancestry[16,17]. Reticulate models of evolutionary history that incorporate gene transfer might

PHYLOGENY
The origin and evolution of a group of organisms, usually of species. Phylogenies are not necessarily tree-like. The term phylogeny is frequently applied to genes and different levels of taxonomic units. These are often labelled as operational taxonomic units.

GENE TREE
A depiction of the history of families of homologous genes. Intragenic recombination can gives rise to non-tree-like gene phylogenies.

HOMOLOGUES
Characters or sequences that are derived from the same ancestral feature.

even provide an opportunity to learn about organisms that became extinct hundreds of million years ago.

It is now known that organismal mutualisms and lineage reticulation are supplemented by horizontal gene transfer (HGT) as processes that lead to the network-like histories of living organisms. Over short time intervals, an organismal line of descent could be defined as a 'plurality consensus' of gene histories. However, these organismal lines of descent are embedded in a web of gene PHYLOGENIES that form connections between the different branches of the tree of life[17,18]. It now appears that all functional categories of genes are susceptible to HGT, even rRNA operons[19] and genes associated with phylum-defining characteristics, such as the photosynthetic machinery[20]. However, not all genes are equally itinerant. Some clearly have a higher propensity for transfer than others[21], and not all groups of organisms experience HGT to the same extent[22,23].

Patterns from HGT versus shared ancestry

One of the predicted outcomes of high levels of HGT between preferred partners is the observation of robust GENE TREES, the implications of which are indistinguishable from the signals produced by recent shared ancestry[19,24]. In particular, two predictions are worth considering: organisms that frequently give or receive genes from sister taxa will group together in most gene phylogenies (that is, the phylogenies of the transferred genes), and organisms not participating in HGT with sister taxa should be 'left basal' by those that are, and these non-participatory lineages should be recovered as deep branching lineages.

The Thermotogales provide an interesting illustration of this point. This group of extreme thermophiles is recovered as a deep branching lineage of Bacteria when using individual gene phylogenies[13,25] as well as using whole-genome-based analyses (see REF. 26 for a recent review). When the genome of *Thermotoga maritima* was sequenced, >20% of the open reading frames were reported to be most similar to HOMOLOGUES from the Archaea[27]. The Thermotogales share their environment mainly with Archaea; FIG. 1a illustrates that many of the proteins encoded in the *T. maritima* genome are less divergent from their archaeal homologues than is found for other Bacteria with comparably sized genomes, such as *Streptococcus thermophilus*. Proteins with slight divergence, which places them in the left-hand tail of the *T. maritima* distribution, presumably correspond to those that have been horizontally transferred from archaeal taxa at some time subsequent to the Archaea–Bacteria split. Imposing a rate of HGT that results in a lower divergence for 3% of loci to the actual *S. thermophilus* gene-divergence diagram illustrates a left-hand tail to the distribution of gene divergences (FIG. 1b) similar to that observed for *T. maritima* (FIG. 1a).

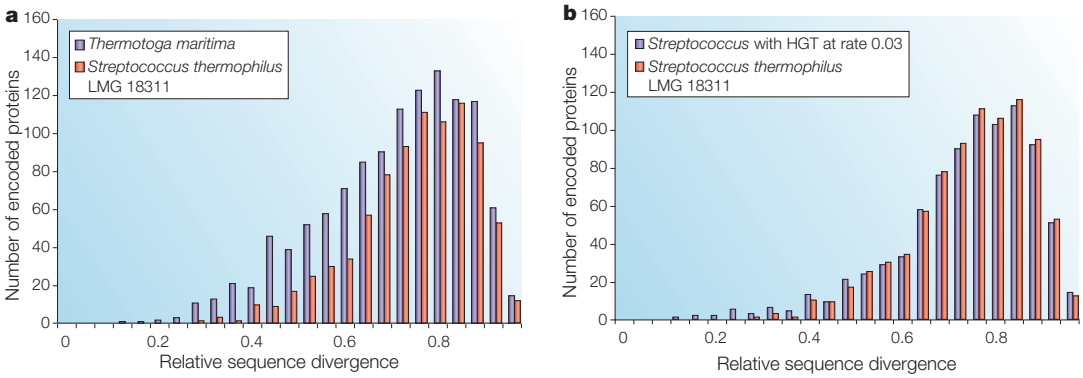

Figure 1 | **The effects of gene transfer on sequence divergence. a** | Histogram showing the number of encoded proteins with different levels of relative sequence divergence. The divergence from the most similar euryarchaeal homologue was calculated for all annotated open reading frames of a Gram-positive bacterium with little interdomain transfer (*Streptococcus thermophilus* LMG 18311, red) and for the extremely thermophilic bacterium *Thermotoga maritima* (blue). For each encoded protein, BLAST searches were carried out against the proteins in five archaeal genomes (*Pyrococcus abyssi*, *Pyrococcus furiosus*, *Archaeoglobus fulgidus*, *Methanocaldococcus jannaschii* and *Methanothermobacter thermautotrophicus*). The *S. thermophilus* genome encodes 1,889 currently annotated open reading frames, 851 of which have a significant match in at least one of the euryarchaeal genomes (E value $<10^{-3}$); the *T. maritima* genome encodes 1,858 proteins, 1,193 of which have a significant match in at least one of the archaeal genomes. The bitscore divided by the alignment lengths was used as a measure of sequence similarity. Relative sequence divergence between two sequences was calculated as (1–similarity(b_a)/similarity(b_b)), where similarity(b_a) is the similarity score for a bacterial sequence with the most similar archaeal sequence, and similarity(b_b) is the similarity score of the bacterial sequence compared with itself. Two thirds of those *T. maritima* genes with <45% sequence divergence are classified as encoding genes that fall into the metabolism category in the COG (clusters of orthologous groups) database[100,101], whereas only a third of all *T. maritima* genes fall into this category. Note that the tail in the distribution, owing to the presence of sequences with little divergence, is absent in the case of *S. thermophilus*, suggesting transfer into the *Thermotoga* lineage as the most probable explanation. **b** | An example of the effect of horizontal gene transfer (HGT) on the distribution of the divergence of genes. In red, the distribution for the percentage amino-acid divergence of genes of *S. thermophilus*. In blue, the expected distribution of divergences of genes given the same gene-specific divergence rates from a donor taxon, but in addition incorporating a low rate of gene transfer from the donor taxon, affecting 3% of the genes. Details of how the expectations were calculated can be obtained from the authors. This calculation is meant to be merely illustrative; the development of precise quantitative methods for testing well specified models of HGT is vital to future studies of particular taxa.

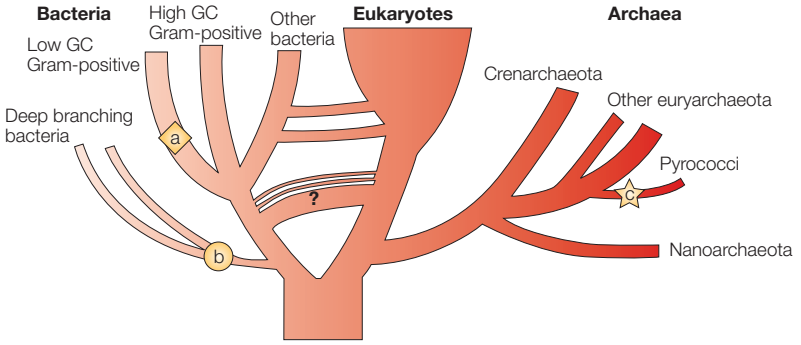

Figure 2 | **The tree of life.** A sketch of the tree of life as it is frequently derived from genome data (for example, REF. 26), with the three possible positions of *Thermotoga maritima* marked according to (**a**) 'concordant' genes (placed with the Gram-positives), (**b**) 16S RNA (and other conserved genes) and whole-genome analyses (placed as an early diverging lineage) and (**c**) phylogenetically discordant genes (placed with the Pyrococci among the Archaea). For further discussion see REF. 28 and text.

Recently, Gophna *et al.*[28] constructed separate genome-based trees from different sets of genes: those that are frequently found to be in phylogenetic agreement with one another and those that are frequently found to be phylogenetically discordant. They found that the phylogenetically discordant genes group *T. maritima* among the Archaea as a sister group to the Pyrococci. By contrast, the concordant genes group *T. maritima* within the Bacteria at the base of the Gram-positive Bacteria (FIG. 2). A phylogenetic affiliation for *T. maritima* similar to that for concordant genes was recovered by Daubin and colleagues[29] using a SUPERTREE approach on stringently selected sets of orthologous genes. According to phylogenetic analysis of infrequently horizontally transferred genes, *T. maritima* seems to be Gram-positive, but frequently transferred genes group it inside the Archaea domain. However, if all genes are considered, *T. maritima* is recovered as a deep branching lineage of Bacteria. These findings indicate that the deep branching bacterial phylogenetic position is an artefact, resulting from analyses that combine genes with different phylogenetic histories.

Surprisingly, *T. maritima* is also frequently recovered as a deep branching lineage of Bacteria when using some of the most trusted individual phylogenetic markers[13,25,30]. Brochier and Phillipe[31] suggest that these results could also be artefacts of phylogenetic reconstruction. Alternatively, some molecular improvements might have spread more recently within the Bacteria by HGT, leaving the Thermotogales with the more ancient and less altered versions (see discussion on species boundaries below). If one considers that, owing to recombination, slowly evolving molecules themselves might be mosaic[19,32–36], then more recent HGT among the non-Thermotogales could explain the similarity between whole-genome phylogenies and single conserved molecules. Interestingly, this explanation rescues the extremely thermophilic Bacteria as a model for early Bacteria: although they no longer represent deep branching organismal lineages, they have perhaps not shared many of the improvements that were recently exchanged among the mesophilic Bacteria.

SUPERTREES
Trees calculated from smaller trees with sets of overlapping operational taxonomic units.

ORTHOLOGUES
Homologues that are related to each other through a speciation event.

PURIFYING SELECTION
Kimura's neutral theory of molecular evolution posits that most variations observed at the molecular level do not provide a selective advantage or disadvantage. However, many nucleotide mutations are never observed in a population because they are associated with a strong selective disadvantage. This is the case for mutations that change a catalytically important amino acid. The selection that prevents these detrimental mutations from becoming fixed in a population is known as purifying selection.

Improving detection of HGT

Several methods have been developed to detect horizontally transferred genes. One of the most popular methods is the detection of codon or nucleotide compositional bias[37–40]. This approach identifies many recently transferred genes[41]. However, it has not been unequivocally shown that HGT is the sole cause of unusual compositional bias. Furthermore, not all recently acquired genes show compositional bias[42]; it is even conceivable that some of those recently acquired genes that increase the fitness of the recipient show a weaker compositional bias. In principle, examination of the phylogenetic conflict among loci is the most direct approach to screen for horizontally transferred genes. For example, heat shock protein homologues (HSP70) group the Archaea among the Bacteria[43–46], and many proteins in *T. maritima* are most similar to ORTHOLOGUES from the Pyrococci[27], presumably because some archaea acquired an HSP70 homologue gene from a bacterium, and the *Thermotoga* lineage incorporated many genes from the Pyrococci or their relatives. Computer programs have been developed that automate the assembly of gene families and reconstruct their phylogeny to detect HGT in genome analyses[47–50]. The use of phylogenetic reconstruction promises more reliable detection of HGT events than simple database searches[51]. With more genome sequences available for closely related organisms, these approaches promise to become even more useful.

However, there are several potential pitfalls to avoid in analyses of phylogenetic conflict, especially for events that happened in the distant past. First, many problems arise from the limited and often noise-riddled phylogenetic information that a gene sequence presents about long-ago periods of evolutionary history. Conserved sequence positions allow the identification of homologues, but a perfectly conserved sequence position contains no phylogenetic information. For example, the amino-acid sequence of histones or ATP-synthase catalytic subunits is nearly identical in closely related species, and therefore useless in reconstructing within-genus relationships. In addition, sequence positions that experience little or no PURIFYING SELECTION will rapidly become saturated with substitutions and will not retain any phylogenetic information[52]. Although all gene families retain information for phylogenetic reconstruction at some phylogenetic depth, in general, this information is insufficient for the reconstruction of relationships at most other phylogenetic depths. Another problem is that gene duplication followed by gene loss can give rise to different gene trees, and therefore conflicting phylogenetic signals that are indistinguishable from those resulting from HGT (FIG. 3). Therefore, phylogenetic incongruence is a well defined screen for HGT, but not unambiguous proof[53].

Taxonomists are frequently divided into 'lumpers' and 'splitters'[54]. At the molecular level, the same tendencies give rise to those who concatenate data to extract even the smallest grain of phylogenetic information that might be distributed over many gene

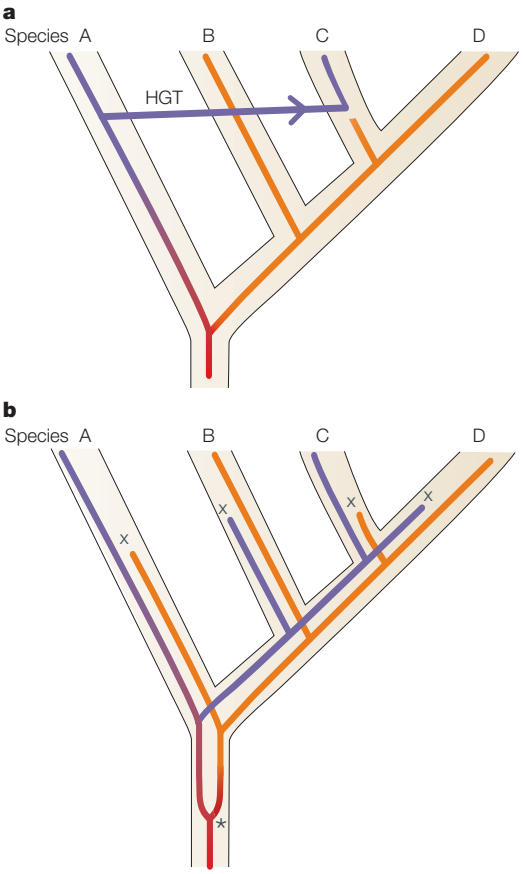

Figure 3 | **Comparison of two explanations for unexpected phylogenetic distribution. a** | The presence of a gene with characteristics that are typical for an unrelated group can be due to horizontal gene transfer (HGT, arrow). **b** | An alternative explanation is an ancient gene duplication (*) followed by differential gene loss (x). The more sister lineages have only the typical gene, the more independent gene-loss events must be postulated under this scenario.

families, and those who direct their attention to the signal contained in individual gene families and who synthesize a consensus only after they have ascertained that there are compatible signals. Much progress has been made in confronting the pitfalls inherent to both approaches, and the different approaches are becoming more similar in their practical implementation.

One problem with concatenation is the selection of data to include. This selection is complicated by the fact that the absence of evidence for transfer cannot be taken as evidence for the absence of transfer. If one applies a stringent measure for conflict, nearly all genes agree with the consensus signal within the limits of confidence. The amount of conflict detected depends on the chosen limits of confidence and on the extent of taxon sampling[55–58]. Tests of compatibility between different trees and the datasets from which these trees were derived[59,60] have become the preferred tool to assess the potential conflict between individual gene families[61], but more sophisticated methods that take a larger number of possible trees into consideration are being developed[62,63].

Fractionating phylogenetic information into smaller sub-analyses (down to the level of quartets) can generate artefacts owing to poor taxon sampling (reviewed in REF. 64). However, evaluation of phylogenies embedded in larger trees effectively addresses this problem[53]. An advantage of analyses that focus on small quanta of phylogenetic information is that plurality consensus signals can be extracted, even though not a single gene family might be in perfect agreement with any other gene phylogeny (supertree approaches). Furthermore, gene families that retain conflicting phylogenetic information can be identified readily, even though these genes might not allow a perfect reconstruction of the gene family's phylogeny at all levels of relationship. For example, spectral analyses of BIPARTITION data, so-called LENTO PLOTS[65], can be applied to all gene families present in a selection of genomes[22]. These Lento plots are histograms that depict support and conflict for the different bipartitions. As bipartitions can readily be separated into those that are compatible (those that can coexist in a tree) and those that are incompatible (bipartitions that cannot coexist in the same tree), these spectra readily identify gene families with conflicting phylogenetic information. The screened and filtered phylogenetic information can be synthesized into consensus trees, as is the case in the many supertree approaches that are being developed[66]. However, if organismal phylogeny is embedded in a web of gene phylogenies that is woven through many gene transfers and other reticulation events, analyses that directly reconstruct a network, rather than a single tree, seem appropriate.

Networks that contain only individual loops that do not share nodes with one another have become known as GALLED TREES. The methodology for construction of these networks of non-interwoven loops has undergone dramatic progress[67–69]. Similar types of analysis that focus on conflicting information contained in a dataset are split-decomposition analyses[70–72]. These analyses provide visually compelling illustrations of phylogenetic ambiguity or conflict contained in data[73], but the loops in these networks cannot readily be interpreted as depicting evolutionary histories. A related approach, recently described[74] and applied to the origin of eukaryotes[75], reaffirmed that eukaryotes contain both archaeal and bacterial genes, and suggested an overall ring structure for organismal phylogeny. Were the bacterial genes found in eukaryotes brought into the eukaryotic cell concomitantly through a small number of fusion and endosymbiotic events? Were they brought in through a steady trickle of smaller HGT events? Or were the bacterial genes present in the most recent common ancestor of all living organisms and lost in the archaeal lineage? And in any case, are HGT recipients receiving genes mostly from a few donors, or from many? At present, these alternatives remain under vigorous debate[76–81] (FIG. 1). Quantitative models of these possibilities would allow testing and evaluation of their likelihood.

BIPARTITION
A bipartition corresponds to an internal branch in a phylogenetic tree. A single bipartition divides the data (sequences, genomes or species) into two groups, but it does not consider the relationships within each of these groups.

LENTO PLOTS
Phylogenetic analyses using bipartition data named after G.M. Lento. For each bipartition, the bipartition spectra give the support for the bipartition as a histogram bar in the positive direction, and the support for all conflicting bipartitions as a bar in the negative direction.

GALLED TREES
Trees that contain local deviations from a strictly furcating pattern.

Deleterious, beneficial or neutral?

Horizontal transfer of functional units can provide the recipient with the tools necessary to occupy new ecological niches. The selective advantage conferred by transferred genes is fundamental to the SELFISH OPERON THEORY[82]. This theory is most clearly manifested in pathogenicity and ecological islands, which are contiguous sets of genes acquired through HGT that form genomic islands of atypical composition. Such sets of genes encode functions conferring traits that allow the colonization of new ecological niches[83]. The discovery of selectable and easily transferred genomic islands gave rise to the expectation that, for genes to be horizontally transferred and successfully integrated into the recipient genome, the transferred genes would need to provide a selective advantage either to the recipient or (in the case of parasitic genetic elements) to themselves[84]. Understanding the frequency with which horizontally transferred genes confer a benefit, are neutral or nearly so, or are deleterious is of great importance[85], both for our understanding of the impact of HGT on the evolution of microorganisms[86] and for the practical purpose of understanding the potential spread of transgenes to natural microbial populations[87].

Comparison of genomes from strains of *E. coli* has revealed that the core of genes present in all strains of this microbial species is surprisingly small[15]. Many of the recently transferred genes are not present in sister taxa[41]. These 'ORFANS' are on average shorter than other genes, contain a higher percentage of A and T nucleotides, and have a codon usage that is similar to that found in phage and plasmids[41,88,89]. Intriguingly, the average codon usage in the recently transferred genes is even more extreme than the average calculated from phage genes[41]. If the atypical composition were to reflect the genomic bias of the previous bacterial host, then atypical genes with a higher GC content should frequently be found, but this is not the case. The recently acquired genes have a higher AT content than the typical chromosomal genes, and this seems to be true even for Bacteria with a high AT content[41]. These findings indicate the presence of a 'vapour' of transient genes that surrounds a stable set of core genes[90]. The genes in the vapour cloud sometimes reside within the bacterial chromosomes, but perhaps more frequently reside in phage and extrachromosomal genetic elements. Daubin *et al.* calculated that these genes have a high turnover rate in the genome[41].

In those instances where acquired genes were present in two or more closely related *E. coli* and *Salmonella enterica* genomes, the ratio of non-synonymous (K_a) to synonymous (K_s) substitutions indicated that most of these genes were under purifying selection; that is, nucleotide substitutions that change the encoded amino acid occur at a lower rate than substitutions that leave the encoded amino acid unchanged by virtue of the redundancy of the genetic code. However, the K_a/K_s ratio for these transferred genes, while indicative of purifying selection, is higher than for other *E. coli* genes. For transferred genes present in both *E. coli* and *S. typhimurium*, the K_a/K_s ratio was calculated as 0.19, whereas the genes classified as 'native' had a K_a/K_s ratio of 0.05 (REF. 89). So, although these apparently transient genes are under purifying selection, this selection is weak. In part, the weak selection might be the consequence of selection against novel deleterious function (for example, protein 'toxicity'[91]) instead of a need to retain a selective advantage that these genes provide to their host.

The notion that many of these non-core genes might be selectively neutral or nearly neutral is also suggested by recent studies of a marine bacterioplankton population of *Vibrio splendidus* by Thompson *et al.*[92] Even though the analysed bacteria all fall into a tight RIBOTYPE cluster with less than 1% sequence divergence in the 16S rRNA gene, the diversity at the genome level is astounding: among the 206 strains tested, 180 unique genotypes were determined by pulse-field gel electrophoresis. Individual genotypes are present at low concentration: Thompson *et al.* estimate that the population (defined as the ribotype cluster) contains >1,000 unique genotypes. Twelve strains that were analysed in more detail differ in genome size between 4.5 and 5.6 Mb. Apparently, none of the detected variations provided sufficient selective advantage to initiate a SELECTIVE SWEEP (also known as a periodic selection event)[93]. Either the ribotype cluster consists of many distinct subpopulations that each occupy a distinct ecological niche, or the astounding genomic variability found in this study is selectively neutral or nearly neutral.

The following picture is emerging: a large amount of gene swapping and gene exchange occurs between chromosomal and non-chromosomal genes. Most of these transfers are nearly neutral to the recipient, some might increase the fitness of phage and viruses (MORONS[41,94]) under some conditions. Within the large pool of recently transferred genes, there are a few genes that increase the fitness of the recipient. These rare transfers can become fixed owing to a selective sweep, and it is only these latter transfers that are usually detected using comparative molecular phylogenies.

Population genetics for prokaryotes

The biological species concept[95] defines a species as a potentially interbreeding group of organisms that are capable of producing fertile offspring. Within such groups, gene phylogenies are seldom congruent, owing to high rates of gene flow and recombination[19,38]. Which phenomena generate cohesion within a prokaryotic species? Two different, but not mutually exclusive, mechanisms have been suggested. First, high levels of gene transfer followed by homologous recombination could play the part that sexual reproduction plays for gene flow in multicellular eukaryotes[38]. In this case, cohesion would be maintained by high levels of genetic exchange. Or, cohesion could also be generated through selective sweeps that occur if a gene that provides a selective advantage to its carrier arises through mutation or gene transfer[96].

SELFISH OPERON THEORY
Explains the formation of operons through gene transfer. According to this theory, genes encoding parts of the same process become clustered because functionally unrelated intervening genes become useless and are deleted following a transfer, and because such clustered genes are more likely to be successfully transferred as a unit compared to genes encoded in distant parts of the genome.

ORFANS
Open reading frames that do not have a recognizable homologue among known sequences.

RIBOTYPE
In analogy to genotype and phenotype, the type of RNA in an organism, usually referring to the type of ribosomal RNA.

SELECTIVE SWEEP
Fixation of an advantageous character in a population. In the absence of recombination, the advantageous character carries with it the whole chromosome and erases diversity within the population.

MORONS
Genetic elements in lambdoid phages acquired through recent gene transfer. These genes are unrelated in function to the genes surrounding them. They were named morons because their addition to the phage genome means that there is 'more DNA' than there is without the element.

The theory for prokaryotic (and perhaps microbial) evolution needs further work: a population genetics and molecular evolution theory for organisms that have no traditional species boundaries, but share genes across considerable evolutionary distance[87]. Clearly, prokaryotes, and possibly single-celled eukaryotes, are under different selective pressures with regard to DNA exchange than most multicellular organisms. Unlike animals, for instance, prokaryotes and single-celled eukaryotes experience recombination less frequently than they reproduce, and the quantity of DNA exchanged is small[96,97]. Consequently, there is little selective advantage in preventing such rare interspecific exchange: the selection coefficient, even if all transfers were completely lethal, could be no larger than the rate of recombination. With potential rates of recombination as low as 10^{-7} to 10^{-10} per generation, and selection coefficients against recombination itself constrained to be this low (if always lethal) or lower (if frequently neutral), it seems reasonable to assume that only the weakest of selection has operated on uptake mechanisms as a direct consequence of the horizontal transfer of deleterious DNA fragments.

The main difference between prokaryotes and multicellular eukaryotes is that the species boundaries for prokaryotes are 'fuzzy'[38,98]. Homologous recombination is not limited to genes exchanged within a species, and ILLEGITIMATE RECOMBINATION can incorporate genes from divergent donors. If a novel gene arises that provides a selective advantage, this invention can be shared between unrelated organisms through HGT. Recombination rates subsequent to interspecies transfer might need to be high in some diverse recipient species, because otherwise the observed recipient species diversity is inexplicable, as it would have been wiped out by selective sweeps in which the advantageous gene carries with it the complete genome[93,99]. The group within which innovations can be exchanged will be different for different genes: some genes will be advantageous only within the environment in which they originated, whereas others will provide a selective advantage even if the gene is transferred across domain boundaries. One result of the larger exchange groups that are created through HGT is an accelerated rate of innovation. Jain *et al.*[23] estimate that the innovation rate increases 10^4- to 10^{10}-fold owing to HGT, and Townsend *et al.*[86] show that recombination across traditional species lines can potentially accelerate the acquisition of adaptively important traits, requiring several amino acid changes by similarly large magnitudes.

The evolution of the *hsp70* (*dnaK*) gene provides an example of an innovation that is apparently spreading to divergent organisms, and it illustrates that gene histories can be different from the history of organismal evolution. Homologues of *hsp70* are found in members of all three domains of life[44], but on closer inspection, there is no other evidence to support the notion that *hsp70* was present in the most recent common ancestor of all organisms. In molecular phylogenies, the archaeal homologues are interspersed within the Bacteria, and many Archaea, including the Crenarchaeota, do not encode an *hsp70* orthologue in their genome. It therefore seems likely that this gene was absent in the archaeal ancestor, and was only acquired more recently by some of the Archaea by HGT[43,45,46]. The most recent common ancestor of all present-day *hsp70* genes seems to have existed more recently than the most recent common ancestor of all organisms.

The conclusion that different genes coalesce to different molecular ancestors and that these molecular ancestors existed in different organismal lineages and at different times is not limited to *hsp70*, but is possibly true for all gene families[17]. Several studies have determined the frequency with which genes belonging to different functional categories are being transferred[21,27,100], frequently using the COG[101] (clusters of orthologous groups) classification, and some environmental and genome characteristics were studied with respect to the influence they have on gene-transfer frequency[23]. To more realistically reconstruct the interplay between HGT and vertical inheritance, more detailed quantitative studies are needed to determine the factors that govern HGT frequency.

Gogarten *et al.*[19] point out that the frequency of successful exchange between taxa will depend on five factors: propinquity, metabolic compatibility, adaptations to their abiotic environment, gene expression systems and gene-transfer mechanisms. At present, most of the evidence regarding the relative importance of these factors is anecdotal, and the only systematic comparative study[23] had limited power, owing to sparse taxon sampling. All of these factors correlate with genetic relatedness and therefore DNA-sequence divergence. For instance, in the context of homology-assisted heterologous recombination, there is a well characterized quantitative effect that greater DNA-sequence divergence results in lower homologous recombination rates in *E. coli*[102], *Bacillus subtilis*[103,104] and *Streptococcus pneumoniae*[105]. Lawrence and Hendrickson[106] characterized short oligonucleotide sequences with asymmetric distribution on the LEADING AND LAGGING DNA STRAND. These sequences might have a role in genome replication, and they are conserved only between closely related organisms, whereas in distantly related organisms the same motif occurs abundantly on either DNA strand. A sequence from a phylogenetically distant donor that contains a recipient's regulatory motif on both strands might incur a selective disadvantage, effectively biasing successful transfers towards more closely related organisms. These observations could be incorporated into a theoretical framework for the evolution of microorganisms that incorporates HGT and relies on DNA-sequence divergence as a quantitative barrier instead of species designation as a qualitative barrier to recombination between microorganisms[86]. Such a framework requires quantitative characterization of the environmental density

ILLEGITIMATE RECOMBINATION
Recombination between two non-homologous DNA segments. Usually, illegitimate recombination is fairly infrequent.

LEADING AND LAGGING DNA STRAND
DNA replication on the leading strand occurs continuously in a 5′ to 3′ direction, whereas DNA replication on the lagging strand occurs discontinuously through the synthesis of short Okazaki fragments.

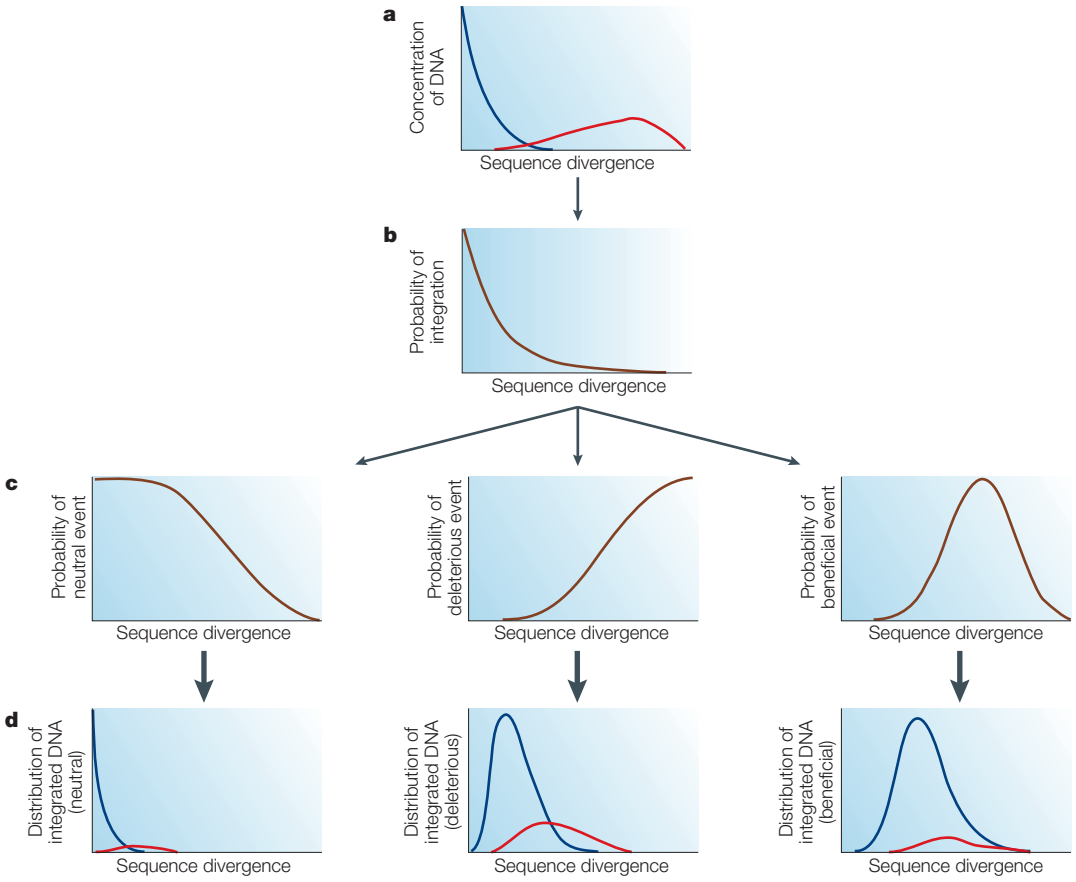

Figure 4 | **Graphical schema for the quantification and modelling of recombination with divergent DNA and horizontal gene transfer.** To develop a 'population genetics' theory for microorganisms that take up divergent DNA, sequence divergence can be considered as a mathematically continuous measure of the species barrier. **a** | Quantification of the environmental distribution of divergent DNA (homogeneous DNA in blue, heterogeneous DNA in red). **b** | Quantification of the probability of integration of divergent DNA into the genome. **c** | Quantification of the probable selective effects of that integration. These three graphs illustrate several hypotheses regarding donor DNA sequence of particular divergences. First, that a sequence with neutral or nearly neutral effects is particularly likely to come from donors that are not highly divergent. Second, that the deleteriousness of incorporated DNA, by contrast, is increasingly likely with divergence. Lastly, that most (rare) beneficial incorporations of DNA will come from some intermediate level of divergence. Because each of these distributions (a–c) have the same *x* axis (sequence divergence), they can be visually or analytically multiplied (a×b×c) to form an evolutionary model, leading to **d** | quantitative predictions relating to the effects of horizontal gene transfer on microbial evolution.

of divergent DNA, the probability of chromosomal integration in the organisms of interest and the selective value of novel microbial traits (FIG. 4; see also the article by C.M. Thomas & K.M. Nielsen in this issue for a more mechanistic discussion of these factors). All of these quantitative characterizations are vital to the further understanding and development of a comprehensive 'population genetics' theory for microbial communities. Such a formalism should encompass gene flow among divergent microorganisms. Recent studies of population-level sequence variation in the archaea *Sulfolobus islandicus* (R.J. Whitaker, D. Grogan & J.W. Taylor, unpublished results) and *Halorubrum* sp.[107] provide evidence of extensive intra- and interpopulation recombination. Quantitative models of HGT need to be constructed and tested against the actual sequence variation observed in such microbial populations.

In addition to surveys of extant variation, construction of useful models of the effect of gene transfer on evolving populations requires experimental work on the ecological effects and selection coefficients associated with non-core genes. Ultimately, not only will the interplay between new survey data, experimental data and models of gene transfer serve to elucidate the dynamics of gene transfer, it will also clarify the degree to which gene transfer has impacted the evolution of microorganisms and will help refine methods for the detection of horizontally transferred genes. Such refinements will help to resolve the ambiguity between HGT and shared ancestry as causes for the patterns that we describe as shared ancestry with phylogenetic trees. Models of HGT should be developed and tested against growing datasets to distinguish between these alternatives.

1. Lamarck, J.-B. *Histoire Naturelle des Animaux sans Vertèbres* (Verdière, Paris, 1815).
2. Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (John Murray, London, 1859).
3. Darwin, C. *Charles Darwin's Notebooks, 1836–1844* (Cornell University Press, Ithaca, 1987).
Annotated version of Darwin's notebooks, including facsimiles of many pages. Gives an insight into Darwin's thoughts as he formulated the idea of evolution by natural selection.
4. Gogarten, J. P. & Olendzenski, L. Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.* **9**, 630–636 (1999).
5. Ehrenreich, A. & Widdel, F. Anaerobic oxidation of ferrous iron by purple bacteria, a new type of phototrophic metabolism. *Appl. Environ. Microbiol.* **60**, 4517–4526 (1994).
6. Ritz, C. M., Schmuths, H. & Wissemann, V. Evolution by reticulation: European dogroses originated by multiple hybridization across the genus *Rosa*. *J. Hered.* **96**, 4–14 (2005).
7. Sang, T., Crawford, D. & Stuessy, T. Documentation of reticulate evolution in peonies (*Paeonia*) using internal transcribed spacer sequences of nuclear ribosomal DNA: implications for biogeography and concerted evolution. *Proc. Natl Acad. Sci. USA* **92**, 6813–6817 (1995).
8. Fuertes Aguilar, J., Rossello, J. A. & Nieto Feliner, G. Molecular evidence for the compilospecies model of reticulate evolution in *Armeria* (Plumbaginaceae). *Syst. Biol.* **48**, 735–754 (1999).
9. Margulis, L. *Symbiosis in Cell Evolution: Microbial Communities in the Archean and Proterozoic Eons* (Freeman, New York, 1995).
10. Margulis, L. & Sagan, D. *Acquiring Genomes: A Theory of the Origin of Species* (Basic Books, New York, 2002).
11. Kropotkin, P. A. *Mutual Aid; a Factor of Evolution* (William Heinemann, London, 1902).
12. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74**, 5088–5090 (1977).
13. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51**, 221–271 (1987).
14. Hilario, E. & Gogarten, J. P. Horizontal transfer of ATPase genes — the tree of life becomes a net of life. *Biosystems* **31**, 111–119 (1993).
An early description of the potentially net-like history of genome evolution.
15. Welch, R. A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **99**, 17020–17024 (2002).
An excellent example of how fully sequenced genomes can be used to understand variation in gene content among closely related organisms.
16. Hartman, H. & Fedorov, A. The origin of the eukaryotic cell: a genomic investigation. *Proc. Natl Acad. Sci. USA* **99**, 1420–1425 (2002).
17. Zhaxybayeva, O. & Gogarten, J. P. Cladogenesis, coalescence and the evolution of the three domains of life. *Trends Genet.* **20**, 182–187 (2004).
Advocates a population-genetics framework that incorporates HGT for the understanding of organismal lineages.
18. Martin, W. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays* **21**, 99–104 (1999).
19. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19**, 2226–2238 (2002).
20. Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y. & Blankenship, R. E. Whole-genome analysis of photosynthetic prokaryotes. *Science* **298**, 1616–1620 (2002).
21. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 3801–3806 (1999).
22. Zhaxybayeva, O., Lapierre, P. & Gogarten, J. P. Genome mosaicism and organismal lineages. *Trends Genet.* **20**, 254–260 (2004).
23. Jain, R., Rivera, M. C., Moore, J. E. & Lake, J. A. Horizontal gene transfer accelerates genome innovation and evolution. *Mol. Biol. Evol.* **20**, 1598–1602 (2003).
24. Olendzenski, L., Zhaxybayeva, O., and Gogarten, J. P. What's in a Tree? Does Horizontal Gene Transfer Determine Microbial Taxonomy? in *Cellular Origin and Life in Extreme Habitats. Vol. 4: Symbiosis* (ed. Seckbach, J.) 67–78 (Kluwer Academic Publishers, Netherlands, 2001).
25. Bocchetta, M., Gribaldo, S., Sanangelantoni, A. & Cammarano, P. Phylogenetic depth of the bacterial genera *Aquifex* and *Thermotoga* inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. *J. Mol. Evol.* **50**, 366–380 (2000).
26. Wolf, Y. I., Rogozin, I. B., Grishin, N. V. & Koonin, E. V. Genome trees and the tree of life. *Trends Genet.* **18**, 472–479 (2002).
27. Nelson, K. E. *et al.* Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329 (1999).
28. Gophna, U., Doolittle, W. F. & Charlebois, R. L. Weighted genome trees: refinements and applications. *J. Bacteriol.* **187**, 1305–1316 (2005).
An illustration that high rates of HGT can impact the inference of lineage history from comparative genomic data.
29. Daubin, V., Gouy, M. & Perriere, G. Bacterial molecular phylogeny using supertree approach. *Genome Inform. Ser. Workshop Genome Inform.* **12**, 155–164 (2001).
30. Kibak, H., Taiz, L., Starke, T., Bernasconi, P. & Gogarten, J. P. Evolution of structure and function of V-ATPases. *J. Bioenerg. Biomembr.* **24**, 415–424 (1992).
31. Brochier, C. & Philippe, H. Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature* **417**, 244 (2002).
32. van Berkum, P. *et al.* Discordant phylogenies within the *rrn* loci of Rhizobia. *J. Bacteriol.* **185**, 2988–2998 (2003).
33. Yap, W. H., Zhang, Z. & Wang, Y. Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J. Bacteriol.* **181**, 5201–5209 (1999).
34. Wang, Y., Zhang, Z. & Ramanan, N. The actinomycete *Thermobispora bispora* contains two distinct types of transcriptionally active 16S rRNA genes. *J. Bacteriol.* **179**, 3270–3276 (1997).
35. Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. & Polz, M. F. Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *J. Bacteriol.* **186**, 2629–2635 (2004).
36. Archibald, J. M. & Roger, A. J. Gene duplication and gene conversion shape the evolution of archaeal chaperonins. *J. Mol. Biol.* **316**, 1041–1050 (2002).
37. Lawrence, J. G. & Ochman, H. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA* **95**, 9413–9417 (1998).
38. Lawrence, J. G. Gene transfer in bacteria: speciation without species? *Theor. Pop. Biol.* **61**, 449–460 (2002).
This article outlines a 'fuzzy' prokaryotic species concept that incorporates HGT. See reference 96 for an alternate conception.
39. Ragan, M. A. Reconciling the many faces of lateral gene transfer. *Trends Microbiol.* **10**, 4 (2002).
40. Lawrence, J. G. & Ochman, H. Reconciling the many faces of lateral gene transfer. *Trends Microbiol.* **10**, 1–4 (2002).
41. Daubin, V., Lerat, E. & Perriere, G. The source of laterally transferred genes in bacterial genomes. *Genome Biol.* **4**, R57 (2003).
Shows that frequently horizontally transferred genes tend to have particular properties.
42. Koski, L. B., Morton, R. A. & Golding, G. B. Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.* **18**, 404–412 (2001).
43. Philippe, H., Budin, K. & Moreira, D. Horizontal transfers confuse the prokaryotic phylogeny based on the HSP70 protein family. *Mol. Microbiol.* **31**, 1007–1009 (1999).
44. Gupta, R. S. & Golding, G. B. Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes. *J. Mol. Evol.* **37**, 573–582 (1993).
45. Gogarten, J. P. Which is the most conserved group of proteins? Homology–orthology, paralogy, xenology, and the fusion of independent lineages. *J. Mol. Evol.* **39**, 541–543 (1994).
46. Gribaldo, S. *et al.* Discontinuous occurrence of the *hsp70* (*dnaK*) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. *J. Bacteriol.* **181**, 434–443 (1999).
47. Sicheritz-Ponten, T. & Andersson, S. G. A phylogenomic approach to microbial evolution. *Nucleic Acids Res.* **29**, 545–552 (2001).
An effective and popular methodology to analyse comparative genomic data within a phylogenetic context.
48. Frickey, T. & Lupas, A. N. PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Res.* **32**, 5231–5238 (2004).
49. Loftus, B. *et al.* The genome of the protist parasite *Entamoeba histolytica*. *Nature* **433**, 865–868 (2005).
50. Huang, J., Mullapudi, N., Sicheritz-Ponten, T. & Kissinger, J. C. A first glimpse into the pattern and scale of gene transfer in Apicomplexa. *Int. J. Parasitol.* **34**, 265–274 (2004).
51. Koski, L. B. & Golding, G. B. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **52**, 540–542 (2001).
52. Graybeal, A. Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. *Syst. Biol.* **43**, 174–193 (1994).
53. Zhaxybayeva, O. & Gogarten, J. P. An improved probability mapping approach to assess genome mosaicism. *BMC Genomics* **4**, 37 (2003).
54. Morowitz, H. *The Wine of Life, and Other Essays on Societies, Energy and Living Things* (Bantam Books Inc., New York, 1979).
55. Snel, B., Bork, P. & Huynen, M. A. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* **12**, 17–25 (2002).
56. Daubin, V., Moran, N. & Ochman, H. Phylogenetics and the cohesion of bacterial genomes. *Science* **301**, 829–832 (2003).
57. Mirkin, B. G., Fenner, T. I., Galperin, M. Y. & Koonin, E. V. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3**, 2 (2003).
58. Novichkov, P. S. *et al.* Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. *J. Bacteriol.* **186**, 6575–6585 (2004).
An early attempt to use data-mining techniques to quantify rates of horizontal and vertical inheritance.
59. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
The gold standard for demonstration of phylogenetic conflict.
60. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116 (1999).
61. Lerat, E., Daubin, V. & Moran, N. A. From gene trees to organismal phylogeny in prokaryotes: the case of the γ-Proteobacteria. *PLoS Biol.* **1**, E19 (2003).
62. Brochier, C., Bapteste, E., Moreira, D. & Philippe, H. Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet.* **18**, 1–5 (2002).
63. Bapteste, E., Boucher, Y., Leigh, J. & Doolittle, W. F. Phylogenetic reconstruction and lateral gene transfer. *Trends Microbiol.* **12**, 406–411 (2004).
64. Daubin, V. & Ochman, H. Quartet mapping and the extent of lateral transfer in bacterial genomes. *Mol. Biol. Evol.* **21**, 86–89 (2004).
65. Lento, G. M., Hickson, R. E., Chambers, G. K. & Penny, D. Use of spectral analysis to test hypotheses on the origin of pinnipeds. *Mol. Biol. Evol.* **12**, 28–52 (1995).
66. Bininda-Emonds, O. R. P. The evolution of supertrees. *Trends Ecol. Evol.* **19**, 315–322 (2004).
67. Wang, L., Zhang, K. & Zhang, L. Perfect phylogenetic networks with recombination. *J. Comput. Biol.* **8**, 69–78 (2001).
68. Nakhleh, L., Warnow, T. & Linder, C. R. Reconstructing reticulate evolution in species — theory and practice. in *Conference on Research in Computational Molecular Biology* 337–346 (RECOMB, San Diego, 2004).
69. Gusfield, D., Eddhu, S. & Langley, C. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *J. Bioinform. Comput. Biol.* **2**, 173–213 (2004).
70. Bandelt, H. J. & Dress, A. W. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.* **1**, 242–252 (1992).
71. Huson, D. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).
72. Bryant, D. & Moulton, V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21**, 255–265 (2004).
73. Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K. & Schuster, S. C. Whole-genome prokaryotic phylogeny. *Bioinformatics* **21**, 2329–2335 (2004).
74. Lake, J. A. & Rivera, M. C. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol. Biol. Evol.* **21**, 681–690 (2004).
75. Rivera, M. C. & Lake, J. A. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**, 152–155 (2004).
76. Andersson, J. O., Sjogren, A. M., Davis, L. A., Embley, T. M. & Roger, A. J. Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. *Curr. Biol.* **13**, 94–104 (2003).

77. Andersson, J. O., Sarchfield, S. W. & Roger, A. J. Gene transfers from nanoarchaeota to an ancestor of diplomonads and parabasalids. *Mol. Biol. Evol.* **22**, 85–90 (2005).
78. Doolittle, W. F. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* **14**, 307–311 (1998).
79. Gogarten, J. P. Gene transfer: gene swapping craze reaches eukaryotes. *Curr. Biol.* **13**, R53–R54 (2003).
80. Martin, W. Gene transfer from organelles to the nucleus: frequent and in big chunks. *Proc. Natl Acad. Sci. USA* **100**, 8612–8614 (2003).
81. Kurland, C. G., Canback, B. & Berg, O. G. Horizontal gene transfer: a critical view. *Proc. Natl Acad. Sci. USA* **100**, 9658–9662 (2003).
82. Lawrence, J. G. & Roth, J. R. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**, 1843–1860 (1996).
83. Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. Genomic islands in pathogenic and environmental microorganisms. *Nature Rev. Microbiol.* **2**, 414–424 (2004).
84. Gogarten, J. P., Senejani, A. G., Zhaxybayeva, O., Olendzenski, L. & Hilario, E. Inteins: structure, function, and evolution. *Annu. Rev. Microbiol.* **56**, 263–287 (2002).
85. Nielsen, K. M. & Townsend, J. P. Environmental exposure, horizontal transfer, and selection of transgenes in bacterial populations. in *Enhancing Biocontrol Agents and Handling Risks, Vol. 339* (eds. Vurro, M. *et al.*) 145–158 (IOS Press, Amsterdam, 2001).
86. Townsend, J. P., Nielsen, K. M., Fisher, D. S. & Hartl, D. L. Horizontal acquisition of divergent chromosomal DNA: consequences of mutator phenotypes. *Genetics* **164**, 13–21 (2003).
Puts forward a quantitative framework for evaluating the effect of HGT on the evolution of prokaryotes.
87. Nielsen, K. M. & Townsend, J. P. Monitoring and modeling horizontal gene transfer. *Nature Biotechnol.* **22**, 1110–1114 (2004).
88. Hooper, S. D. & Berg, O. G. Duplication is more common among laterally transferred genes than among indigenous genes. *Genome Biol.* **4**, R48 (2003).
Shows that frequently horizontally transferred genes tend to have particular properties.
89. Daubin, V. & Ochman, H. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res.* **14**, 1036–1042 (2004).
90. Philippe, H. & Douady, C. J. Horizontal gene transfer and phylogenetics. *Curr. Opin. Microbiol.* **6**, 498–505 (2003).
91. Hightower, L. E. Heat shock, stress proteins, chaperones, and proteotoxicity. *Cell* **66**, 191–197 (1991).
92. Thompson, J. R. *et al.* Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307**, 1311–1313 (2005).
93. Majewski, J. & Cohan, F. M. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. *Genetics* **152**, 1459–1474 (1999).
94. Hendrix, R. W., Lawrence, J. G., Hatfull, G. F. & Casjens, S. The origins and ongoing evolution of viruses. *Trends Microbiol.* **8**, 504–508 (2000).
95. Mayr, E. *Systematics and the Origin of Species* (Columbia University Press, New York, 1942).
96. Cohan, F. M. What are bacterial species? *Annu. Rev. Microbiol.* **56**, 457–487 (2002).
Outlines a species concept defined by selective sweeps. For an alternate conception, see reference 38.
97. Majewski, J. Sexual isolation in bacteria. *FEMS Microbiol. Lett.* **199**, 161–169 (2001).
98. Hanage, W. P., Fraser, C. & Spratt, B. G. Fuzzy species among recombinogenic bacteria. *BMC Biol.* **3**, 6 (2005).
99. Cohan, F. M. Sexual isolation and speciation in bacteria. *Genetica* **116**, 359–370 (2002).
100. Zhaxybayeva, O. & Gogarten, J. P. Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses. *BMC Genomics* **3**, 4 (2002).
101. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
102. Vulic, M., Dionisio, F., Taddei, F. & Radman, M. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc. Natl. Acad. Sci. USA* **94**, 9763–9767 (1997).
103. Zawadzki, P., Roberts, M. S. & Cohan, F. M. The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust. *Genetics* **140**, 917–932 (1995).
104. Majewski, J. & Cohan, F. M. The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*. *Genetics* **148**, 13–18 (1998).
105. Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M. & Dowson, C. G. Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. *J. Bacteriol.* **182**, 1016–1023 (2000).
106. Lawrence, J. G. & Hendrickson, H. Lateral gene transfer: when will adolescence end? *Mol. Microbiol.* **50**, 739–749 (2003).
107. Papke, R. T., Koenig, J. E., Rodriguez-Valera, F. & Doolittle, W. F. Frequent recombination in a saltern population of *Halorubrum*. *Science* **306**, 1928–1929 (2004).

Acknowledgements

We thank O. Zhaxybayeva, K. Nielsen, P. Lapierre, W.F. Doolittle, J. Lawrence and F.M. Cohan for many stimulating and relevant discussions. Work in J.P.G.'s laboratory was supported by the National Science Foundation's Microbial Genetics programme and the NASA Applied Information Systems Research and Exobiology Programmes.

Competing interests statement

The authors declare no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:
Entrez: http://www.ncbi.nlm.nih.gov/Entrez
Archaeoglobus fulgidus | *Bacillus subtilis* | *Methanocaldococcus jannaschii* | *Methanothermobacter thermautotrophicus* | *Pyrococcus abyssi* | *Pyrococcus furiosus* | *Streptococcus pneumoniae* | *Streptococcus thermophilus* | *Thermotoga maritima*

FURTHER INFORMATION

J. Peter Gogarten's laboratory: http://gogarten.uconn.edu
Jeffrey P. Townsend's laboratory:
http://web.uconn.edu/townsend
Access to this interactive links box is free online.