Horizontal gene transfer and phylogenetics

Hervé Philippe* and Christophe J Douady†

The initial analysis of complete genomes has suggested that horizontal gene transfer events are very frequent between microorganisms. This could potentially render the inference, and even the concept itself, of the organismal phylogeny impossible. However, a coherent phylogenetic pattern has recently emerged from an analysis of about a hundred genes, the so-called 'core', strongly suggesting that it is possible to infer the phylogeny of prokaryotes. Also, estimation of the frequency of horizontal gene transfers at the genome level in a phylogenetic context seems to indicate that it is rather low, although of significant biological impact. Nevertheless, it should be emphasized that the history of microorganisms cannot be properly represented by the phylogeny of the core, which represents only a tiny fraction of the genome. This history, even if horizontal gene transfers are rare, should be represented by a network surrounding the core phylogeny.

Addresses
*Canadian Institute for Advanced Research, Département de Biochimie, Université de Montréal, Montréal, Qc., H3C 3J7, Canada
†Genome Atlantic, and Department of Biochemistry & Molecular Biology, Dalhousie University, Sir Charles Tupper medical building, Halifax, NS, B3H 4H7, Canada
Correspondence: Hervé Philippe
e-mail: herve.philippe@umontreal.ca

Current Opinion in Microbiology 2003, **6**:498–505

This review comes from a themed issue on
Genomics
Edited by Eduardo A Groisman and S Dusko Ehrlich

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2003.09.008

Abbreviations
COG clusters of orthologous groups
HGT horizontal gene transfer
MY million years

Introduction

Horizontal (or lateral) gene transfer (HGT) is commonly assumed to be an important, if not the dominant, force in shaping prokaryotic genomes [1–5,6•]. However, in spite of a tremendous number of publications, its actual contribution and consequences remain highly debated [7•,8•]. In fact, since its first recognition the assumed role of HGTs in prokaryotic evolution has changed many times [1]. But in the late nineties, the discovery of potentially massive gene exchanges in completely sequenced genomes shifted the paradigm again [9]. This resurrection was based mainly on the observation of deviant nucleotide composition and anomalous phylogenetic distribution (in particular, variation of gene content between closely related organisms) suggesting that a large proportion of genes (up to 20%) can be rapidly exchanged [10,11•,12•]. Therefore, if all the genes are equally subjected to HGT, this implies that foreign DNA could replace a complete genome in only a few hundred million years [13], erasing any evidence of ancient vertical inheritance. Doolittle followed this idea through and suggested that 'the history of life cannot be properly represented as a tree' [9]. Following this statement, numerous reviews and commentary papers were devoted to this subject (e.g. [1–5,14,15•,16,17]) and some extreme viewpoints were proposed including 'it is clear that genes have flowed through the biosphere, as in a global organism' [18]. However, until recently, relatively few primary papers have addressed the fundamental questions of how to detect an HGT, how many HGTs have occurred, and whether all the genes/taxa are equally affected?

In this paper, we first review the recent progress in detecting and quantifying HGTs. We focus then on HGT impacts and consequences on phylogenetics, and conclude by reviewing evidence against, and in favor, of a universal tree of life.

How to detect horizontal gene transfers

Most of the methods for detecting HGT fall into four categories: deviant composition, anomalous phylogenetic distribution, abnormal sequence similarity (i.e. greatest similarity with a gene from a distantly related species) and incongruent phylogenetic trees (reviewed in [16,19,20]). However, the number of overlapping putatively transferred genes detected by these approaches can be smaller than that expected by chance alone [21]. It is possible that the variability in outcome of these methods demonstrates more a difference in the targeted HGTs than inconsistency. For example, it is not surprising that HGTs revealed by phylogenetic analyses and involving orthologous gene replacement will not be detected by methods focusing on anomalous phylogenetic distribution. Likewise, we cannot expect to detect, via deviant GC content methods, HGTs between two very distant GC-rich bacteria. Thus, depending on the methods, different types and relative ages of HGTs will be detected [21,22].

Current methods can also be flawed by many false-positives. For example, intragenomic base content variations may bias detection based on GC content abnormality [23,24•]. When considering HGTs detected through BLAST similarity, variability in the rates of evolution

or even genome size have been shown to greatly affect results [25]. Koski and Golding [26] even stated that 'genes appearing to be the most similar based on BLAST hits are often not each others closest relative phylogenetically'. Similarly, products of phylogenetic based methods may be incongruent just because of tree reconstruction artifacts. For example, genes encoded by plastid genomes, for which HGTs are highly unlikely, yield significantly incongruent phylogenies [27]. Thus, the most reliable inference of recent HGTs is where a gene is present in one genome but not in several closely related genomes (anomalous phylogenetic distribution), because assuming a vertical inheritance would imply several independent losses. However, even if this method is unlikely to be prone to artifacts, orthologous replacement (i.e. one gene being supplanted by a homolog of a different origin) will not be detected and operators will still have to decide how many independent losses are too many. More importantly, HGTs are as such very recent and so it is difficult to know if they will be fixed in the population (see below). Albeit difficult in practice, phylogenetic inference remains the best way to detect fixed HGT.

As trivial as this might seem, we should always keep in mind that most of the genes contain too little signal to identify HGTs. In particular, many genes have evolved so rapidly that they simply cannot be aligned even between closely related organisms, others contain only a limited conserved region, which contains little phylogenetic signal and finally others are present in less than four taxa (the minimum number of taxa required to have potentially different trees, as under this number at most one tree exists). Therefore, all phylogenetic analyses, even at the genome scale, will deal with a small portion of genes contained in the genome and not address the frequency of HGTs for the majority of the genes.

How to quantify horizontal gene transfers

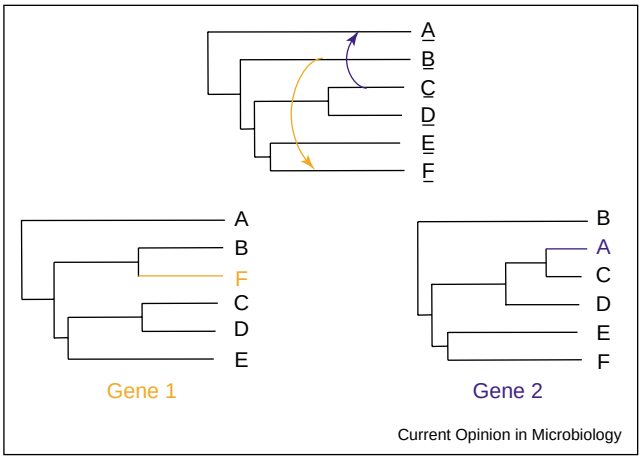
Two recent studies have tried to assess the amount of HGTs using anomalous phylogenetic pattern techniques [7•,8•]. The first analyzed six Archaea and eleven Proteobacteria [7•] while the second examined 26 taxa spanning the three domains of life [8•]. Both approaches used parsimony-based criterion to identify the respective number of genes vertically inherited, lost and laterally transmitted. After selecting all sets of orthologous genes contained by these organisms (custom made [7•] or extracted from the clusters of orthologous groups of proteins (COG) database [8•]), they reconstructed for each gene the most parsimonious evolutionary scenario (i.e. requiring the least loss and HGT). Whatever the penalty given to HGT (relative to gene loss), the number of horizontally transferred genes was found to be less than 10% of the number of vertically inherited genes [7•]. Similarly, for the set of 26 taxa, a mean of slightly more than one HGT per gene family was computed [8•]. However, it is more surprising that they produced fairly different results considering HGT versus gene loss predominance. Snel and co-workers [7•] for instance concluded that 'although it is necessary to invoke HGT to explain the content of present day genomes, gene loss, gene genesis and simple vertical inheritance are quantitatively the most dominant processes in shaping the genome'. Alternatively, Mirkin *et al.* [8•] stated that 'their results seem to be compatible with approximately equal likelihoods of HGT and gene loss in the evolution of prokaryotes'.

Obviously, these two analyses are rather simplistic in their design. First, vertical inheritance is supposed to have no cost, a sensible assumption but one that should prevent the assessment of frequency of HGT over vertical descent. Second, and perhaps more importantly, branch lengths of the organismal topology and the history of the gene family are not taken into account, nor are orthologous replacements. Most of these limitations, inherent to parsimony methods, could be dealt with by using Likelihood or Bayesian models (i.e. for any given gene, the cost of an event could be inversely proportional to its frequency and events could be made more likely along long branches). Furthermore, it remains possible that their conclusions cannot be generalized and only represent particularities of their data (taxa and set of orthologous genes). However, they definitively represent interesting new and global ways to decipher between the multiple forces of prokaryotic evolution.

Major impact of horizontal gene transfer on phylogeny

It should be emphasized that HGTs, even if they are rare, have a major impact on phylogenetic inference. Figure 1

Figure 1



Phylogenetic incongruities generated by horizontal transfers. Orange and Blue arrows highlight two hypothetical horizontal gene transfers (HGTs) affecting genes 1 and 2, respectively. As illustrated in the bottom part, the resulting phylogenies will be completely incongruent (no partition in common), although the number of HGTs invoked is limited.

shows two genes that have followed exactly the same history except for one HGT in each. As a result, phylogenies inferred from gene 1 and from gene 2 are completely incongruent (i.e. there is no grouping of taxa or 'partition' in common). Moreover, in practice, the level of incongruities is amplified by the effect of tree reconstruction artifacts. Therefore, contrary to common belief, observing highly incongruent phylogenies is not obligatory evidence of massive HGTs. Because of the major impact of HGTs on phylogenetic structure, even a small amount of transfers can be highly problematic and could be sufficient to prevent the use of a tree to represent the evolution of prokaryotes.

If a gene is transferred into a genome that contains a homolog (not necessarily an ortholog), it is possible that the two copies undergo homologous recombination, leading to a chimerical gene. In this case, a network is the valid mathematical representation of the evolution of this gene. This problem is well known in population biology, particularly for viruses (e.g. [28]). Several methods have been devised for detecting recombination and appear to be efficient, especially on a small evolutionary scale (see [29] for review). Homologous recombination mainly occurs between members of a population and this should not seriously affect phylogenetic inference at high taxonomic level.

However, reality is much more complex because some gene portions are highly conserved, because of strong functional pressure, and homologous genes from very distantly related organisms can therefore undergo homologous recombination. This is particularly problematic because these highly conserved genes (e.g. rRNA, elongation factors and HSP70) are often used to infer organismal phylogeny, precisely because of this conservation that facilitates alignment. Only very few examples of recombination between distantly related genes have been reported so far: for chaperonin within Archaea [30], for rRNA within high GC Gram-positive bacteria [31] and Rhizobia [32]. Nevertheless, one should note that this kind of recombination is very difficult to detect. For instance, a highly conserved segment of only 100 nucleotides recombined for a rRNA gene will generate very few (e.g. five) positions showing an incongruent phylogenetic pattern over 1500 positions. This number can be of the same order of magnitude that is expected randomly by multiple independent substitutions (homoplasy) and it will be almost impossible to find significant evidence in favor of the recombination hypothesis. In practice, recombination is much easier to recover when at least a divergent region is involved. Thus, recombination is likely to be a minor problem when inferring phylogeny between very distantly related organisms, but surely adds some noise in the dataset. However, as shown recently [32], because of recombination, it is not possible to infer organismal phylogeny from a single gene.

Because of HGTs, it has been stated that 'the history of life cannot properly be represented as a tree' [9]. However, it is important to look at not only the evolution of the genomes but also the evolution of membranes [33]. The fusion of organisms is highly exceptional (if we exclude hybridization between closely related taxa). Therefore, at the level of the membranes, the history of life has to be represented as a tree and debate centers on whether it is possible to infer the membrane-phylogeny (or envelope phylogeny) from information contained in the genome and on whether it is meaningful to use the membrane-phylogeny if HGTs are rampant.

An emerging picture

Several phylogenetic analyses at the genome level (or at large scale) have been performed and, with a few exceptions, lead to the inference of a phylogenetic structure in good agreement with the one inferred from rRNA. These analyses can be classified in two categories depending on the primary information used: gene presence/absence or gene sequence comparison.

Gene-based evidence

In the first category, orthologous genes are detected through BLAST search (e.g. the COG database [34]) and constitute the elements of the data matrix. The information for inferring phylogeny is the presence/absence of genes (gene content) or of pairs of orthologous genes (gene order) and the tree can be inferred with distance and maximum parsimony methods. The gene order phylogenies (which can be applied only to prokaryotes) are the least in agreement with the rRNA tree, especially at large evolutionary scale. This is due to the fact that gene order evolved rapidly, because of numerous gene rearrangements (e.g. [35]). Therefore, the saturation level is reached because of multiple rearrangements affecting the same gene and the phylogenetic signal has disappeared. HGTs also add noise to the gene order phylogeny. This could be very problematic as operons are often transferred on block [36] (only the whole operon generally provides a selective advantage by furnishing new metabolic capacities), simultaneously generating several aberrant gene pairs.

Phylogenies based on gene content usually display greater agreement with the rRNA tree. In particular, the monophyly of the three domains of life is always recovered, and the monophyly of the major phyla (e.g. Spirochaetes and low GC Gram-positive bacteria) is generally found. However, when HGTs are frequent between two particular lineages (e.g. between Thermoplasmatales and Sulfolobales or between hyperthermophilic bacteria and archaea), the gene content trees are biased and tend to cluster these groups, even if they are not closely related (as defined by small subunit [SSU] rRNA). For example, the branching order within the Bacteria is biased because of similarities between

hyperthermophilic bacteria and archaea, yielding to the early emergence of *Aquifex* and *Thermotoga* [37]. In addition, gene loss can also generate artifacts, as proposed in the case of the *Mycoplasma* [38•] and of *Caenorhabditis* [37]. Despite methodological efforts, in particular the use of homologs instead of orthologs [37], the problems in phylogenetic inference generated by differential HGTs and gene loss have not been overcome. In fact, the rare incongruities between gene content trees and the rRNA tree have been proposed as a way to detect massive HGTs [39•].

Alignment based evidence

As underlined a long time ago [40,41], the best evidence in favor of the existence of a tree of life is that different markers provide essentially the same tree. Although it was already possible to make this test at the genome level several years ago, only a few studies have recently been published, because phylogenetic analyses are difficult to automate. The first limitation is the selection of the genes. Orthologs can be automatically detected as best reciprocal BLAST hits [42] or as homologs present only once in all the studied genomes (tolerating multiple copies only when they are very similar) [43••]. These approaches have been criticized, as a closest blast hit does not imply phylogenetic proximity [26] and time-consuming phylogenetic analysis should be preferred [44••]. However, as a large number of genes are used, some errors in orthology establishment are likely to have limited impact. The second limitation concerns the alignment and the selection of the homologous segments, the so-called 'unambiguously aligned regions', and is more problematic. Alignments are performed automatically and generally refined manually. The ambiguous regions were conserved [42,45,46] and manually [44••,47] or automatically [43••] removed. This variation may have a significant impact on the phylogeny, and recent progresses, in particular in the selection of unambiguously aligned regions [48–50], will improve future studies and facilitate their comparison.

Some multigene approaches [42,45,46,51] suggest that there is no phylogenetic structure, at least among some major prokaryotic groups. They observe that the different possible topologies (three for four taxa and 15 for five taxa) are recovered by a similar number of genes. However, phylogenies with a limited number of taxa are difficult to infer, and it is common to see the very same gene supporting different topologies when the taxa sample is modified [52,53]. This cannot be leveled as a criticism of Nesbo *et al.* [51] who analyzed many of the strongly conflicting genes with many more species as a control for this. But, in the case of the results of Raymond *et al.* [46], the taxa sample problem is significant (Douady and Philippe, unpublished results). More importantly, the selected taxa often belong to groups for which rRNA suggests that their successive divergences occurred relatively rapidly. With a phylogeny displaying a short internal branch and long terminal branches, one expects that stochastic effects and also the impact of tree reconstruction artifacts are very important and therefore that different genes support different topologies, just because of their short size. Indeed, the very same pattern was observed for plastid genes, for which HGTs are very unlikely [27]. A reanalysis of the 188 genes used by Raymond *et al.* [46] shows that almost none of the phylogenies are significantly supported (Douady and Philippe, unpublished results). By contrast, if the same approach is applied to taxa for which the internal branches are long, the various genes, in their great majority, provide strong support for a single phylogenetic tree ([45] and Douady and Philippe, unpublished results). Thus, the lack of preferential support for a single phylogeny [42,45,46] could be due to the lack of phylogenetic signal (i.e. speciation events closely spaced in time) and to limitations of tree reconstruction methods, and should not be viewed as direct evidence of rampant HGT.

By contrast, other multigene approaches that are based on a much larger taxonomic sampling (about 40 taxa) yielded a very similar and congruent topology with rRNA phylogeny [43••,44••,47,54]. Combinations of information from different genes have been performed either before phylogenetic reconstruction, through concatenation of the sequences, or after, through a supertree approach. Supertree combines phylogenetic trees inferred from different sets of taxa into a single phylogeny that contains all the taxa [55]. Using 730 genes, Daubin *et al.* [43••] inferred a supertree in which all the major groups (the three domains Eukarya, Archaea and Bacteria, low GC Gram-positives, High GC Gram-positives, Spirochaeta, Proteobacteria, Crenarchaeota and Euryarchaeota) were found to be monophyletic. Using a more limited sampling of 23 genes (because of the stringent criterion requiring the presence of a single copy in all of the 45 taxa studied), Brown *et al.* [47] obtained the same results from the analysis of the concatenated sequences. However, the relationships between the major bacterial phyla were different, but always poorly supported. Brochier and co-workers [44••,54] analyzed Bacteria and Archaea separately to reduce the impact of tree reconstruction limitations (especially the long branch attraction phenomenon generated by the very long branches connecting the three domains) and focused on the proteins involved in translation. In these studies, the protein-based phylogenies were compared with trees based on large subunit (LSU) and SSU rRNA for the same set of taxa. Except for the fast evolving *Mycoplasma* and the weakly supported interphyla relationships, the two phylogenies were identical.

Consequences for prokaryotic genome evolution

The very good congruence between phylogenies depicted by Daubin, Brown and Brochier, obtained from

different datasets with slightly different approaches, strongly suggests the persistence of a phylogenetic signal, despite HGTs. However, a tree based on the concatenation of various genes infers the average of the histories contained in each gene and does not demonstrate that each gene underwent the same history. In fact, Brochier *et al.* [44••] concatenated genes that are likely to have experienced HGTs (e.g. tRNA synthetases) and the corresponding phylogeny was very similar to the rRNA tree, albeit to a lesser extent than for the tree based on a concatenation of genes selected *a priori* without HGTs (e.g. ribosomal proteins). It is therefore of prime importance to verify that each gene of the concatenation has followed the same history (except for stochastic effects and tree reconstruction artifacts). Only two methods have been developed and applied to a large set of genes. In the first, each gene is described by the likelihood values for a representative set of topologies and the results are summarized through a principal component analysis to a two-dimensional scatter plot [44••]. In the second, the Robinson-Foulds topological distance between the phylogenies inferred from two genes is computed and the resulting matrix for all the pairs of genes is displayed by a principal coordinates analysis [43••]. Both methods suggest that a significant proportion of the genes used have undergone the same phylogenetic history and that a few detectable HGTs have affected the remaining ones. Thus, it seems that genes could pertain to three categories: the hard core that is composed of genes that are never transferred (or at least not transferred at the considered scale), the soft core with genes rarely transferred (maybe about one atypical node out of 100) and the shell genes where all genes susceptible to HGTs belong and for convenience sake all genes that do not contain enough information to be classified as core genes.

Conclusions

There is a big ideological and rhetorical gap between researchers believing that HGTs are so frequent that phylogeny is useless and others believing that HGTs are rare and constitute an additional minor noise when inferring phylogeny. Clearly, the heterogeneity of genome composition between closely related strains (only 40% of the genes in common with three *E. coli* strains [11•]) and the very congruent phylogenetic structure inferred from ~100 genes represent convincing evidence in favor of these two extreme opinions, which should nevertheless be reconciled.

Clearly, a simple answer to this paradox is that these genes correspond to two different classes (the shell and the core). The core of genes that underwent no (or very few) HGTs can be used to infer what most would assume to be the 'membrane' phylogeny. However, if one wants to fully understand prokaryotic genome evolution we will need to develop new tools (such as networked trees) that allow mapping of shell-gene history onto the 'core' tree. Equating the core phylogeny with the tree of life when soft and hard core genes together are likely to represent only 1–5% of the genome is not satisfying. As Doolittle points out (personal communication) this would be equivalent of claiming that mitochondrial genes, because they have (or should have) the same phylogenies represent the true tree by which we should relate living humans.

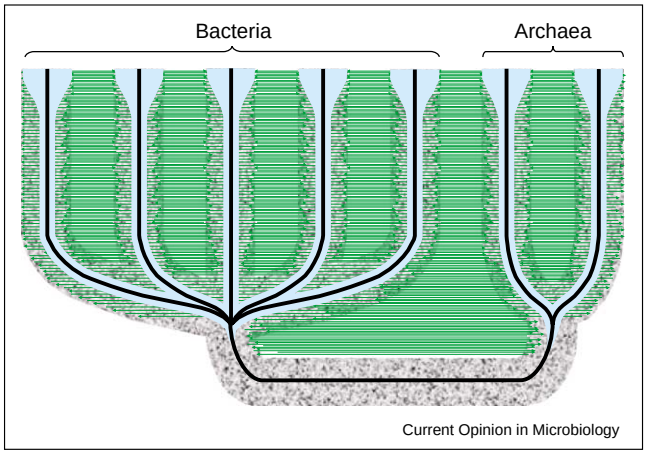
Gogarten *et al.* [15•] suggested a way to reconcile recurrent pattern and rampant HGTs. This would become possible if taxa were to exchange preferentially between themselves and that the frequency of transfer was the structural signal itself. Two taxa are more similar than a third one, not because they share a more recent common ancestor, but because they exchange genes more frequently than with the third one. In its present formulation, this hypothesis is rather difficult to test, especially as the evolution of the process over time has not been made explicit, but it provides an interesting alternative.

A contrasting view is that most of the genes contained in one genome, but not in closely related ones, are in fact only transiently present in the genome and will never be fixed in the population. In fact, even if some HGTs accelerate genome innovation [6•,10,15•], the great majority of HGTs may be neutral (and perhaps often deleterious) and the corresponding genes are rapidly lost through inactivating mutations. Interestingly, bacteriophages could constitute the reservoir for these strain-specific genes [56,57•], providing a continuous influx of genes for which the fate is almost always rapid disappearance.

In any case, it is of prime interest to estimate the frequency of fixed HGTs. There is little doubt that global research programs, such as those initiated by the groups of Huynen and Koonin [7•,8•], will work towards more realistic assumptions, making full use of the power of probabilistic phylogenetic inference methods (see [58]). It will be equally crucial to determine if HGT frequency can fluctuate over time and with which amplitude. The significance of HGTs both in terms of topological rearrangement (Figure 1) and biological impact [59] will have to be further accessed, as should be the HGT fixation rate between closely related taxa versus ecological neighbors. However, other directions such as improvement of the phylogenetic methods and the pursuit of data acquisition should remain a top priority, especially if our goal is to understand prokaryotic evolution and not only evolution of pathogenic organisms (e.g. [12•]).

In conclusion, let us assume that HGTs are rare, for instance one successful HGT per million years (MY) per lineage. During the evolution of prokaryotes (at least 2000 MY), a lineage will have acquired at least 2000 genes by HGTs. As the genome size of prokaryotes is likely to have remained fairly constant throughout time (a few thousand genes), this indicates that about half of the

Figure 2



Schematic representation of the history of prokaryotic genomes. Green arrows represent horizontal gene transfers (HGTs). The black line indicates the hard-core genes, which are likely to have the same phylogeny as the envelope during the evolution of prokaryotes. The blue area corresponds to the soft-core genes, which have only undergone very few HGTs during their entire history and are useful to infer membrane phylogeny at small and intermediate evolutionary scale. The shell genes are indicated in gray marbled texture, to indicate that phylogeny is not the correct way to display their history. Surfaces are drawn to be very roughly proportional to their size.

genome of a given lineage does not have the same history as the 'envelope phylogeny' (inferred from the core genes). If we take a much greater value of 100 successful HGTs per MY, at least 200 000 HGTs will have occurred and the history of the genome would be almost completely different from the phylogeny of the cores (Figure 2). Yet, it should be noted that a rate of 100 HGTs per MY equates to roughly one HGT for one million cell divisions (vertical inheritance), which is still extremely low. Therefore, even if recent work has been able to infer the envelope phylogeny from the core genes and if HGTs turn out to be extremely rare (one per MY), the history of prokaryotes (especially their genomes) cannot be represented by a phylogeny, but should be represented by a large network, surrounding a thin scaffold envelope (Figure 2).

Update

A recent work presented by Daubin *et al.* [60] seems to deny HGT preponderance in prokaryotic genome evolution. The analysis of seven sets of four genomes from closely related species revealed that orthologous genes yielded almost exclusively phylogenies either congruent with the rRNA tree or unresolved. Less than 1% of the phylogenies are incongruent, except for Rhizobia (5%) and Streptococci (23%), suggesting that HGTs rarely affect orthologous genes. However, a large number of genes (between 200 and 1200, depending on the set of species) have been acquired by HGTs. This is in excellent agreement with previous works that show the importance of HGTs and the existence of a core of non-transferred genes. But the analyses of Daubin *et al.* [60] exclusively focus on a very low taxonomic level (the top part of Figure 2) and extrapolating their conclusion to the history of prokaryotes may be hazardous.

Acknowledgements

We are grateful to Yan Boucher, W Ford Doolittle, David Moreira and Andrew Roger for comments and suggestions. This work was supported by Canadian Research Chairs, Genome Atlantic (an affiliate of Genome Canada) and the Killam foundation.

References and recommended reading

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

- • of special interest
- •• of outstanding interest

1. Doolittle WF, Boucher Y, Nesbo CL, Douady CJ, Andersson JO, Roger AJ: **How big is the iceberg of which organellar genes in nuclear genomes are but the tip?** *Philos Trans R Soc Lond B Biol Sci* 2003, **358**:39-57.

2. Eisen JA: **Assessing evolutionary relationships among microbes from whole-genome analysis**. *Curr Opin Microbiol* 2000, **3**:475-480.

3. Lake JA, Jain R, Rivera MC: **Mix and match in the tree of life**. *Science* 1999, **283**:2027-2028.

4. Kurland CG: **Something for everyone. Horizontal gene transfer in evolution**. *EMBO Rep* 2000, **1**:92-95.

5. Brown JR: **Ancient horizontal gene transfer**. *Nat Rev Genet* 2003, **4**:121-132.

6. • Jain R, Rivera MC, Moore JE, Lake JA: **Horizontal gene transfer accelerates genome innovation and evolution.** *Mol Biol Evol* 2003, in press.

This is the first paper that tries to quantify the factors influencing the success of horizontal gene transfer. Surprisingly, genome size and GC content are major factors.

7. • Snel B, Bork P, Huynen MA: **Genomes in flux: the evolution of archaeal and proteobacterial gene content**. *Genome Res* 2002, **12**:17-25.

The authors and authors of [8•] present a method that estimates the overall rate of horizontal gene transfer, gene loss, gene genesis and vertical inheritance.

8. • 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.

See Snel *et al.* (2002) [7•].

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

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

11. • Welch RA, Burland V, Plunkett G III, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J *et al.*: **Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli***. *Proc Natl Acad Sci U S A* 2002, **99**:17020-17024.

A three-way genome comparison of *E. coli* strains reveals that only 39% of their combined set of proteins are shared.

12. • Nesbo CL, Nelson KE, Doolittle WF: **Suppressive subtractive hybridization detects extensive genomic diversity in *Thermotoga maritima***. *J Bacteriol* 2002, **184**:4475-4488.

Using suppressive subtractive hybridization, the authors corroborate the results of Welch *et al.* (2002) [11•] for free-living organisms.

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

14. Charlebois RL, Beiko RG, Ragan MA: **Microbial phylogenomics: branching out**. *Nature* 2003, **421**:217.

15. • Gogarten JP, Doolittle WF, Lawrence JG: **Prokaryotic evolution in light of gene transfer**. *Mol Biol Evol* 2002, **19**:2226-2238.
The consequence of horizontal gene transfer for prokaryotic evolution is discussed. New and provocative theories are presented.

16. Eisen JA: **Horizontal gene transfer among microbial genomes: new insights from complete genome analysis**. *Curr Opin Genet Dev* 2000, **10**:606-611.

17. Wolf YI, Rogozin IB, Grishin NV, Koonin EV: **Genome trees and the tree of life**. *Trends Genet* 2002, **18**:472-479.

18. de la Cruz F, Davies J: **Horizontal gene transfer and the origin of species: lessons from bacteria**. *Trends Microbiol* 2000, **8**:128-133.

19. Ragan MA: **Detection of lateral gene transfer among microbial genomes**. *Curr Opin Genet Dev* 2001, **11**:620-626.

20. Koonin EV, Makarova KS, Aravind L: **Horizontal gene transfer in prokaryotes: quantification and classification**. *Annu Rev Microbiol* 2001, **55**:709-742.

21. Ragan MA: **On surrogate methods for detecting lateral gene transfer**. *FEMS Microbiol Lett* 2001, **201**:187-191.

22. Lawrence JG, Ochman H: **Reconciling the many faces of lateral gene transfer**. *Trends Microbiol* 2002, **10**:1-4.

23. Guindon S, Perriere G: **Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes**. *Mol Biol Evol* 2001, **18**:1838-1840.

24. • Daubin V, Perriere G: **G+C3 structuring along the genome: a common feature in prokaryotes**. *Mol Biol Evol* 2003, **20**:471-483.
A demonstration that GC content is not homogeneous along the prokaryotic chromosome, probably because of functional constraints, which have yet to be identified. The GC content anomaly does not constitute a reliable indicator of horizontal gene transfers.

25. Salzberg SL, White O, Peterson J, Eisen JA: **Microbial genes in the human genome: lateral transfer or gene loss?** *Science* 2001, **292**:1903-1906.

26. Koski LB, Golding GB: **The closest BLAST hit is often not the nearest neighbor**. *J Mol Evol* 2001, **52**:540-542.

27. Vogl C, Badger J, Kearney P, Li M, Clegg M, Jiang T: **Probabilistic analysis indicates discordant gene trees in chloroplast evolution**. *J Mol Evol* 2003, **56**:330-340.

28. Bailes E, Gao F, Bibollet-Ruche F, Courgnaud V, Peeters M, Marx PA, Hahn BH, Sharp PM: **Hybrid origin of SIV in chimpanzees**. *Science* 2003, **300**:1713.

29. Posada D, Crandall KA, Holmes EC: **Recombination in evolutionary genomics**. *Annu Rev Genet* 2002, **36**:75-97.

30. Archibald JM, Roger AJ: **Gene duplication and gene conversion shape the evolution of archaeal chaperonins**. *J Mol Biol* 2002, **316**:1041-1050.

31. Yap WH, 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* 1999, **181**:5201-5209.

32. van Berkum P, Terefework Z, Paulin L, Suomalainen S, Lindstrom K, Eardly BD: **Discordant phylogenies within the rrn loci of *Rhizobia***. *J Bacteriol* 2003, **185**:2988-2998.

33. Cavalier-Smith T: **Membrane heredity and early chloroplast evolution**. *Trends Plant Sci* 2000, **5**:174-182.

34. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes**. *Nucleic Acids Res* 2001, **29**:22-28.

35. Zivanovic Y, Lopez P, Philippe H, Forterre P: **Pyrococcus genome comparison evidences chromosome shuffling-driven evolution**. *Nucleic Acids Res* 2002, **30**:1902-1910.

36. Lawrence J: **Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes**. *Curr Opin Genet Dev* 1999, **9**:642-648.

37. House CH, Runnegar B, Fitz-Gibbon ST: **Geobiological analysis using whole genome-based tree building applied to the Bacteria, Archaea and Eukarya**. *Geobiology* 2003, **1**:15-26.

38. • Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV: **Genome trees constructed using five different approaches suggest new major bacterial clades**. *BMC Evol Biol* 2001, **1**:8.
Using five largely independent approaches, the authors show the emergence of a recurrent phylogenetic pattern.

39. • Korbel JO, Snel B, Huynen MA, Bork P: **SHOT: a web server for the construction of genome phylogenies**. *Trends Genet* 2002, **18**:158-162.
Same work as Wolf *et al*. (2001) [38•], but using only the gene order and gene content approach.

40. Schwartz RM, Dayhoff MO: **Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts**. *Science* 1978, **199**:395-403.

41. Woese CR, Gibson J, Fox GE: **Do genealogical patterns in purple photosynthetic bacteria reflect interspecific gene transfer?** *Nature* 1980, **283**:212-214.

42. Zhaxybayeva O, Gogarten JP: **Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses**. *BMC Genomics* 2002, **3**:4.

43. •• Daubin V, Gouy M, Perriere G: **A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history**. *Genome Res* 2002, **12**:1080-1090.
The only phylogenetic study at the genome level that used refined phylogenetic tree reconstruction methods (e.g. maximum likelihood). The authors suggest that, unexpectedly, horizontal gene transfers affect similarly operational and informational genes.

44. •• Brochier C, Bapteste E, Moreira D, Philippe H: **Eubacterial phylogeny based on translational apparatus proteins**. *Trends Genet* 2002, **18**:1-5.
A detailed phylogenetic analysis of the protein involved in translation for a rich sample of 45 bacteria. The congruency between the genes is evaluated, suggesting that the phylogeny inferred from concatenation is not a mere average of different history, but really reflects the organismal phylogeny.

45. Raymond J, Zhaxybayeva O, Gogarten JP, Blankenship RE: **Evolution of photosynthetic prokaryotes: a maximum-likelihood mapping approach**. *Philos Trans R Soc Lond B Biol Sci* 2003, **358**:223-230.

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

47. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ: **Universal trees based on large combined protein sequence data sets**. *Nat Genet* 2001, **28**:281-285.

48. Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O: **Towards a reliable objective function for multiple sequence alignments**. *J Mol Biol* 2001, **314**:937-951.

49. Cline M, Hughey R, Karplus K: **Predicting reliable regions in protein sequence alignments**. *Bioinformatics* 2002, **18**:306-314.

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

51. Nesbo CL, Boucher Y, Doolittle WF: **Defining the core of nontransferable prokaryotic genes: the euryarchaeal core**. *J Mol Evol* 2001, **53**:340-350.

52. Philippe H, Douzery E: **The pitfalls of molecular phylogeny based on four species, as illustrated by the Cetacea/Artiodactyla relationships**. *J Mamm Evol* 1994, **2**:133-152.

53. Adachi J, Hasegawa M: **Instability of quartet analyses of molecular sequence data by the maximum likelihood method: the Cetacea/Artiodactyla relationships**. *Mol Phylogenet Evol* 1996, **6**:72-76.

54. Matte-Tailliez O, Brochier C, Forterre P, Philippe H: **Archaeal phylogeny based on ribosomal proteins**. *Mol Biol Evol* 2002, **19**:631-639.

55. Bininda-Emonds ORP, Gittleman JL, Steel MA: **The (super)tree of life: procedures, problems, and prospects**. *Annu Rev Ecol Syst* 2002, **33**:265-289.

56. Ohnishi M, Kurokawa K, Hayashi T: **Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors?** *Trends Microbiol* 2001, **9**:481-485.

57. • Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR *et al.*: **Origins of highly mosaic mycobacteriophage genomes**. *Cell* 2003, **113**:171-182.
Information obtained from sequencing ten mycobacteriophage genomes suggests that phages may represent a largely unexplored and underestimated genetic biodiversity, which could constitute a reservoir of genes for prokaryotes.

58. Addario-Berry L, Hallett M, Lagergren J: **Towards identifying lateral gene transfer events.** *Pac Symp Biocomput* 2003:279-290.

59. Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau MER, Nesbø CL, Case RJ, Doolittle WF: **Lateral gene transfer and the origins of prokaryotic groups.** *Annu Rev Genet* 2003, in press.

60. Daubin V, Moran NA, Ochman H: **Phylogenetics and the cohesion of bacterial genomes**. *Science* 2003, **301**:829-832.