Phylogenetics and the cohesion of bacterial genomes

Vincent Daubin, Nancy Moran, Howard Ochman

Phylogenetics and the Cohesion of Bacterial Genomes

Vincent Daubin,[1] Nancy A. Moran,[2] Howard Ochman[1]*

Gene acquisition is an ongoing process in many bacterial genomes, contributing to adaptation and ecological diversification. Lateral gene transfer is considered the primary explanation for discordance among gene phylogenies and as an obstacle to reconstructing the tree of life. We measured the extent of phylogenetic conflict and alien-gene acquisition within quartets of sequenced genomes. Although comparisons of complete gene inventories indicate appreciable gain and loss of genes, orthologs available for phylogenetic reconstruction are consistent with a single tree.

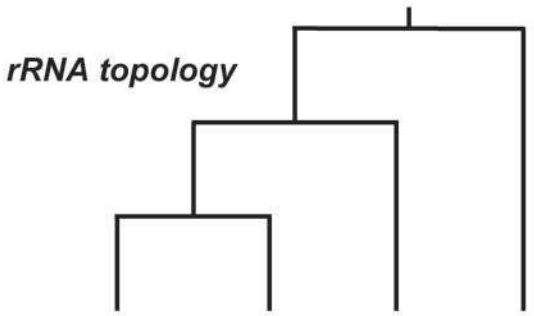
In all but the most reduced bacterial genomes, there is a substantial fraction of genes whose distributions and compositional features indicate that they originated by lateral gene transfer (LGT) (*1*). There is also clear evidence of LGT between distantly related organisms based on phylogenetic studies involving large taxonomic samples (*2*). Given these findings, incompatibility of phylogenies within and among bacterial phyla based on different genes has routinely been ascribed to LGT (*3–10*). However, building molecular phylogenies for distantly related species is often a difficult task, and choice of phylogenetic methods, genes, or taxa can yield different results. For example, there is still no consensus on the monophyly of rodents (*11*, *12*) or the branching order of amniotes (*13*, *14*), and these groups are young compared to bacterial phyla. In addition, distinguishing between orthologous genes (sequences that trace their divergence to the splitting of organismal lineages) and paralogous (duplicated) genes becomes increasingly difficult when considering more distantly related taxa.

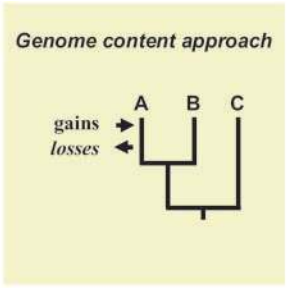
The effects of LGT have been extended from the deepest to the shallowest levels of bacterial relationships. Indeed, the similarities in gene sequence and gene content that define widely accepted bacterial taxa have been proposed to reflect boundaries to gene transfer, rather than vertical transmission and common organismal ancestry (*10*). Thus, LGT may overwhelm attempts to reconstruct the relationships among bacterial taxa. The claim that the history of bacteria might be more faithfully depicted as a net than as a tree (*7*) relies upon the postulate that the substantial incidence of acquired DNA within genomes is the basis for findings of phylogenetic incongruence among genes. However, the genes detected as recently transferred are, by and large, different from those used to build species phylogenies. The former are disproportionately A+T-rich, have restricted phylogenetic distributions, and usually encode accessory functions. In contrast, species phylogenies are based on genes with wide taxonomic distributions and having key roles in cellular processes. However, such differences are often ignored when considering the impact of LGT on bacterial relationships. Although the incidence of recently acquired DNA in bacterial genomes is the most direct indication of extensive LGT among species (*1*), the question of whether the incongruence in gene phylogenies is linked to the amount of new DNA in a genome has not been addressed.

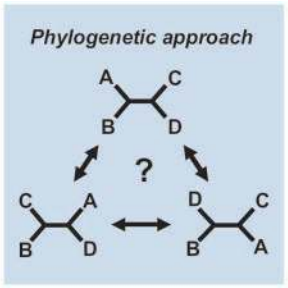
To investigate the relation between DNA acquisition and phylogenetic incongruence, we selected quartets of related, sequenced genomes whose phylogenetic relationships, based on small subunit ribosomal RNA (SSU rRNA) sequences, display the branching topology shown in Fig. 1. For each quartet, we inferred both the number of recently acquired and lost genes (based on their phylogenetic distributions) and the proportion of ortholog phylogenies supporting lateral transfers. We applied a conservative method for identifying orthologs by including only those genes having a single significant match per genome, thus minimizing the risks of including hidden paralogs descending from within-genome duplication events. This contrasts with the commonly used "reciprocal best-hit method" (*15*) to infer orthology, which can yield misleading results (*16*), especially when paralogs experience different evolutionary rates. We retained all quartets of species for which >25% of the genes from the smallest genome were recovered as orthologs. We then tested which of the three possible trees was significantly supported for each ortholog family, using the Shimodaira-Hasegawa (SH) (*17*) test implemented in Tree-puzzle 5.1 (*18*) at the 5% level of significance (*19*). This method tests if an alignment significantly supports a tree by estimating the confidence limits of the likelihood estimates of the topologies.

[1]Department of Biochemistry and Molecular Biophysics, [2]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721, USA.

*To whom correspondence should be addressed. E-mail: hochman@email.arizona.edu







		Genome A	Genome B	Genome C	Genome D	Complete genome content: Genome A gains (losses)	Complete genome content: Genome B gains (losses)	Ortholog phylogenies: rRNA topology	Ortholog phylogenies: LGT topology	Ortholog phylogenies: Non-resolving	Ortholog phylogenies: Total orthologs
Interspecies	Escherichia	*E. coli 933*	*E. coli K12*	*S. typhimurium*	*Y. pestis KIM*	522 *(22)*	144 *(114)*	1589	9	116	1715
	Salmonella	*S. enterica*	*S. typhimurium*	*E. coli K12*	*Y. pestis KIM*	259 *(38)*	203 *(13)*	1603	3	97	1703
	Enterobacteria	*E. coli K12*	*S. typhimurium*	*Y. pestis KIM*	*V. cholerae*	549 *(54)*	601 *(29)*	923	0	222	1145
	Buchnera	*B. aphidicola Ap*	*B. aphidicola Sg*	*B. aphidicola Bp*	*E. coli K12*	9 *(10)*	1 *(11)*	237	0	169	406
	Streptococcus	*S. pyogenes*	*S. agalactiae*	*S. pneumoniae*	*L. lactis*	232 *(56)*	337 *(19)*	93	28	564	686
	Rhizobia	*S. meliloti*	*A. tumefasciens*	*M. loti*	*B. melitentis*	214 *(117)*	519 *(113)*	616	34	360	1010
	Staphylococcus	*S. aureus MW2*	*S. aureus Mu50*	*S. epidermidis*	*B. subtilis*	71 *(4)*	128 *(2)*	737	0	72	809
Intraspecies	***C. pneumoniae***	*C. pneumoniae CWL029*	*C. pneumoniae AR39*	*C. pneumoniae J138*	*C. trachomatis*	nd	nd	60	52	520	632
	E. coli	*E. coli 933*	*E. coli K12*	*E. coli CFT*	*S. typhimurium*	nd	nd	177	149	1888	2215
	S. aureus	*S. aureus N315*	*S. aureus Mu50*	*S. aureus MW2*	*S. epidermidis*	nd	nd	340	6	1225	1571

Fig. 1. Two methods for assessing LGT in bacterial genomes, applied to available quartets of closely related, fully sequenced bacterial taxa. The reference topology, based on SSU rRNA, is shown in the upper left, with taxon names listed in the rows below. The yellow box contains the numbers of gene acquisitions in genomes A and B, as determined by parsimony in comparisons of complete genome contents. The blue box contains the numbers of orthologous genes supporting a topology that conflicts with the reference topology. "Interspecies" and "Intraspecies" comparisons represent quartets of taxa in which phylogenetic incongruence can be explained, respectively, by a transfer from another species or from another strain of the same species. For intraspecies comparisons, numbers of acquired and lost genes were not calculated because of uncertainty about the actual tree topology (nd, not determined). (*B. aphidicola* strains are entirely isolated in different hosts and were thus considered as different species despite having a single name. In *B. aphidicola*, amounts of gene loss and gene gain are similar, suggesting that LGT is overestimated due to independent losses of genes.)

For each quartet of species, the sets of orthologs were divided into three categories: (i) those supporting the reference ribosomal RNA tree (termed "rRNA topology" in Fig. 2), (ii) those supporting one of the two alternative topologies ("LGT topology"), and (iii) those not significantly supporting any topology ("Nonresolving"). Note that all transfers to lineages connecting A and B from outside of the ABC clade, or from the C lineage, result in an "LGT topology." For all quartets, the proportion of ortholog phylogenies supporting a hypothesis of LGT is always small, and often zero (Fig. 2). In some of the interspecies comparisons, a few alignments support alternative topologies. In some cases, we identified groups of two or more adjacent genes that support the same LGT topology, consistent with the transfer of an operon (e.g., an operon encoding three subunits of a glutamyl–tRNA amidotransferase in the streptococci). Although some of the alternate topologies probably reflect LGT, the frequencies of such topologies are strongly correlated with the ratio of external and internal branch lengths ($r^2 = 0.928$; $P < 0.0005$), but not with the distance between sequences in the rRNA tree ($P > 0.3$). This suggests that most cases of alternate topologies represent false-positives due to reconstruction artifacts rather than the accumulation of LGT events with time.

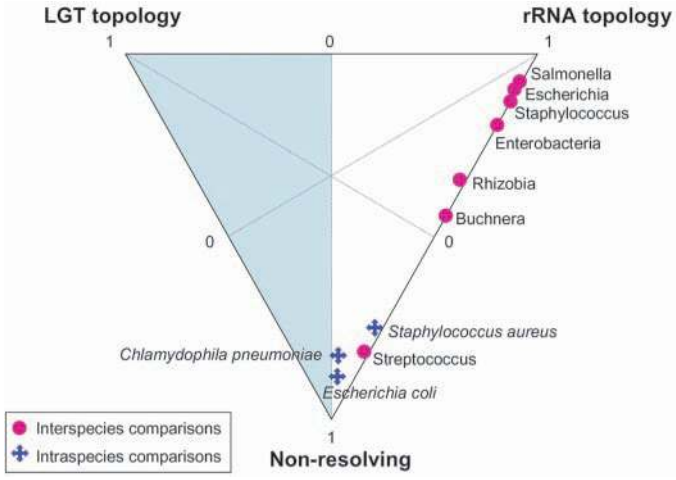


Fig. 2. Relative frequencies of the three categories of alignments, i.e., those supporting the reference phylogeny (SSU rRNA), those supporting an alternate phylogeny (LGT), and those with no statistical support for any phylogeny. Points represent quartets of genomes for which orthologous genes have been inferred, aligned, and evaluated at the nucleic acid sequences level based on the SH test implemented in Puzzle 5.1 (*19*). The left part of the plot (in blue) represents the area where LGT predominates.

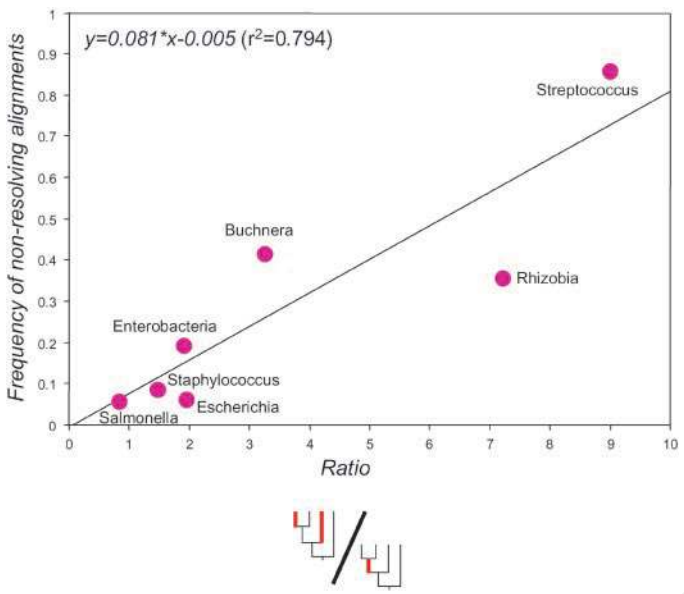


Fig. 3. Relation between relative length of the internal and terminal branches in the SSU rRNA tree and the frequency of nonresolving alignments for the interspecies comparisons considered in Fig. 1.

The only quartets indicating substantial levels of LGT are the intraspecies comparisons of strains of *Escherichia coli* and of *Chlamydophila pneumoniae*. This result indicates genome-wide incidence of homologous recombination, extending previous studies based on individual genes (*20*, *21*), and supports the proposal that the integrity and definition of some bacterial species can be inferred from the incongruency of gene trees (*22*).

Thus, orthologous genes comprise mainly two groups: those supporting the rRNA topology and those containing insufficient phylogenetic signal. The intraspecies comparisons have a higher proportion of nonresolving alignments (Fig. 2), which may result from both intragenic recombination and low levels of divergence. In contrast, the interspecies comparisons display wide variation in the proportion of nonresolving alignments. This lack of resolution reflects saturation rather than recombination, as supported by the increasing proportion of nonresolving alignments when terminal branches are long relative to internal branches in the rRNA tree (Fig. 3).

Although these results are limited to relatively closely related species, they highlight methodological artifacts that may underlie some previous claims that LGTs between distantly related species are sufficiently frequent that the history of bacterial lineages cannot be depicted by traditional phylogenies (*4*, *5*): First, previous studies adopted a likelihood-mapping method that examined differences in the likelihood of alternate phylogenies without considering their confidence intervals, thereby inflating the perceived incidence of LGT. Applying such an approach to the sets of orthologs identified in the present study yielded similar results for the interspecies comparisons involving closely related species (e.g., *Escherichia*, *Salmonella*) but led to a three- and fivefold increase in estimated levels of LGT in the *Rhizobia* and in the *Streptococci*, respectively, under the most stringent threshold ($P > 0.99$). Hence, likelihood mapping does not provide an accurate measure of LGT when applied to distantly related species. Second, although arguably there would be an increased incidence of LGT between very distantly related taxa owing to the time elapsed since their divergence, deep phylogenies such as these are the most likely to contain short internal branches. Thus, our results suggest that LGT among prokaryotic phyla can be overestimated, even when applying a conservative statistical test. More generally, phylogenetic studies based on small numbers of distantly related taxa may yield spurious results (*23*, *24*). As shown previously (*4*), the addition of species to well-supported quartet phylogenies will often decrease statistical support for LGT. Finally, even when full genome sequences are available, the use of a poor criterion for orthology, such as the reciprocal best-hit method, might lessen the ability to eliminate paralogous genes in distant species and could lead to false conclusions of LGT. Hence, evidence for LGT should be considered cautiously when relying on poor taxonomic sampling.

Based on a parsimony comparison of complete genomes from closely related organisms, we have also estimated the number of gene acquisitions and gene losses in each lineage. Using a BLASTP query (*25*), we identified recently acquired genes as those with no significant matches in the sister group and the closest outgroup (*26*). Levels of gene acquisition are often high, sometimes exceeding 10% of the genome (Fig. 1), providing independent support of previous estimates based on compositional features of individual genomes (*1*). The introduced genes identified by this method are principally acquired from outside of the group of species considered. If comparable levels of transfer were affecting the sets of orthologs that we identified, topological incongruence would be more frequent and would be expected to correlate with the proportion of acquired genes in a genome. For example, *Escherichia* and *Salmonella* show no evidence of interspecies transfer affecting orthologous gene families despite the high incidence of alien genes (Fig. 1).

Thus, not all genes are equally subject to LGT. Based on previous analyses of prokaryotic phyla, it was suggested that some functional gene classes, such as those involved in complex molecular interactions (termed "informational genes"), are less prone to LGT (*3*, *8*). Our results provide evidence for a more pronounced dichotomy of genes, consisting of "acquired" and "ortholog" classes. These classes differ in fundamental features: More than 70% of acquired genes encode proteins of uncharacterized functions, whereas ~80% of genes in the ortholog class possess functional annotations (of which only 10 to 15% would be classified as "informational" genes). Because little evidence of LGT is found for orthologous genes, we conclude that LGT is concentrated in a class of genes that are not candidates for phylogenetic analysis.

The discovery of large amounts of alien genes in bacteria has led to the prediction that LGT causes phylogenetic disruption in bacteria. Whereas invoking LGT is a last resort for explaining an observation of phylogenetic incongruence in animals or plants, LGT is often the first choice to explain incongruence in prokaryotes, often without regard to the significance of the support accorded to different trees (*6*). LGT is an important driving force in prokaryotic evolution and adaptation (*1*, *9*, *27*), but it is not always the explanation for instances of phylogenetic conflict. Regardless, if used critically, sequence data offer great promise for reconstructing the evolutionary relationships of bacterial lineages.

References and Notes

1. H. Ochman, J. G. Lawrence, E. A. Groisman, *Nature* **405**, 299 (2000).
2. J. R. Brown, W. F. Doolittle, *Microbiol. Mol. Biol. Rev.* **61**, 456 (1997).
3. R. Jain, M. C. Rivera, J. A. Lake, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3801 (1999).
4. C. L. Nesbo, Y. Boucher, W. F. Doolittle, *J. Mol. Evol.* **53**, 340 (2001).
5. O. Zhaxybayeva, J. P. Gogarten, *BMC Genomics* **3**, 4 (2002).
6. J. Raymond, O. Zhaxybayeva, J. P. Gogarten, S. Y. Gerdes, R. E. Blankenship, *Science* **298**, 1616 (2002).
7. W. F. Doolittle, *Science* **284**, 2124 (1999).

8. M. C. Rivera, R. Jain, J. E. Moore, J. A. Lake, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6239 (1998).
9. W. F. Doolittle, *Trends Cell Biol.* **9**, M5 (1999).
10. J. P. Gogarten, W. F. Doolittle, J. G. Lawrence, *Mol. Biol. Evol.* **19**, 2226 (2002).
11. D. Graur, W. A. Hide, W.-H. Li, *Nature* **351**, 649 (1991).
12. D. Huchon *et al.*, *Mol. Biol. Evol.* **19**, 1053 (2002).
13. S. B. Hedges, L. L. Poling, *Science* **283**, 998 (1999).
14. S. Hughes, D. Mouchiroud, *J. Mol. Evol.* **53**, 70 (2001).
15. A. R. Mushegian, E. V. Koonin, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10268 (1996).
16. L. B. Koski, G. B. Golding, *J. Mol. Evol.* **52**, 540 (2001).
17. H. Shimodaira, M. Hasegawa, *Mol. Biol. Evol.* **16**, 1114 (1999).
18. K. Strimmer, A. von Haeseler, *Mol. Biol. Evol.* **13**, 964 (1996).
19. For each set of orthologs, both nucleotide and amino acid alignments were evaluated. Both methods yielded very similar results, and only the results based on nucleotide sequences are presented. Specifically, the protein alignments did not yield significantly smaller numbers of nonresolving or LGT topologies. The more permissive (*17*) Kishino-Hasegawa test, either two-sided (*28*) or one-sided (*29*), produced very similar results.
20. K. L. Millman, S. Tavare, D. Dean, *J. Bacteriol.* **183**, 5997 (2001).
21. D. S. Guttman, D. E. Dykhuizen, *Science* **266**, 1380 (1994).
22. D. E. Dykhuizen, L. Green, *J. Bacteriol.* **173**, 7257 (1991).
23. H. Philippe, E. Douzery, *J. Mamm. Evol.* **2**, 133 (1994).
24. J. Adachi, M. Hasegawa, *Mol. Phylogenet. Evol.* **6**, 72 (1996).
25. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
26. A protein was considered absent from a genome when it had no match >10% of the bit score of the protein against itself. A protein was considered present in a genome when it had a match >50% of the bit score of the protein against itself, based on the BLOSUM62 matrix. Losses estimate the probability that a gene present in the common ancestor of the three taxa (ABC) was lost in one. Because gene gains greatly outnumber gene losses in all quartets (except *B. aphidicola*), the probability of overestimating gene acquisition, due to independent loss by two taxa, is very low.
27. J. G. Lawrence, H. Ochman, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9413 (1998).
28. H. Kishino, M. Hasegawa, *J. Mol. Evol.* **29**, 170 (1989).
29. N. Goldman, J. P. Anderson, A. G. Rodrigo, *Syst. Biol.* **49**, 652 (2000).

30. We thank S. Santos, E. Lerat, and two anonymous reviewers for comments. This research was supported by U.S. Department of Energy grant DEFG0301ER63147 (H.O.).