Lateral Gene Transfer, Genome Surveys, and the Phylogeny of Prokaryotes

Doolittle argued that to construct phylogenies, an organism should be regarded as either less or more than the sum of its genes (*1*). The argument is based on the observation that gene phylogenies are rarely consistent with one another because, among others, of lateral gene transfer (LGT). Creating phylogenies from sequence data in which an organism is described exactly as the sum of its genes is not, however, the only approach (*2*). Rather than creating phylogenies based on sequence identity for separate genes, this alternative creates a distance-based phylogeny at the genome level by comparing the fraction of genes shared between genomes (*2*). The resulting phylogeny (Fig. 1) of completely sequenced genomes (for an overview, see http://www.tigr.org/tdb/tdb.html) is remarkably similar to the phylogenies that are based on 16*S* ribosomal RNA (*3*). Not only is the trichotomy between Eukarya, Bacteria, and Archaea present, but within each of these taxa the clusters generally recognized as being monophyletic and for which multiple genomes are available, all have high bootstrap values (the Proteobacteria and their branching order, the Spirochaetales, the low (G+C) Gram-positive bacteria, and the Euryarchaeota). This method does not resolve the major branchings of the Bacteria, but neither do the phylogenies based on sequences that do not show LGT resolve this part of Bacterial phylogeny with a high degree of confidence.

LGT of genes that are not yet present in a genome, and the parallel loss of orthologous genes in distant phylogenetic branches, reduce the phylogenetic pattern in the gene content. We argue that the rate of these processes is not so high as to preclude a phylogenetic view of genome evolution. Genome phylogeny based on gene content disregards the evolutionary history of genes. It is analogous to distance-based phylogenies of sequences that disregard the origin of amino acids. In the absence of a model of sequence or genome evolution, such approaches have been shown to be very useful. In discussions about genome phylogeny and gene phylogenies, it is difficult to see the forest (genome phylogeny) for the trees (gene phylogenies) (*4*). Higher order approaches that are complementary to gene phylogenies and that stress the complete genome aspect and the relations between the genes should be taken into consideration (*5*).

Martijn Huynen
Berend Snel
Peer Bork
European Molecular Biology Laboratory
Meyerhofstrasse 1
69117 Heidelberg, Germany
Max-Delbrück-Centrum
for Molecular Medicine
13122 Berlin Buch
Germany
E-mail: huynen@embl-heidelberg.de

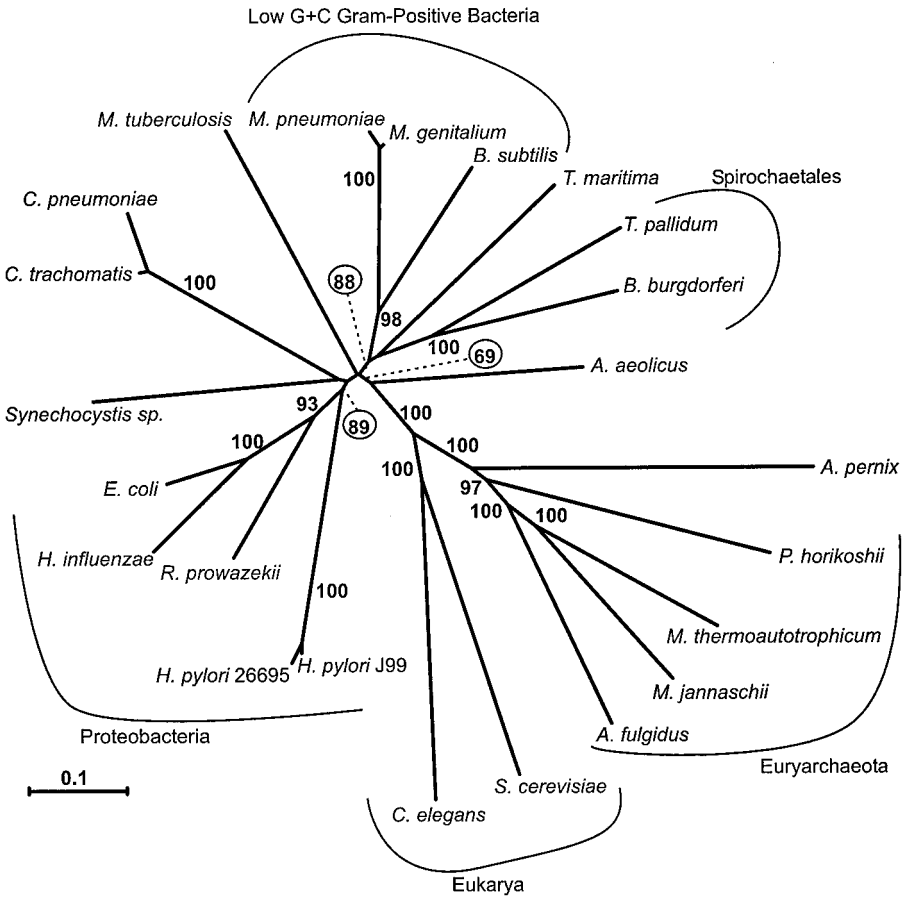

Fig. 1. Genome phylogeny based on gene content. A Fitch-Margoliash (*6*) tree was made from a genome distance matrix. Distances were calculated based on the number of genes shared between two genomes divided by the number of genes in the smallest genome. The number of shared genes between two genomes is calculated using an operational definition of orthology. Two genes from two genomes are considered orthologous when they have the highest significant level of pairwise similarity to each other compared to their similarity to the other genes in each other's genome. Two genes can be orthologous to a single gene from another genome when their alignments do not overlap [see (*2*)].

References
1. W. F. Doolittle, *Science* **284**, 2124 (1999).
2. B. Snel, P. Bork, M. A. Huynen, *Nature Genet.* **21**, 108 (1999).
3. G. J. Olsen, C. R. Woese, R. Overbeek, *J. Bacteriol.* **176**, 1 (1994).
4. E. Pennisi, *Science* **284**, 1305 (1999).
5. M. A. Huynen and P. Bork, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 5849 (1998).
6. W. M. Fitch and E. Margoliash, *Science* **155**, 279 (1967).

12 July 1999; accepted 9 September 1999

From the time that ribosomal RNA molecules were first observed and described (*1*) until the 1970s, they were the only high molecular weight cellular RNAs that could readily be isolated. Historically, the great abundance of these RNAs and the repetitiveness of their genes, particularly in eukaryotes, made rDNA a popular choice as the molecular yardstick for determining evolutionary relationships between diverse organisms. While the essential and unvarying role of rRNA in protein synthesis has become clear, it remains to be seen whether the sequences of rRNA molecules should form the basis for a "universal tree of life" (*2*).

W. F. Doolittle (*3*) highlighted that lack of clarity in his review of the universal tree and emphasized the problem posed by what appears to be rampant lateral gene transfer (LGT), particularly among and between the Archaea and Eubacteria. In so doing,

Doolittle echoed a growing chorus that has interpreted conflicts between the rDNA tree and those inferred from other molecular sequences, as evidence that genes have transferred among organisms quite freely throughout the course of cellular evolution.

The case for widespread LGT during microbial evolution appears to be supported by the observations of Lawrence and Ochman (*4*) that segmental inhomogeneities exist in base composition and codon usage within the *E. coli* genome. Yet, although these authors suggested that atypical segments are relics from past LGT events, other explanations are possible. For example, similar regions of exceptional DNA composition in other microbial genomes have been attributed to the preference for genes to be transcribed in parallel with leading strand replication, which can result in a compositional bias between DNA strands (*5*). In genomes so organized, the consequences of chromosomal inversions could mistakenly be attributed to LGT; it is certainly plausible that comparable mechanistic explanations will yet surface for the variant sequences in *E. coli.*

However, regardless of the source of these sequence blocks in *E. coli*, such data from specialized enteric bacteria are not a sufficient basis for concluding that pervasive LGT has shaped the broader course of cellular evolution. That generalization rests squarely on conflicts between the rDNA universal tree and comparative analyses of other gene sequences, or on less phylogenetically rigorous heuristic comparisons of total genome content using search algorithms that look for similar sequences. We and others (*6*) have demonstrated that deep branches in rDNA phylogenies suffer from long-branch attraction and other sources of artifactual behavior. Much of the ancient topology of the universal tree seems, therefore, unreliable. Yet rDNA has been subjected to the most extensive analyses of all genes used to examine ancient evolutionary relationships. What, then, is to be made of conflicts that arise from the less thoroughly investigated gene sequences?

There are many well-documented cases of lateral gene transfer. There are also many demonstrated examples of erroneous molecular phylogeny. It may turn out that LGT was, and is, as rampant as Doolittle and others have suggested. However, it remains unclear whether this conclusion should be drawn from initial examinations of the vast number of gene sequences that are pouring in from diverse organisms. As an alternative to sequence-based phylogenies, Gutpa (*7*) used the presence or absence of signature sequences (clearly homologous insertions and deletions in many different genes) to derive the tree of life. Although somewhat different from the rDNA model, the evolutionary pattern that emerges from his investigations is, remarkably consistent internally and shows little evidence of LGT. Martin (*8*) suggested that acceptance of the existence of vast amounts of LGT may lead to an understanding of "the principles which must govern the distribution of genes across bacterial genomes." But if methodological problems are the dominant cause of the apparent conflicts between gene histories, misattributing them to LGT will obscure those governing principles. It would seem prudent to develop alternative methodologies and to explore other possible hypotheses before settling on LGT.

There were two problems implicit in the rDNA universal tree when it first took shape: one was an untested hypothesis—the specific evolutionary topology it proposed for the tree of life; the other was an assumption of the insufficiency of the methods and data used to reliably infer the nature of ancient evolutionary events. While conflicts arising from other molecular sequence data have caused Doolittle and others to question the topology and even the concept of the universal tree, the possibility that the cause of the conflicts may lie largely with the methods used has generally been overlooked. This leaves lateral gene transfer as the most obvious explanation for the perceived discrepancies between gene histories. A precipitous acceptance of such widespread LGT places evolutionary biologists in the untenable position of adopting an unfalsifiable hypothesis, at least in terms of the techniques of comparative sequence analyses that currently dominate the field of molecular evolution. Any phylogenetic pattern inferred from any given gene can be fit to some suitable mix of conventional intraspecies gene transmission and interorganismal genetic promiscuity. Thus, unless more reliable evidence is uncovered, the scientific method requires that we invoke the idea of ubiquitous LGT only as a last resort.

John W. Stiller
Benjamin D. Hall
Department of Genetics
University of Washington
P. O. Box 357360
Seattle, WA 98195, USA

References

1. B. D. Hall and P. Doty, *J. Mol. Biol.* **1**, 111 (1959); C. G. Kurland, *J. Mol. Biol.* **2**, 83 (1960).
2. C. R. Woese, *Microbiol Rev.* **51**, 221 (1987); C. R. Woese, O. Kandler, M. L. Wheelis, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 4576 (1990).
3. W. F. Doolittle, *Science* **284**, 2124 (1999).
4. L. G. Lawrence and H. Ochman, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9413 (1998).
5. J. R. Lobry, *Mol. Biol. Evol.* **13**, 660 (1996); J. O. McInerney, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 10698 (1998).
6. M. Hasegawa and T. Hashimoto, *Nature* **326**, 411 (1993); H. Phillipe and A. Adoutte, in *Evolutionary Relationships Among Protozoa*, G. H. Coombs, K. Vickerman, M. A. Sleigh, A. Warren, Eds. (Chapman & Hall, London, 1998), pp. 25–56; J. W. Stiller and B. D. Hall, *Mol. Biol. Evol.*, in press.
7. R. S. Gupta, *Microbiol. Mol. Biol. Rev.* **62**, 1435 (1998).
8. W. Martin, *Bioessays* **21**, 99 (1999).

29 July 1999; accepted 9 September 1999

Doolittle (*1*) argued that the lateral gene transfer (LGT) problem is so acute for prokaryotes that, except for the distinction between *Archaea* (or archaebacteria) and *Bacteria* (or eubacteria), it precludes determination of the evolutionary relationships among various prokaryotic taxa. To infer that the presence of a particular gene in a given organism is due to LGT, one must first assume a model for the relationship among organisms. Since Woese *et al.*'s (*2*) three-domain proposal is the current paradigm, any gene phylogeny inconsistent with this model would be attributed to LGT. Although the basic premise that Archaea are totally distinct from Bacteria is supported by several important characteristics (*1, 2*), there now exists compelling evidence supporting an alternative relationship among prokaryotes (*3*). Emerging from extensive analyses of conserved insertions and deletions (signature sequences) in many highly conserved proteins, this alternate view points to a specific relationship of archaebacteria to the Gram-positive bacteria, both of which possess similar cell structure (bounded by a single membrane) and are distinct from the Gram-negative bacteria (bounded by two different membranes). The characteristics that distinguish archaebacteria from eubacteria (genes involved in information transfer, cell wall and membrane compositions) are primarily those that are the main targets of antibiotics produced by Gram-positive bacteria, and these could have evolved in Gram-positive bacteria in response to strong antibiotic selection pressure (*3*). Once such a possibility is recognized, many observations that are currently attributed to LGT may be interpreted differently. With the use of signature sequence analyses, it has proven possible to define the main taxa among eubacteria and to deduce their orders of evolution from a common ancestor, indicated as follows: Archaebacteria $\Leftrightarrow$ Gram-positive bacteria $\Rightarrow$ *Deinococcus-Thermus* $\Rightarrow$ Green nonsulfur bacteria $\Rightarrow$ Cyanobacteria $\Rightarrow$ *Spirochetes* $\Rightarrow$ *Chlamydia-Cytophaga*-Green sulfur bactria $\Rightarrow$ *Proteobacteria*-ε, δ $\Rightarrow$ *Proteobacteria*-α, β, γ (*3*). This deduced relationship among prokaryotes is consistent with all available data including cell morphology and molecular phylogenies, and it generally supports the inferences based on rRNA phylogenies (*3*). Assuming this alternative model, LGT did not pose a serious problem in the interpretation of data for the different gene sequences (*3*). The implication is that LGT, although widespread, is not so pervasive that it precludes the determination of the evolutionary relationships among organisms. The model based on signature sequences provides a contrasting but reasonable alternative for interpreting the genome data as they accumulate.

Radhey S. Gupta
Bohdan J. Soltys
Department of Biochemistry

McMaster University
Hamilton, Canada L8N 3Z5
E-mail: gupta@mcmaster.ca

References
1. W. F. Doolittle, *Science* **284**, 2124 (1999).
2. C. R. Woese, O. Kandler, M. L. Wheelis, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 4576 (1990).
3. R. S. Gupta, *Microbiol. Mol. Biol. Rev.* **62**, 1435 (1998); R. S. Gupta, T. Mukhtar, B. Singh, *Mol. Microbiol.* **32**, 893 (1999).

16 July 1999; accepted 9 September 1999

Response: The comments of Huynen, Snel, and Bork (*1*); Stiller and Hall (*2*); and Gupta and Soltys (*3*) address many of the issues that make the lateral gene transfer (LGT) question interesting on so many levels.

Huynen *et al*. (*1*) suggest that even in the face of extensive LGT we might still reconstruct phylogeny by considering total gene content—the more genes any two genomes share, the closer their relationship, and thus the closer the relationship of the organisms that replicate them. Even though analyzing genome data in this way produces a tree-like branching diagram not unlike the "standard model" based on rRNA, the approach seems misconceived. If there were only one true phylogenetic tree (a single genealogical signal for all genes), disagreement among the data would reflect only statistical or methodological noise and it would be sensible to use such an aggregated measure. However, we already know that genes have different histories and that there are, indeed, several conflicting genealogical signals in a genome. Hence, it seems inappropriate to equate the history of a whole genome with the history of only a portion—even the majority—of its genes.

Gene content comparisons will be very useful for taxonomy: they give us the best possible measure of "overall (phenetic) similarity" between organisms. But the tree-like branching pattern which can be derived from such similarity measurements is not an organismal (or genome) phylogeny. The mere fact that we can derive a unique "tree" does not mean that a bifurcating branching process (of that same or any topology) gave rise to contemporary organisms or their genomes. This would be analogous to the construction of a tree-like pattern relating Halifax, Toronto, Montreal, and Vancouver on the basis of the percentage of shared surnames in their telephone directories. This pattern would tell us something useful about similarities between their populations, but would it be a "phylogeny" of the cities?

Ironically, LGTs will contribute substantially to the apparent phylogenetic signal in such a gene content tree. There is no signal from the genes to which all genomes have a copy, and this includes the transcription and translation-related genes on which the universal tree is thought to rest most securely. Close relationships, like that between the two *Helicobacter pylori* strains in Huynen *et al.* (*1*, figure 1) must be based on the sharing of genes not found in sister taxa (even *E. coli*, which, overall, has almost three times as many open reading frames). Some of the *Helicobacter*-specific genes will indeed be new sequences, created by duplications since the *Helicobacter/Escherichia* divergence, but many others will have orthologs scattered among other genomes further away than *E. coli*'s, as expected if LGT had occurred "at the base" of the *Helicobacters.*

Stiller and Hall (*2*) fear that the current LGT bandwagon has become a juggernaut. Indeed, many discordant trees are just bad trees—not evidence of LGT. Stiller and Hall's alternative explanation for Lawrence and Ochman's (*4*) results deserves more rigorous testing (or rebuttal by Lawrence and Ochman). Their concerns for results based on BLAST scores or other simple heuristic comparisons are well taken. However, the strongest evidence for the importance of LGT among prokaryotes is not within-genome inhomogeneities or unexpected BLAST scores, but the varying gene content of the collection of genomes now available, the very same sort of data addressed by Huynen *et al*. (*1*). Each new prokaryotic genome that appears contains dozens, if not hundreds, of genes not found in the genomes of its nearest sequenced relatives but found elsewhere among Bacteria or Archaea. Sometimes we can attribute this to loss, from near relatives, of genes present in a common ancestor, but each time we do we add to the burden of genes that must have been carried by the last common ancestor of all prokaryotes. This burden seems unreasonably large, although a rigorous analysis is needed. The patchy distribution of genes and gene clusters in the prokaryotic world (*5, 6*) almost certainly reveals an underlying process of recurrent loss and gain.

Stiller and Hall, rightly in my view, endorse insertions and deletions as the sort of shared-derived characters on which gene genealogies might be more solidly built. Gupta (*7*) presented phylogenies based on such signatures. Gupta and Soltys (*3*) claim in their comment that with these phylogenies, "LGT did not pose a serious problem in the interpretation of data for the different gene sequences." LGT is not a problem for them because they freely embrace it. Explaining why many gene trees show archaea intermingled with Gram-positive bacteria, Gupta (*7*) notes that "it is necessary to propose that some lateral or horizontal gene transfers have taken place." He suggests that either

> genes for many of the proteins for which archaebacteria show a polyphyletic branching within the gram-positive bacteria . . . have been transferred from low G+C gram-positive bacteria to methanogens and thermoacidiphilic archaebacteria and from high G+C gram-positive bacteria to the halophiles, . . . [or] that genes for many functions that indicate a monophyletic nature of archaebacteria were transferred from one or more gram-positive bacteria that originally evolved such changes into others.

Among these "genes for many functions" are included many or most of those involved in transcription, translation, and probably replication, tRNA modification, lipid metabolism, and much else. This is, to my knowledge, the most comprehensive and radical gene transfer event or episode ever proposed.

As long as prokaryotic populations split into new "species" at a high frequency compared to the rate at which LGT supplements or replaces the genes in their genomes, we expect tree-like behavior. Living species and more inclusive taxa will show some consistency in the patterns of similarity and difference shown by most of their genes. But it is possible, and interesting to contemplate (although not proven), that LGT may be frequent enough and promiscuous enough (in terms of genes transferred and phylogenetic distances bridged) that few patterns will exhibit such consistency over 3.5 billion to 4 billion years. Could physical environment and ecology be as strong a determinant of a genome's composition as phylogeny? Bacteria and Archaea seem, in spite of many obvious LGTs, to retain domain-specific gene complements, but all sequenced archaeal genomes are from hyperthermophiles. It is intriguing that recently sequenced hyper-thermophilic bacterial genomes show substantially more archaeal genes than the many mesophilic bacterial genomes sequenced earlier (*8, 9*). Might we expect that when a mesophilic crenarachaeal genome sequence appears, it will prove to be a treasure trove of mesophilic bacterial genes?

W. Ford Doolittle
Department of Biochemistry and Molecular Biology
Dalhousie University
Halifax, Nova Scotia B3H 4H7

References
1. Technical comment by M. Huynen *et al*. on W. F. Doolittle, *Science* **284**, 2124 (1999).
2. Technical comment by J. W. Stiller and B. D. Hall on W. F. Doolittle, *Science* **284**, 2124 (1999).
3. Technical comment by R. S. Gupta and B. J. Soltys on W. F. Doolittle, *Science* **284**, 2124 (1999).
4. J. G. Lawrence and H. Ochman, *Proc. Natl. Acad. Sci.* **95**, 9413 (1998).
5. R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **278**, 631 (1997).
6. T. Gaasterland and M. A. Ragan, *Micr. Comp. Genomics* **3**, 199 (1998).
7. R. S. Gupta, *Microbiol. Mol. Biol. Rev.* **62**, 1435 (1998).
8. G. Deckert *et al*., *Nature* **392**, 353 (1998).
9. K. E. Nelson *et al*., *Nature* **399**, 323 (1999).

21 August 1999; accepted 28 September 1999