RESEARCH ARTICLE

Lateral Transfer of Genes and Gene Fragments in Prokaryotes

Cheong Xin Chan,[1] Robert G. Beiko,*† Aaron E. Darling,*[2] and Mark A. Ragan**

*Institute for Molecular Bioscience and ARC Centre of Excellence in Bioinformatics, The University of Queensland, Brisbane, Queensland, Australia; †Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

Lateral genetic transfer (LGT) involves the movement of genetic material from one lineage into another and its subsequent incorporation into the new host genome via genetic recombination. Studies in individual taxa have indicated lateral origins for stretches of DNA of greatly varying length, from a few nucleotides to chromosome size. Here we analyze 1,462 sets of single-copy, putatively orthologous genes from 144 fully sequenced prokaryote genomes, asking to what extent complete genes and fragments of genes have been transferred and recombined in LGT. Using a rigorous phylogenetic approach, we find evidence for LGT in at least 476 (32.6%) of these 1,462 gene sets: 286 (19.6%) clearly show one or more "observable recombination breakpoints" within the boundaries of the open reading frame, while a further 190 (13.0%) yield trees that are topologically incongruent with the reference tree but do not contain a recombination breakpoint within the open reading frame. We refer to these gene sets as observable recombination breakpoint positive (ORB$^+$) and negative (ORB$^-$) respectively. The latter are prima facie instances of lateral transfer of an entire gene or beyond. We observe little functional bias between ORB$^+$ and ORB$^-$ gene sets, but find that incorporation of entire genes is potentially more frequent in pathogens than in nonpathogens. As ORB$^+$ gene sets are about 50% more common than ORB$^-$ sets in our data, the transfer of gene fragments has been relatively frequent, and the frequency of LGT may have been systematically underestimated in phylogenetic studies.

Introduction

Coherent transmission of genetic material from parent to offspring defines a genomic or organismal lineage. In morphologically complex eukaryotes, almost all genetic transmission follows a vertical (parent-to-offspring) pattern, but many prokaryotes can acquire genetic material potentially originating from outside a lineage and integrate it into the host genome through recombination. Although open issues remain concerning the quantitative extent and physiological consequences of lateral genetic transfer (LGT; also known as horizontal genetic transfer), numerous studies now show that LGT can be quantitatively important and has contributed in many instances to genomic and physiological innovation (Woese 2000; Falkowski et al. 2008; Fournier and Gogarten 2008; Ragan and Beiko 2009). For example, genes encoding antibiotic resistance can be readily acquired and spread within populations in highly selective environments (Grundmann et al. 2006; Barlow 2009). If LGT has been frequent, many genes or genomic regions will exhibit incoherent phylogenetic histories, perhaps undermining the very concept of a genomic lineage (Doolittle 1999; Gogarten et al. 2002; Wolf et al. 2002; Gogarten and Townsend 2005; Ciccarelli et al. 2006; Koonin 2009).

Various approaches have been taken to identify regions of lateral origin within genomes and to quantify the frequency of LGT. A particularly powerful approach has involved inferring the phylogenetic tree of each orthologous gene or protein family and comparing this tree (or its individual topological features, e.g., internal edges) against an accepted external reference topology such as a species tree (Kunin and Ouzounis 2003; Creevey et al. 2004; Beiko et al. 2005b; Lerat et al. 2005; Zhaxybayeva et al. 2006; Shi and Falkowski 2008; Nesbø et al. 2009). Other approaches to quantify the extent of LGT are based on nucleotide composition or codon usage patterns (Mrázek and Karlin 1999; Nakamura et al. 2004; Kechris et al. 2006), inferred gene gain and loss events (Snel et al. 2002; Kunin and Ouzounis 2003; Hao and Golding 2006; Iwasaki and Takagi 2009), and assumptions about ancestral genome sizes (Dagan and Martin 2007).

During LGT, exogenous genetic material is first introduced into the recipient cell and then integrated into the new host via recombination. The integrated genetic material can constitute stretches of noncoding DNA, fragments of genes (Bork and Doolittle 1992; Inagaki et al. 2006), entire genes (Hartl et al. 1992), multiple (entire or fragmentary) adjacent genes (Igarashi et al. 2001; Omelchenko et al. 2003), operons, transposable chromosomal elements, plasmids and other naturally occurring extrachromosomal elements, and pathogenicity islands. Lateral origins have been suggested for stretches of DNA ranging from seven nucleotides (Denamur et al. 2000) to more than 3 Mbp in length (Lin et al. 2008). Evidence for megabase-scale LGT incorporated into the chromosome as hundreds of smaller fragments has also been reported (Didelot et al. 2007). Nonetheless, phylogenetic studies of LGT persist in considering an intact gene (or protein) as the unit of analysis. Only very recently has the frequency of within-gene lateral transfer and recombination in prokaryotes been investigated at multigenome scale (Chan et al. 2009). Recombination of gene fragments (in addition to recombination of whole genes) within the context of LGT among distantly related taxa has not been rigorously studied.

Here we report the first systematic study of recombination of both genes and gene fragments across a broad diversity of sequenced prokaryote genomes, within the

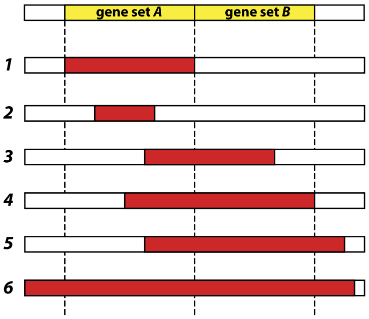
[1] Present address: Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, NJ 08901.

[2] Present address: Genome Center, University of California Davis, Davis, CA 95616.

Key words: lateral genetic transfer, horizontal genetic transfer, genetic recombination, genome evolution, comparative genomics.

E-mail: m.ragan@uq.edu.au.

Genome. Biol. Evol. Vol. 2009:429–438.
doi:10.1093/gbe/evp044
Advance Access publication November 4, 2009



	gene set *A*	gene set *B*
1	ORB⁻	-
2	ORB⁺	-
3	ORB⁺	ORB⁺
4	ORB⁺	ORB⁻
5	ORB⁺	ORB⁻
6	ORB⁻	ORB⁻

FIG. 1.—Definition of ORB$^+$ and ORB$^-$ gene sets. This simplified example shows six cases of LGT, each involving an orthologous region of six genomes (the white rectangles); adjacent genes *A* and *B* lie fully within this region in each genome, with gene boundaries as shown by the vertical dashed lines. Each case 1–6 corresponds to the presence of a single genetically recombined region (colored in red); thus, for the LGT event depicted in case 4, a recombination breakpoint is detected in gene set A (but not in gene set B), and the tree inferred for gene set B is topologically incongruent with the reference tree. We therefore label gene set A as ORB$^+$ and gene set B as ORB$^-$ as shown in the table on the right. ORB$^+$ cases are shaded in light blue.

conceptual framework of LGT. We define a "recombination breakpoint" to be a boundary of a genetic region introduced by an LGT event and incorporated via recombination into a genome. When by our approach we infer a recombination breakpoint to exist within the boundaries of a gene, we refer to that breakpoint as an "observable recombination breakpoint" and classify the corresponding set of orthologous genes as ORB$^+$. On the other hand, a gene set inferred to have undergone LGT but lacking a detectable internal recombination breakpoint is classified as ORB$^-$. Figure 1 illustrates how we classify recombination events that involve two adjacent genes, and extension to longer stretches of genes is straightforward. Thus, the designation ORB$^+$ indicates lateral transfer of a fragment of one or more genes, whereas ORB$^-$ indicates transfer of the whole gene (and possibly also of genomic sequence extending beyond that gene). Intuitively, we expect ORB$^+$ gene sets to have suffered more structural (and perhaps functional) disruption in the course of LGT than have the ORB$^-$ sets. Here we report the frequencies of gene sets that are ORB$^+$ and ORB$^-$, and discuss how each is correlated with annotated function and with phyletic group. To minimize, to the extent possible, the complications of paralogy and to increase the confidence with which we infer LGT events, we focus here on sets of single-copy, that is, putatively orthologous, genes. Because our approach requires multiple sequence alignment, by "gene" we necessarily refer specifically to the corresponding open reading frame; in prokaryotes these tend to be coextensive, or nearly so.

Materials and Methods

Data

From 144 completely sequenced prokaryote genomes, we generated 22,437 putatively orthologous protein sets of size $N \geq 4$ via a hybrid clustering approach (Harlow et al. 2004). We aligned these sequence sets (Beiko et al. 2005b) and validated the alignments using a pattern-centric objective function (Beiko et al. 2005a). These protein sequence alignments were converted into DNA sequence alignments by retrieving the corresponding nucleotide sequences from GenBank (http://www.ncbi.nlm.nih.gov/) and arranging the nucleotide triplets to parallel exactly the protein alignment in each case, yielding 18,809 gene sets ($N \geq 4$) containing a total of 139,707 genes. We require $N \geq 4$ because 4 is the minimum size that can yield distinct topologies; however, this is true only if every sequence in the set is unique. Therefore, we identified sets containing two or more identical sequences and removed (at random) all but one copy of each identical sequence; this further reduced our data set to 16,639 gene sets ($N \geq 4$) with 119,695 genes (open reading frames). In every case, the identical copies removed from consideration represented organisms either in the same genus (99.7%) or within the *Escherichia–Shigella* genus pair (0.3%); many represent different strains within the same species (89.1%). It is possible that some of these represent (within-gene or whole-gene) LGT, but such cases could not have been detected by our (or any other existing) approach in any case.

To minimize erroneous inference arising from paralogous sequences, we further restricted the data set to those 1,462 gene sets for which each member represents a different genome. In this data set, these sets of single-copy genes range in size from 4 to 52 members, totaling 11,128 sequences (supplementary fig. S1, Supplementary Material online).

The average length of the DNA alignments containing four or more sequences is 893 nt (shortest 72 nt; longest 28,317 nt).

Detecting Within-Gene (Fragmentary) Genetic Transfer

We applied a two-phase strategy for detecting recombination (Chan et al. 2007) which in previous studies has been shown to yield high-accuracy inference of recombination events using simulated data (Chan et al. 2006, 2007). In the first phase, PhiPack (Bruen et al. 2006) was used to detect the occurrences of recombination based on discrepancies of phylogenetic signal within the sequence alignments. This program incorporates *P* values of the Neighbor Similarity Score statistics in Reticulate (Jakobsen and Easteal 1996), the MaxChi test (Maynard Smith 1992), and PHI (Bruen et al. 2006). Data sets with at least two of the three *P* values $\leq$0.10 were considered as positive for recombination and were taken forward to the second phase of analysis.

In the second phase, for each sequence set that showed evidence of recombination (above), a Bayesian phylogenetic approach was used to delineate recombination breakpoints; this was implemented in DualBrothers (Minin et al. 2005) run with Markov chain Monte Carlo (MCMC) chain length = 2,500,000, burnin = 500,000, window_length = 5, and Green's constant C = 0.25. The tree search space for each run of DualBrothers was defined by a list of unrooted tree topologies inferred using MrBayes (Huelsenbeck and Ronquist 2001). We applied MrBayes to sliding windows of length 100 alignment positions, incremented by 50 positions per step. We set the MrBayes parameters as follows: MCMC chain length = 2,500,000, burnin = 500,000, nucmodel = 4by4, rates = gamma, ngammacat = 4. Topologies within a 90% Bayesian confidence interval were combined from each window and limited to 1,000 trees maximum. The posterior output of DualBrothers was then used to identify gene sets exhibiting evidence of within-gene transfer.

Detecting Whole-Gene (nonfragmentary) Genetic Transfer

For each gene set, we inferred a Bayesian phylogenetic tree (see below) and compared its topology against that of a reference tree; whole-gene transfer was inferred if the topologies were significantly discordant. These individual gene-set trees were inferred from DNA alignments (above). As reference, we used the Matrix Representation with Parsimony (MRP) supertree (Ragan 1992) computed from all well-supported (Bayesian posterior probability [BPP] $\geq$ 0.95) bipartitions among all individual protein-set trees in these 144 genomes (Beiko et al. 2005b). The individual gene-set trees were inferred using MrBayes (Huelsenbeck and Ronquist 2001) with MCMC chain length = 2,500,000, burnin = 500,000, and model = K2P (Kimura 1980) as described in Beiko et al. (2005b). We assessed possible discordance between individual gene-set trees and the reference supertree topology under likelihood models captured in the 1) Shimodaira–Hasegawa test (Shimodaira and Hasegawa 1999), 2) the Kishino–Hasegawa test (Kishino and Hasegawa 1989; Goldman et al. 2000), and 3) expected likelihood weights (Strimmer and Rambaut 2002), all as implemented in Tree-Puzzle 5.1 (Strimmer and von Haeseler 1996). Discordance was inferred if any topology was rejected by two or more of the three maximum likelihood tests at a confidence interval of 95% ($P \leq 0.05$) and was taken as prima facie evidence of whole-gene lateral transfer. In independent analyses, we also assessed possible topological discordance using the approximately unbiased (AU) test (Shimodaira 2002).

Tendency for LGT and Its Correction for Gene-Set Size

We devised a simple statistical test to evaluate the tendency for a gene set to undergo LGT, given its size. If there is an intrinsic background probability of LGT across all gene families, then the probability of observing transfer in a large data set is greater than the probability of observing transfer in a smaller one because more sequences mean more opportunities for observing a genetic transfer event. To compare frequencies of recombination across gene sets of different sizes, we implement an approach similar to the "linear normalization" described by Price et al. (2005), which was used to correct for homologous gene-set size in assessing the accuracy of homology search. For a fair comparison among gene sets of different sizes, we define the expected probability of recombination for a gene set as follows. Let f denote the observed frequency of recombination across the entire data set; in our case, $f = 489/1{,}462 = 0.334$. Let n_y denote the observed number of gene sets of size y. Then, for gene sets ranging in size from 4 to S members, we have:

$$F(\text{Recomb}|x) = \min\left(fx \sum_{y=4}^{S} \frac{n_y}{yn_y}, 1\right),$$

in which x represents a gene-set size, and $F(\text{Recomb}|x)$ is the expected frequency of recombination for a gene set of size x. The equation above can be used to define a null distribution of recombinant gene-set sizes, D_{null}. Using the Kolmogorov–Smirnov test (Durbin 1973), we then test whether the observed distribution of gene-set sizes (D_{obs}) could be generated by our D_{null}. Based on a set of 100 simulated D_{null} distributions that account for gene-set sizes, we find that D_{obs} is unlikely to have been generated by our D_{null} (median $P = 2.9 \times 10^{-9}$, mean $D = 0.2$). In a similar manner, we also independently compared the D_{obs} distributions obtained from the ORB$^+$ and ORB$^-$ gene sets against the corresponding D_{null} distributions.

Functional Analysis of Gene Sets

Functional information for each protein sequence was retrieved from the Comprehensive Microbial Resource (CMR) at The J. Craig Venter Institute (JCVI) Web site (http://cmr.jcvi.org/) and is based on JCVI role identifiers (Mainrole). Over- or underrepresentation of functional categories and taxonomic groups was based on the probability of observing a defined number of target groups (or categories) in a subsample, given a process of sampling without replacement from the whole data set (as defined in each case: see text) under a hypergeometric distribution (Johnson et al. 1992). The probability of observing x number of a particular target category is described as:

$$P(k = x) = f(k; N, m, n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}},$$

where N represents the total population size, m the size of the target category within the population, n the total size of the subsample, and k the size of the target category within the subsample.

Results

For discovery of LGT events in prokaryote genomes, we extracted a subset of the 22,437 putatively orthologous gene sets used in a previous large-scale study (Beiko et al. 2005b) on LGT in 144 phyletically diverse prokaryote genomes (see Materials and Methods). This subset, 1,462 gene sets, was restricted to sets of single-copy genes, that is, genes that are sufficiently unique within their respective genomes to make it unlikely that they have arisen by gene duplication. By applying this restriction we ensure, to the greatest extent possible, that any inferred recombination is due to LGT, not to the presence of (or recombination with) a paralogous gene copy.

These 1,462 gene sets range in size from 4 to 52 members each (supplementary fig. S1, Supplementary Material online); 1,229 (84.1%) of the sets contain $\leq$10 sequences, with almost a quarter of the 1,462 (362, 24.7%) of size 4. Gene sets of size <4 were excluded from the analysis, as they do not contribute to meaningful phylogenetic inference. Each of the 1,462 gene sets was examined for evidence of LGT, as described below.

Gene Sets with ORBs

We applied a two-phase strategy (Chan et al. 2007) for detecting recombination in each of the 1,462 sets of single-copy genes. In the first phase, we used three statistical measures (Maynard Smith 1992; Jakobsen and Easteal 1996; Bruen et al. 2006) to search for evidence of phylogenetic discrepancy (i.e., a recombination signal) within each gene set. If at least two of the three tests show a P value $\leq$0.10 in support of recombinant ancestry, the gene set was passed on to a second phase of recombination inference. In the second phase, we utilized a Bayesian phylogenetic approach, implemented in the software program DualBrothers (Minin et al. 2005), to locate recombination breakpoints more precisely in the putatively recombinant gene sets. DualBrothers employs reversible-jump MCMC and a dual multiple change-point model to identify, within a set of sequences, contiguous regions that share a common tree topology and the boundaries (recombination breakpoints) between regions that support different topologies (Suchard et al. 2003; Minin et al. 2005).

Instances of recombination discovered using this approach are thus ORB$^+$ gene sets, as at least one end of a topologically distinct region (i.e., a recombination breakpoint) occurs within the sequence set used in our analysis. ORB$^-$ gene sets escape detection because no point of topological discontinuity can be inferred inside the gene set.

Our first-phase screening produced preliminary evidence of recombination in 426 (29.1%) of these 1,462 gene sets, and Bayesian inference of recombination breakpoints was applied to those 426. Following the classification system reported in a previous study (Chan et al. 2009), we found clear evidence of recombination breakpoints within 286 of these gene sets (19.6% of 1462), where "clear evidence" is defined as BPP support $\geq$0.500 for the dominant topology (as defined internally with respect to the individual alignment) on at least one side of the inferred breakpoint. We found a further 80 cases (5.5%) in which a breakpoint was located, but no sequence region supports a single topology with BPP $\geq$0.500; we classified these as inconclusive and excluded them from further consideration. Finally, we observed 60 cases (4.1%) for which recombination was indicated in the initial screening, but no recombination breakpoint could be identified.

Figure 2*A* shows the size distribution of these 286 gene sets; the most-populated classes are of eight (42 sets, 14.7% of 286) and six sequences each (39 sets, 13.6%). After normalization by gene-set size, we found that small ORB$^+$ gene sets are observed more frequently than expected (median $P < 10^{-5}$, mean $D = 0.2$ in 100 comparisons); this may indicate a higher susceptibility of small gene sets to lateral transfer of gene fragments and/or may reflect a greater sensitivity of recombination detection methods when sequence sets are small.

Gene Sets with no ORB

We inferred phylogenetic trees for all 1,462 gene sets and compared the inferred topology with a reference tree. The reference (species) tree (Beiko et al. 2005b) was generated using MRP (Ragan 1992), yielding a supertree that summarizes all well-supported (BPP $\geq$ 0.95) bipartitions among the 22,432 trees of putatively orthologous sets in these 144 prokaryote genomes (Beiko et al. 2005b). In the absence of a detectable recombination breakpoint within the gene, phylogenetic discordance between a well-supported gene tree and the reference supertree can be treated as lateral transfer of the entire gene (and potentially of flanking intergenic regions and adjacent genes as well). Using a combination of three statistical tests (see Materials and Methods) we found 342 gene sets to be topologically incongruent with the reference tree, suggestive of LGT. Among these 342 gene sets, 152 were independently inferred as ORB$^+$ (above); the remaining 190 (13.0%) are thus ORB$^-$. The size distributions of these sets (fig. 2*B*) show that, on a per-gene basis, gene sets with fewer members are potentially more susceptible not only to within-gene lateral transfer (ORB$^+$ gene sets, above) but also to transfer of whole genes, than expected under our null model (median $P < 10^{-3}$, mean $D = 0.2$ in 100 comparisons).

Thus, in total, among the 1,462 single-copy gene sets, we found clear evidence of LGT in 476 (32.6%) using our approach, of which 286 (60.1%) are ORB$^+$ and 190 (39.9%) are ORB$^-$.

Note that 134 gene sets inferred as ORB$^+$ were not found to show strong topological incongruence with the reference tree. To the extent that this result is more broadly representative, it suggests that phylogenetic analyses based on entire genes may overlook almost half of the gene sets actually affected by within-gene LGT. This obviously raises questions, beyond the scope of the present investigation, about the extent and nature of the genetic regions responsible for incongruent signal (e.g., their size, contiguity, and statistical support), and whether protein sequence sets may be similarly susceptible.

The AU test (Shimodaira 2002) is reported to better adjust for Type I error than other statistical tests for comparing tree topologies, recognizing, however, that the appropriateness of any test depends also on the nature of the data and on a priori assumptions about tree topologies (Goldman et al. 2000). We therefore independently substituted the AU test in place of our multiple-test criterion to compare topologies: 489 gene sets (33.4% of 1462) showed strong discordance vis-à-vis the reference supertree, 160 (10.9% of 1462) of which are in the ORB$^+$ set and 329 (22.5% of 1462) of which are ORB$^-$. Thus, under the AU test, the complete LGT footprint is 286 ORB$^+$ (as above) plus 329 ORB$^-$ for a total of 615 gene sets (42.1% of 1462), 29% greater than the 476 identified using the multiple-test approach above. The 489 discordant gene sets

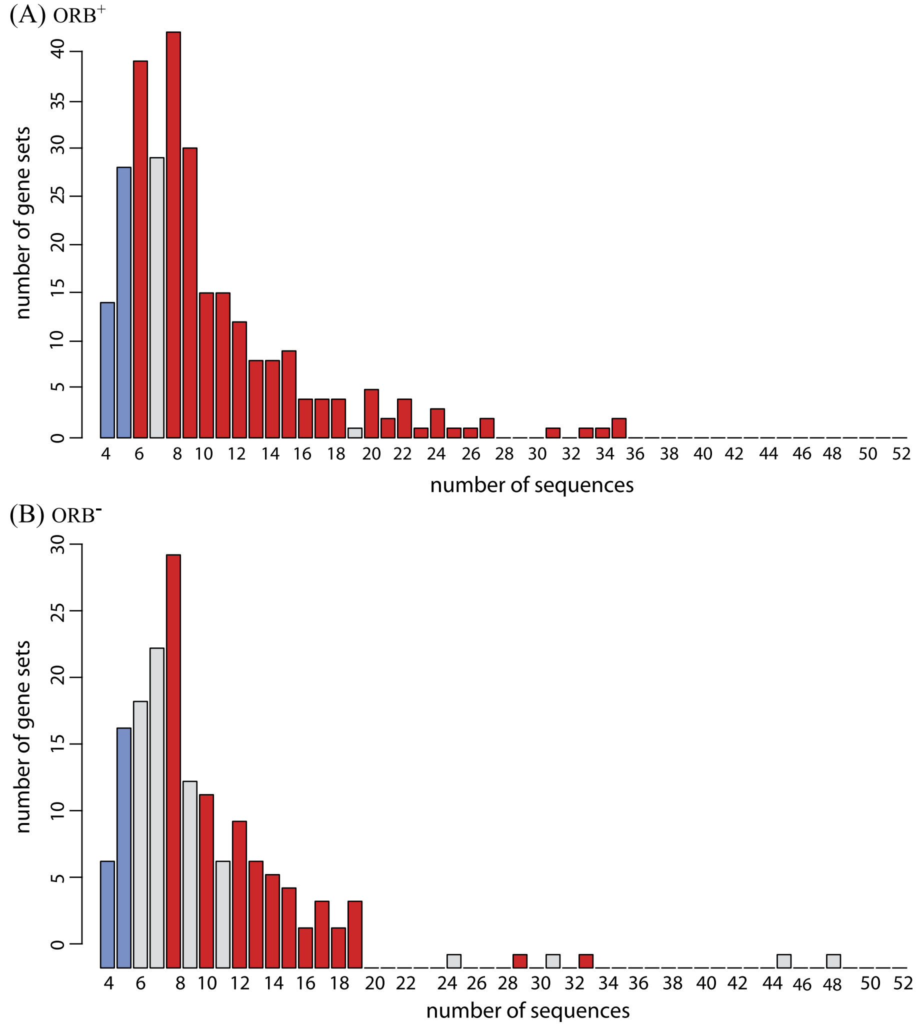


Fig. 2.—Size distribution of (*A*) ORB$^+$ and (*B*) ORB$^-$ gene sets. The solid red bars indicate overrepresented size classes; the solid blue bars indicate underrepresented classes; and the gray bars indicate classes neither over- nor underrepresented in comparison with the corresponding size-class frequency over all 1,462 gene sets, at $P \leq 0.05$.

include all 342 discordant sets identified using our multiple-test approach, demonstrating that our estimate of the extent of whole-gene transfer in prokaryotes is of high confidence and perhaps conservative. Interestingly, the proportion of gene sets inferred as ORB$^+$ but not recovered as topologically incongruent with the reference supertree is reduced only slightly (9%), from 134 under our multiple-test criterion to 126 under AU.

Functional Biases of ORB$^+$ and ORB$^-$ Gene Sets

We used annotations from the JCVI CMR (http://cmr.jcvi.org/) to assign a functional category (JCVI role category) to the protein associated with each gene in the 476 gene sets in which we inferred LGT. Details are provided in Materials and Methods. Figure 3 shows the proportions of proteins in each functional category, broken down by ORB$^+$ or ORB$^-$ classification.

Hypothetical proteins (i.e., corresponding to genes that show no significant similarity to genes from other organisms) constitute the major overrepresented category in both the ORB$^+$ and ORB$^-$ gene sets. A relatively tiny category of proteins related to viral functions (including transduction of DNA by phages) is the only other category similarly overrepresented among proteins corresponding to the ORB$^-$ gene sets. On the contrary, proteins involved in a range of biosynthetic, metabolic, protein-synthetic, transport, and binding functions are significantly underrepresented. Proteins that function in energy metabolism are underrepresented only in the case of ORB$^+$ gene sets (fig. 3*A*), whereas those engaged in DNA metabolism, central intermediary metabolism, and transcription are underrepresented only for ORB$^-$ sets (fig. 3*B*).

Phyletic Biases of LGT Leading to ORB$^+$ and ORB$^-$ Gene Sets

We next asked whether ORB$^+$ and ORB$^-$ gene sets are over- or underrepresented in particular taxa. Figure 4 shows the taxonomic origins (National Center for Biotechnology

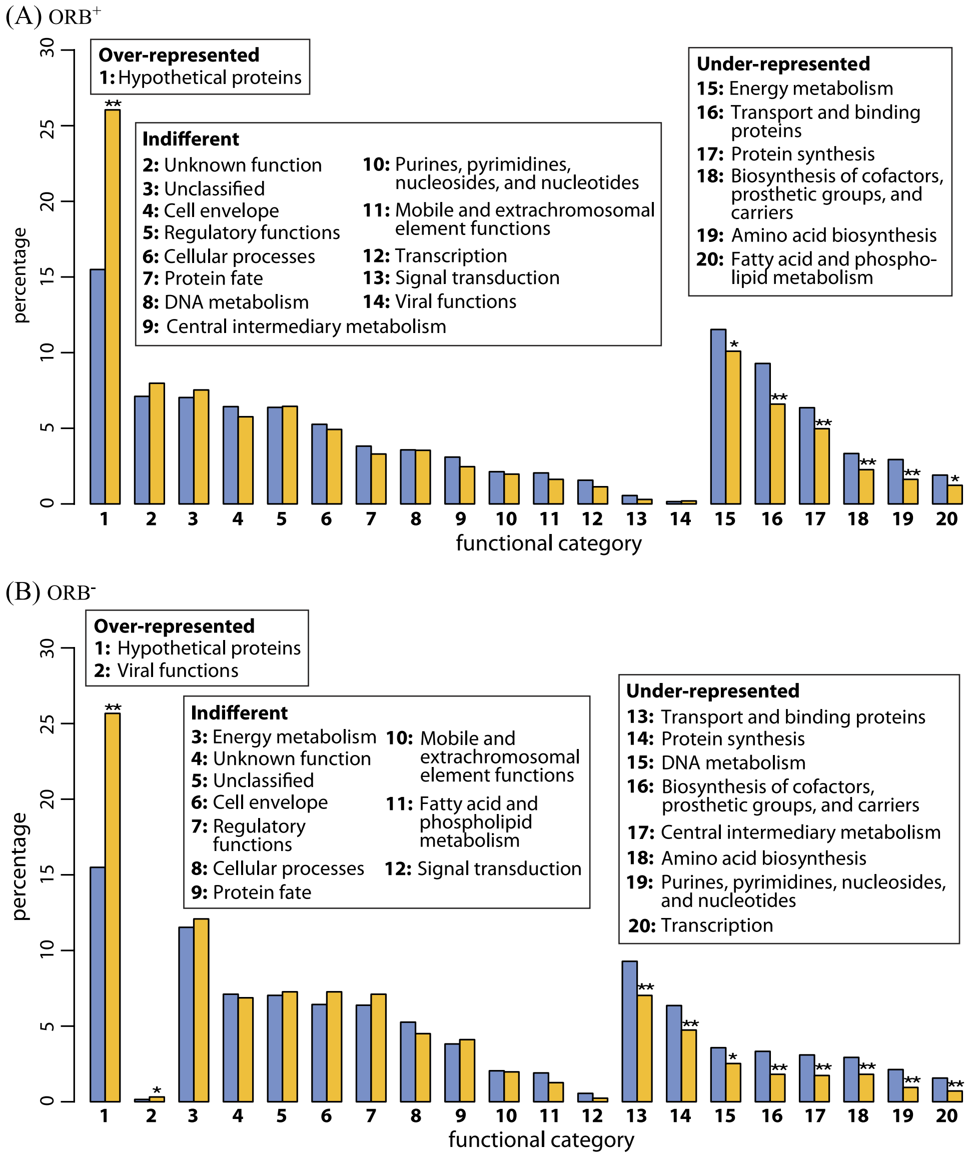


FIG. 3.—Representation of functional categories assigned to protein sequences corresponding to the (*A*) ORB$^+$ and (*B*) ORB$^-$ gene sets (solid yellow bars). The solid blue bars show the representation of these same functional categories in the full data set (16,639 gene sets of size $\geq$4, 119,695 proteins). Categories are numbered (differently for *A* and *B*) as shown in the boxes. Significance of over- or underrepresentation is represented by single ($P \leq 0.05$) and double asterisks ($P \leq 0.01$).

Information level-4 taxa) of proteins that correspond to each gene in these two groups of gene sets. For clarity, the corresponding proportions are not shown over the entire 144-genome (16,639 gene sets) data set; over- and underrepresentation ($P \leq 0.05$) are indicated by red and blue coloration, respectively. Our results reveal that sets of single-copy genes affected by LGT contain a significantly ($P \leq 0.05$) higher-than-expected proportion of genes originating from High–G+C Firmicutes, Planctomycetes, and Spirochaetales. This is true for both ORB$^+$ and ORB$^-$ sets.

Other taxonomic groups are overrepresented in only one of the two types of gene set. The data and our approach do not allow us to extrapolate with certainty, but to the extent that these single-gene sets are representative of complete genomes, the cyanobacteria, chlamydiales, and crenarchaeotes appear to be relatively receptive to introgression of gene fragments, whereas euryarchaeotes, chlorobi, and members of *Thermotoga* and *Aquifex* have been relatively receptive to transfer of entire genes or multigene regions. We note that many of the latter taxa are extremophiles, suggesting that further analysis relating LGT patterns and mechanisms (e.g., conjugation, transduction) to environmental factors may be warranted.

Sequences from low–G+C Firmicutes are underrepresented in both the ORB$^+$ and ORB$^-$ gene sets. Genes of Proteobacteria (the high-level taxon most abundantly represented in our data set) are underrepresented only in the ORB$^-$ set. Table 1 shows the genomes of individual isolates that are significantly enriched with either ORB$^-$ or ORB$^+$ gene set or both, in our data set. Many overrepresented species are pathogens.

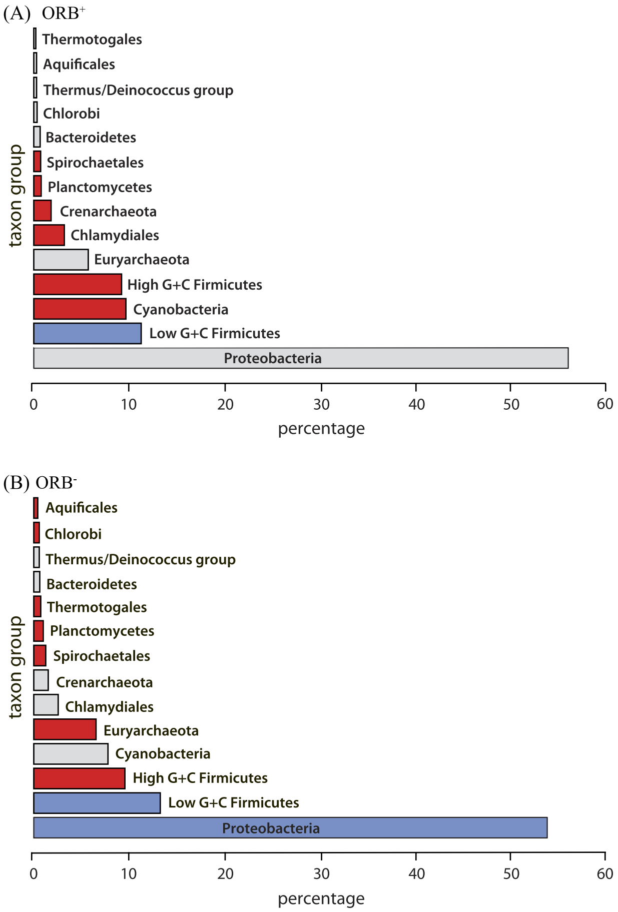


FIG. 4.—Taxonomic origins (National Center for Biotechnology Information level-4 taxa) of genes in the (*A*) ORB$^+$ and (*B*) ORB$^-$ gene sets. Overrepresentation relative to the 16,639 gene sets is shown in red; underrepresentation is shown in blue; gray indicates that there is neither over- nor underrepresentation at $P \leq 0.05$.

Table 1
Species That Are Overrepresented ($P \leq 0.05$) in ORB+ and/or ORB- Gene Sets, in Comparison with Their Contribution to the 16,639 Gene Sets

ORB$^+$ Gene Sets	ORB$^-$ Gene Sets
Nostoc sp. PCC 7120	*Salmonella enterica* subsp. *enterica* serovar Typhi Ty2
Streptomyces avermitilis MA-4680	*Nitrosomonas europaea* ATCC 19718
Shewanella oneidensis MR-1	*Yersinia pestis* KIM
Yersinia pestis CO92	*Methanothermobacter thermautotrophicus*
Synechococcus sp. WH 8102	*Leptospira interrogans* serovar lai str. 56601
Thermosynechococcus elongatus BP-1	*Thermotoga maritima*
Pasteurella multocida	*Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586
Methanococcus jannaschii	*Chlorobium tepidum* TLS
Methanopyrus kandleri AV19	*Aquifex aeolicus*
Halobacterium sp. NRC-1	*Chlaymydophila pneumonia* J138
Chlamydophila pneumoniae CWL029	*Treponema pallidum*
	Streptococcus pyogenes MGAS315
Photorhabdus luminescens subsp. *laumondii* TTO1	
Haemophilus ducreyi 35000HP	
Pirellula sp.	
Borrelia burgdorferi	
Mycoplasma pulmonis	

NOTE.—Species are listed in descending order, from the most overrepresented to the least overrepresented, separately for the ORB$^+$ and ORB$^-$ gene set. The species with underlined citations are pathogens. The five species listed separately at the bottom of the table are overrepresented in both sets of ORB$^+$ and ORB$^-$.

Discussion

Our results demonstrate that most single-copy genes from diverse prokaryotes show no strongly supported phylogenetic discordance or evidence of intergenome recombination, consistent with the idea that a cohesive signal may be present in these genomes. However, a substantial minority of these single-copy gene sets shows clear evidence of LGT; at least 19.6% contain an observable recombination breakpoint (ORB$^+$), and a further 13.0% were phylogenetically discordant but do not contain an observable recombination breakpoint (ORB$^-$). In previous studies, estimates of the frequency of LGT range widely: 2% (Ge et al. 2005), 13% (Beiko et al. 2005b), 16% (Kunin and Ouzounis 2003), 60% (Lerat et al. 2005), to as high as 90% (Mirkin et al. 2003) of genes or bipartitions. In a study based on inference of ancestral genome sizes (Dagan and Martin 2007), all genes in prokaryotes were proposed to have had undergone LGT at some point in their histories. Several factors contribute to this range of estimates, including but not limited to the choice of methodological approach and sampling of genes and genomes (Ragan and Beiko 2009). Different methodologies can produce not only different estimates of the extent of LGT but also incompatible lists of lateral genes on the same data set (Ragan 2001). The phylogenetic approach to detection of LGT is firmly grounded in biological principle (the same principles as those responsible for inheritance and diversification of lineages) and can be carried out in a statistically rigorous manner, although systematic biases, for example, surrounding the model of sequence change, may still intrude. ORBs might also arise within a gene set due to genetic conversion subsequent to an LGT event resulting in different evolutionary rates being inferred for different gene regions (Chan et al. 2007) or via duplication of horizontally transferred gene fragments. In previous analysis of the same data sets (Chan et al. 2009), we did not observe rate differences of a magnitude likely to confound these analyses; and by limiting our data set to single-copy gene families, we ensure, to the extent possible, that inferred ORBs have not arisen from recombination with a paralogous sequence.

A limitation of the phylogenetic approach as adopted in previous studies (e.g., Beiko et al. 2005b; Lerat et al. 2005), however, has been the underlying assumption that the unit of genetic transfer is an entire gene. Topological discordance between a gene-set tree and the reference topology has been interpreted as prima facie evidence that a gene has been transferred from one lineage into another. Here, we have employed a phylogenetic approach but without restricting the unit of transfer to be a whole gene and show that among these diverse prokaryotic species, LGT can involve the recombination of a fragment smaller than a gene and/or the interruption of an existing gene. Indeed, over the sets of single-copy genes in these 144 prokaryotic genomes, gene sets with observed recombination breakpoints are about 50% more frequent than those of inferred lateral origin for which no breakpoint is observed, suggesting that LGT more commonly interrupts genes than preserves them intact.

Although our approach relaxes assumptions that the gene is the fundamental unit of lateral transfer, it suffers certain limitations of its own that must be considered when interpreting these results. First, our approach does not consider the arrangement or grouping of genes within modern or ancestral genomes: a single LGT event transferring multiple genes, whether wholly or in part, would appear as multiple LGT events. As a consequence, it is unwarranted to interpret our results as indicating a number or frequency of transfer events. Second, the taxonomic sampling of our data set is based on organisms for which whole-genome sequences are available. For the most part, strains have been selected for sequencing based on their availability in culture collections and/or their pathogenicity or economic significance, and thus are unlikely to be representative of extant (much less past) microbial diversity. It is unknown whether taxon availability and sampling may bias inference regarding the frequency of LGT in different clades. Third, our method for discovering recombination breakpoints within genes requires orthologous gene (open reading frame) sets to be aligned and does not take heterogeneity in sequence composition or nonstationarity of the substitution process into account. As our data set covers a broad range of genomic G+C content, such issues are clearly relevant here, and in extreme cases may have led to erroneous inference of breakpoints. In addition, our approach would not have detected the transfer of genes or gene fragments that share high sequence similarity, that is, recombination within gene regions that exhibit weak (or unsubstantiated) phylogenetic signal, and as such will tend to underestimate the extent of LGT. As data sets grow and models for phylogenetic inference become more realistic, it will be necessary to revisit the issue of within-gene LGT. Our present analysis should be considered a best-practice examination of within-gene LGT, given current data sets and inference methods.

The data set used in this study is a subset of that of Beiko et al. (2005b), who concluded that some 13%–14% of bipartitions are discordant and potentially affected by LGT. Here we report that at least 32.6% of gene sets are affected by LGT. These two numbers are not directly comparable for three reasons: 1) our present subset contains only sets of single-copy genes; 2) our data set is smaller, having 74% as many gene sets and 54% as many sequences; and 3) we base our analyses on gene sets, not on bipartitions, as genetic transmission involving within-gene recombination can be only partially mapped into the paradigm of bipartitions and subtrees. Neither we nor Beiko et al. (2005b) attempted to estimate LGT in paralogous gene families or among very closely related genomes. Again we are reminded of the multifaceted trade-offs between methodological rigor and the goal of a more global estimate of frequency of LGT in prokaryotes. Our findings must therefore be interpreted carefully, especially with regard to LGT in duplication-rich clades such as Proteobacteria (Gevers et al. 2004). Other studies have reported high rates of LGT for some Proteobacteria (Gogarten et al. 2002; Lerat et al. 2003).

Our results also indicate that, after normalization for numbers of genes in each set, small gene sets with evidence of LGT are more frequent than expected under a random model. This is the case for both whole-gene (the ORB$^-$ gene sets) and within-gene transfer (the ORB$^+$ sets). Nonetheless, we urge caution in interpreting this result. Gene-set size is correlated with degree of sequence divergence; many, although not all, small gene sets are constituted of sequences from closely related organisms (genomes) that have only recently diverged from a common ancestor (Pushker et al. 2004). These sequences presumably are, or have recently been, preferentially susceptible to homologous recombination events that changed the sequence very little and thus remain cryptic to our, and indeed all other, approaches. Other small gene sets exhibit patchy phyletic distributions most parsimoniously explained by LGT. Further, if multiple instances of LGT are common in individual gene sets (of any size), setting a conservative upper bound on the value of the normalizing factor, as we do here, could skew the D_{obs} distribution toward high values. Interestingly, we observe in our data that large gene sets (size $N > 10$) tend to have a greater number of distinct breakpoints (β) than small gene sets: The mean β is 6.6, compared with a mean β of 1.3 in gene sets of $N \leq 10$ ($P < 10^{-16}$). Our findings describe the propensity of a gene set to have suffered LGT, not the number of LGT events, and therefore, our result must be interpreted strictly against that definition. For ORB$^-$ gene sets, the shortest constrained subtree prune-and-regraft distance (Beiko and Hamilton 2006) might serve as surrogate for number of LGT events, but novel approaches would presumably be required to estimate this for ORB$^+$ sets. Finally, the sensitivity of different recombination detection approaches in light of gene-set size has not been explored in depth; if adding more sequences to a set diminishes the probability of detecting a single recombination event (e.g., due to the use of an overly conservative multiple-test correction such as the Bonferroni), then large sets will be more susceptible to false negatives.

We observed only modest differences in functional bias between within-gene and whole-gene transfers in these gene sets, with hypothetical proteins very significantly ($P \leq 0.01$) overrepresented in both cases. Gene products classified by the JCVI CMR as of "unknown function" (i.e., corresponding to genes that show significant similarity to genes in other organisms, for which the function has not been identified) are not significantly over- (or under-)represented, suggesting that an exogenous or hybrid origin does not significantly decrease (or increase) annotation of a functional role category. We also found that genes encoding viral functions are more likely to be laterally transferred in their entirety than as fragments. A similar trend is observed for pathogenic bacteria, which are prominent among the organisms that contribute disproportionately to gene sets affected by nonfragmentary transfer. Genes that encode virulence factors (e.g., toxins, adhesions, and invasins) are commonly located on mobile genetic elements such as plasmids and transposons, or in specific genomic region called pathogenicity islands (Ilyina and Romanova 2002; Hacker et al. 2004).

Genes annotated as involved in DNA metabolism, transcription, and protein synthesis are underrepresented among the ORB$^-$ gene sets. Of these, only the "protein synthesis" functional category is underrepresented among the ORB$^+$ sets. The complexity hypothesis (Jain et al. 1999) postulates that "informational" proteins involved in processes related to transcription and translation, including many in these three categories, typically function in the cell within large multiprotein complexes; they must interact in finely tuned ways with many other biomolecules, and it may be a consequence that their genes are less likely to be susceptible to successful LGT than are genes encoding the putatively less interactive "operational" proteins. Our results do not speak directly to the validity of this hypothesis, but suggest that any bias against transfer of informational genes may be expressed more strongly in the case of whole-gene than within-gene genetic transfer.

Supplementary Material

Supplementary figure S1 is available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Funding

This work was supported by the Australian Research Council (CE0348221 to M.A.R.) and National Science Foundation (DBI-0630765 to A.E.D.). R.G.B. is supported by the Canada Research Chairs program.

Acknowledgments

We thank Vladimir Minin for valuable advice on the use of DualBrothers.

Literature Cited

Barlow M. 2009. What antimicrobial resistance has taught us about horizontal gene transfer. Methods Mol Biol. 532:397–411.

Beiko RG, Chan CX, Ragan MA. 2005a. A word-oriented approach to alignment validation. Bioinformatics. 21:2230–2239.

Beiko RG, Hamilton N. 2006. Phylogenetic identification of lateral genetic transfer events. BMC Evol Biol. 6:15.

Beiko RG, Harlow TJ, Ragan MA. 2005b. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A. 102:14332–14337.

Bork P, Doolittle RF. 1992. Proposed acquisition of an animal protein domain by bacteria. Proc Natl Acad Sci U S A. 89:8990–8994.

Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics. 172:2665–2681.

Chan CX, Beiko RG, Ragan MA. 2006. Detecting recombination in evolving nucleotide sequences. BMC Bioinform. 7:412.

Chan CX, Beiko RG, Ragan MA. 2007. A two-phase strategy for detecting recombination in nucleotide sequences. S Afr Comput J. 38:20–27.

Chan CX, Darling AE, Beiko RG, Ragan MA. 2009. Are protein domains modules of lateral genetic transfer? PLoS ONE. 4:e4524.

Ciccarelli FD, et al. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science. 311:1283–1287.

Creevey CJ, et al. 2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes? Proc R Soc Lond B Biol Sci. 271:2551–2558.

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

Denamur E, et al. 2000. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell. 103:711–721.

Didelot X, Achtman M, Parkhill J, Thomson NR, Falush D. 2007. A bimodal pattern of relatedness between the *Salmonella paratyphi* A and *typhi* genomes: convergence or divergence by homologous recombination? Genome Res. 17:61–68.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science. 284:2124–2128.

Durbin J. 1973. Distribution theory for tests based on the sample distribution function. Philadelphia (PA): SIAM.

Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive earth's biogeochemical cycles. Science. 320:1034–1039.

Fournier GP, Gogarten JP. 2008. Evolution of acetoclastic methanogenesis in *Methanosarcina* via horizontal gene transfer from cellulolytic Clostridia. J Bacteriol. 190:1124–1127.

Ge F, Wang LS, Kim J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3:e316.

Gevers D, Vandepoele K, Simillon C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. 12:148–154.

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

Gogarten JP, Townsend JP. 2005. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol. 3:679–687.

Goldman N, Anderson JP, Rodrigo AG. 2000. Likelihood-based tests of topologies in phylogenetics. Syst Biol. 49:652–670.

Grundmann H, Aires-de-Sousa M, Boyce J, Tiemersma E. 2006. Emergence and resurgence of meticillin-resistant *Staphylococcus aureus* as a public-health threat. Lancet. 368: 874–885.

Hacker J, et al. 2004. Pathogenomics of mobile genetic elements of toxigenic bacteria. Int J Med Microbiol. 293:453–461.

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

Harlow TJ, Gogarten JP, Ragan MA. 2004. A hybrid clustering approach to recognition of protein families in 114 microbial genomes. BMC Bioinform. 5:45.

Hartl DL, Lozovskaya ER, Lawrence JG. 1992. Nonautonomous transposable elements in prokaryotes and eukaryotes. Genetica. 86:47–53.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Igarashi N, et al. 2001. Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. J Mol Evol. 52:333–341.

Ilyina TS, Romanova YM. 2002. Bacterial genomic islands: organization, function, and evolutionary role. Mol Biol. 36:171–179.

Inagaki Y, Susko E, Roger AJ. 2006. Recombination between elongation factor 1-alpha genes from distantly related archaeal lineages. Proc Natl Acad Sci U S A. 103:4528–4533.

Iwasaki W, Takagi T. 2009. Rapid pathway evolution facilitated by horizontal gene transfers across prokaryotic lineages. PLoS Genet. 5:e1000402.

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

Jakobsen IB, Easteal S. 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput Appl Biosci. 12:291–295.

Johnson NL, Kotz S, Kemp AW. 1992. Univariate discrete distributions. New York: Wiley.

Kechris KJ, Lin JC, Bickel PJ, Glazer AN. 2006. Quantitative exploration of the occurrence of lateral gene transfer by using nitrogen fixation genes as a case study. Proc Natl Acad Sci U S A. 103:9584–9589.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in *Hominoidea*. J Mol Evol. 29:170–179.

Koonin EV. 2009. Darwinian evolution in the light of genomics. Nucleic Acids Res. 37:1011–1034.

Kunin V, Ouzounis CA. 2003. The balance of driving forces during genome evolution in prokaryotes. Genome Res. 13:1589–1594.

Lerat E, Daubin V, Moran NA. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol. 1:e19.

Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 3:e130.

Lin CH, Bourque G, Tan P. 2008. A comparative synteny map of *Burkholderia* species links large-scale genome rearrangements to fine-scale nucleotide variation in prokaryotes. Mol Biol Evol. 25:549–558.

Maynard Smith J. 1992. Analyzing the mosaic structure of genes. J Mol Evol. 34:126–129.

Minin VN, Dorman KS, Fang F, Suchard MA. 2005. Dual multiple change-point model leads to more accurate recombination detection. Bioinformatics. 21:3034–3042.

Mirkin BG, Fenner TI, Galperin MY, Koonin EV. 2003. 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.

Mrázek J, Karlin S. 1999. Detecting alien genes in bacterial genomes. Ann N Y Acad Sci. 870:314–329.

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

Nesbø CL, et al. 2009. The genome of *Thermosipho africanus* TCF52B: lateral genetic connections to the Firmicutes and Archaea. J Bacteriol. 191:1974–1978.

Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV. 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement *in situ*. Genome Biol. 4:R55.

Price GA, Crooks GE, Green RE, Brenner SE. 2005. Statistical evaluation of pairwise protein sequence comparison with the Bayesian bootstrap. Bioinformatics. 21:3824–3831.

Pushker R, Mira A, Rodríguez-Valera F. 2004. Comparative genomics of gene-family size in closely related bacteria. Genome Biol. 5:R27.

Ragan MA. 1992. Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol. 1:53–58.

Ragan MA. 2001. On surrogate methods for detecting lateral gene transfer. FEMS Microbiol Lett. 201:187–191.

Ragan MA, Beiko RG. 2009. Lateral genetic transfer: open issues. Philos Trans R Soc B. 364:2241–2251.

Shi T, Falkowski PG. 2008. Genome evolution in cyanobacteria: the stable core and the variable shell. Proc Natl Acad Sci U S A. 105:2510–2515.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51:492–508.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16:1114–1116.

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

Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc R Soc Lond B Biol Sci. 269:137–142.

Strimmer K, von Haeseler A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol. 13:964–969.

Suchard MA, Weiss RE, Dorman KS, Sinsheimer JS. 2003. Inferring spatial phylogenetic variation along nucleotide sequences: a multiple change-point model. J Am Stat Assoc. 98:427–437.

Woese CR. 2000. Interpreting the universal phylogenetic tree. Proc Natl Acad Sci U S A. 97:8392–8396.

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

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

William Martin, Associate Editor

Accepted October 31, 2009