

Patterns of Bacterial Gene Movement

Weilong Hao and G. B. Golding

Department of Biology, McMaster University, Hamilton, Ontario, Canada

Lateral gene transfer has emerged as an important force in bacterial evolution. A substantial number of genes can be inserted into or deleted from genomes through the process of lateral transfer. In this study, we looked for atypical occurrence of genes among related organisms to detect laterally transferred genes. We have analyzed 50 bacterial complete genomes from nine groups. For each group we use a 16s rRNA phylogeny and a comparison of protein similarity to map gene insertions/deletions onto their species phylogeny. The results reveal that there is poor correlation of genes inserted, deleted, and duplicated with evolutionary branch length. In addition, the numbers of genes inserted, deleted, or duplicated within the same branch are not always correlated with each other. Nor is there any similarity within groups. For example, in the *Rhizobiales* group, the ratio of insertions to deletions in the evolutionary branch leading to *Agrobacterium tumefaciens* str. C58 (Cereon) is 0.52, but it is 39.52 for *Mesorhizobium loti*. Most strikingly, the number of insertions of foreign genes is much larger in the external branches of the trees. These insertions also greatly outnumber the occurrence of deletions, and yet the genome sizes of these bacteria remain roughly constant. This indicates that many of the insertions are specific to each organism and are lost before related species can evolve. Simulations of the process of insertion and deletion, tailored to each phylogeny, support this conclusion.

Introduction

The number of genes in bacterial genomes are not stable (Mira, Ochman, and Moran 2001) but rather, over evolutionary time many genes can be deleted, duplicated, or inserted. There are two general classes of mechanisms involved in the generation of evolutionary novelty (Dutta and Ran 2002). One is the duplication/modification of genes and the other is the acquisition of foreign genes via lateral gene transfer (LGT). Genes that arise from LGT may carry out new functions and may result in adaptation due to a changing environment. A large number of genes can be inserted into or deleted from genomes through the process of lateral transfer. Some genes may even be transferred from very distantly related species, such as from bacteria to eukaryotes (Keeling and Doolittle 1997; Jain et al. 2002; Gogarten 2003; Klotz and Loewen 2003). For example, Martin et al. (2002) estimate 18% of the nuclear genes in *Arabidopsis* are cyanobacterial in origin. As more genome sequences become available, more information about the nature of lateral gene transfer can be obtained. Insertions and/or deletions resulting from lateral gene transfer are widely thought to contribute to genome size variation and to differing pathogenic traits (Rode et al. 1999).

In bacteria, LGT may be more important as a mechanism of evolutionary change than small-scale change within genes. Laterally transferred genes can play a crucial role in the ability of bacteria to invade novel niches. Some transferred genes are involved in a bacterium's ability to cause different diseases, and occasionally transferred genes can lead directly to the evolution of pathogenic strains from nonpathogenic strains (Hacker et al. 1990; Hacker and Kaper 2000; Ochman et al. 1996).

There are several methods for detecting lateral gene transfer. One is based on atypical sequence characteristics,

such as codon usage bias and G + C content. This is because the bias in codon usage is often species specific (Lawrence and Ochman 1998) and because genomic DNA from different organisms has a specific mean G + C content (Garcia-Vallvé, Romeu, and Palau 2000). Using this method, more than 18% of genes in *Escherichia coli* were found to be laterally transferred genes (Lawrence and Ochman 1998), as were approximately 24% of genes in *Thermotoga maritima* (Garcia-Vallvé, Romeu, and Palau 2000). But codon usage bias and G + C content are not always sufficiently reliable to detect laterally transferred genes (Koski, Morton, and Golding 2001). Another method by which to detect LGT is to reconstruct a phylogeny for the genes. This method uses a comparison of phylogenetic trees among individual genes within closely related species to recognize unusual origins. This approach is very powerful; nevertheless it may fail as a result of a variety of potential errors such as a difficulty in recognizing orthologous genes.

Here we have taken a simpler approach of determining the presence/absence of genes in related bacteria. In most cases, enough genome rearrangements have occurred to prevent the inference of the number of insertion/deletion events but the change in the total number of genes due to insertions/deletions can be determined. We have mapped the number of gene insertions, deletions, and gene duplications onto the phylogeny for the species based on a 16s rRNA phylogeny. Unlike the relationship between divergence of sequence and gene order degradation in closely related bacterial genomes (Suyama and Bork 2001), the insertions, deletions, and duplications are not strongly correlated with evolutionary branch length. Insertions are shown to be most numerous in external branches.

Methods

Fifty complete genome sequences were selected as the basis for an analysis of gene insertion and deletion (table 1). Genomes were selected to be within the same group based on perceived relatedness and, whenever possible, consisted of congeneric species (when at least

Key words: lateral gene transfer, bacterial genomes, phylogeny, molecular evolution.

E-mail: golding@mcmaster.ca

Mol. Biol. Evol. 21(7):1294–1307. 2004

doi:10.1093/molbev/msh129

Advance Access publication March 28, 2004

Table 1
Related Genomes Analyzed

Group Genus	Species (Strain)	Outgroup(s) ^a
<i>Pseudomonadaceae</i>		
<i>Pseudomonas</i>	<i>P. aeruginosa</i> <i>P. fluorescens</i> <i>P. putida</i> <i>P. syringae</i>	<i>H. influenzae</i> <i>V. cholerae</i>
<i>Chlamydiaceae</i>		
<i>Chlamydia</i>	<i>C. muridarum</i> <i>C. trachomatis</i>	<i>L. interrogans</i>
<i>Chlamydophila</i>	<i>C. caviae</i> <i>C. pneumoniae</i> AR39 <i>C. pneumoniae</i> CWL029 <i>C. pneumoniae</i> J138	
<i>Mycoplasmataceae</i>		
<i>Mycoplasma</i>	<i>M. gallisepticum</i> <i>M. genitalium</i> <i>M. penetrans</i> <i>M. pneumoniae</i> <i>M. pulmonis</i>	<i>C. acetobutylicum</i> <i>C. perfringens</i>
<i>Ureaplasma</i>	<i>U. urealyticum</i>	
<i>Streptococcaceae</i>		
<i>Lactococcus</i>	<i>L. lactis</i>	<i>E. faecalis</i>
<i>Streptococcus</i>	<i>S. agalactiae</i> NEM316 <i>S. agalactiae</i> 2603V/R <i>S. mutans</i> <i>S. pneumoniae</i> R6 <i>S. pneumoniae</i> TIGR4 <i>S. pyogenes</i> M1 GAS <i>S. pyogenes</i> MGAS315 <i>S. pyogenes</i> MGAS8232	
<i>Rhizobiales</i>		
<i>Agrobacterium</i>	<i>A. tumefaciens</i> C58 (Cereon) circular <i>A. tumefaciens</i> C58 (U) circular	<i>C. crescentus</i> <i>R. prowazekii</i>
<i>Brucella</i>	<i>B. melitensis</i> <i>B. suis</i>	
<i>Mesorhizobium</i>	<i>M. loti</i>	
<i>Sinorhizobium</i>	<i>S. meliloti</i>	
<i>Actinomycetales</i>		
<i>Corynebacterium</i>	<i>C. efficiens</i> <i>C. glutamicum</i>	<i>B. longum</i>
<i>Mycobacterium</i>	<i>M. leprae</i> <i>M. tuberculosis</i> CDC1551 <i>M. tuberculosis</i> H37Rv <i>S. coelicolor</i>	
<i>Streptomyces</i>		
<i>Bacillaceae</i>		
<i>Bacillus</i>	<i>B. anthracis</i> <i>B. cereus</i> <i>B. halodurans</i> <i>B. subtilis</i>	<i>L. innocua</i> <i>L. monocytogenes</i>
<i>Oceanobacillus</i>	<i>O. iheyensis</i>	
<i>Staphylococcaceae</i>		
<i>Staphylococcus</i>	<i>S. aureus</i> N315 <i>S. aureus</i> MW2 <i>S. aureus</i> Mu50 <i>S. epidermidis</i>	<i>L. innocua</i> <i>L. monocytogenes</i>
<i>Campylobacterales</i>		
<i>Campylobacter</i>	<i>C. jejuni</i>	<i>C. crescentus</i>
<i>Helicobacter</i>	<i>H. hepaticus</i> <i>H. pylori</i> 26695 <i>H. pylori</i> J99	<i>X. fastidiosa</i> 9a5c

^a Generic names of the outgroups are given in table 2.

four genomes were known). The genome sequences were downloaded from TIGR (<http://www.tigr.org/>). If one of the genomes was annotated, we made use of this annotation to extract all of the putative protein sequences from the genome. The similarity of these proteins to the proteins in the other genomes was measured via the BLASTP algorithm (Altschul et al. 1997). All potential matches between genes were required to be hits with reciprocal expect values less than 10^{-20} . In addition, they were required to have a match that extended over more than 85% of the length of the query protein. Different thresholds for an expect value can be expected to cause some minor variation in the identification of potential homologs, but a comparatively stringent level was chosen for this study (Bansal and Meyer 2002). Protein sequences that satisfied these criteria were deemed to be in the same gene family. If the number of genes in the gene family changed within the group of organisms, the extra copies were assumed to be gene duplications. It is of course quite possible that they are insertions of an extraneous gene that has a close similarity to an existing gene, but the assumption of duplication rather than transfer is more conservative.

It should be noted that the method being used here relies on sequence similarity. So orthologous gene replacement (Koonin, Mushegian and Bork 1996) or unique genes will not be detected using this method. In addition, lineage-specific fast evolving genes (Pesole et al. 1999; Koonin et al. 2004) will be treated as unique genes and will not be considered.

If not all of the genomes were annotated (as is the case for example with *Pseudomonas fluorescens*), the proteins of an annotated genome were used to query the nucleotide sequence of the unannotated genome. This was done using the TBLASTN algorithm; querying the presence of known (annotated) proteins within the translated nucleotide sequence of the genome. The largest annotated genome was used for the TBLASTN algorithm to ensure that the largest number of potential genes would be revealed. All nucleotide sequences that had significant matches (according to the criteria above) to the annotated protein sequences were extracted and translated. To ensure that any other potential coding sequences were identified, the nucleotide sequences between significant matches to the annotated proteins (i.e., the gaps) were extracted if they were greater than 100 nucleotides long. These nucleotide sequences were then translated in all six frames and searched for long open reading frames (greater than 30 amino acids).

All translated proteins that had significant matches with the annotated genome were extracted. All the sequences which had matches with select representatives from the bacterial domain of life (table 2) were also extracted. Based on these searches a database of genes for each species was established. If they had a significant hit to each other, all the matches were referred to as being in the same gene family. The best BLAST hit is not always the best indicator for the nearest phylogenetic neighbor or even as an ortholog (Koski and Golding 2001), and so to avoid the effect of paralogs, we clustered all potential homologs in the same genome as a gene family. As above,

Table 2
Representative Bacterial Genomes

<i>Agrobacterium tumefaciens</i> C58 (U. W) chr circular ^a	<i>Bifidobacterium longum</i>
<i>Buchnera aphidicola</i> Sg	<i>Caulobacter crescentus</i>
<i>Clostridium acetobutylicum</i>	<i>Clostridium perfringens</i>
<i>Enterococcus faecalis</i>	<i>Escherichia coli</i> K12-MG1655
<i>Escherichia coli</i> O157:H7 EDL933	<i>Escherichia coli</i> O157:H7 VT2-Sakai
<i>Fusobacterium nucleatum</i>	<i>Haemophilus influenzae</i>
<i>Helicobacter pylori</i> 26695	<i>Lactococcus lactis</i>
<i>Leptospira interrogans</i>	<i>Listeria innocua</i>
<i>Listeria monocytogenes</i>	<i>Mycobacterium tuberculosis</i> CDC1551
<i>Neisseria meningitidis</i>	<i>Pasteurella multocida</i>
<i>Rickettsia prowazekii</i>	<i>Salmonella enterica</i> serovar Typhi CT18
<i>Synechocystis</i> sp	<i>Thermotoga maritima</i>
<i>Vibrio cholerae</i> (chr I)	<i>Xylella fastidiosa</i> 9a5c

^a Strains are indicated only where more than one was available.

differing numbers of paralogs in closely related species are attributed to duplication rather than to LGT. The “single link” method of Friedman and Hughes (2003) was used to define the protein families. This groups genes that share similarity to any other single member. For example, if genes A and B are in a family, and genes B and C are in another family, then A, B, and C are in a family. Again if there are differing numbers of genes in a family, this is taken as evidence for gene duplication.

With this procedure, only genes known to exist (or at least annotated) in another organism are detectable. For example, all genes that we have indicated as present in *Pseudomonas fluorescens* have a significant similarity with a gene from either *P. aeruginosa*, *P. putida*, *P. syringae*, or with species from the representative group of bacteria. Any potentially novel genes in *P. fluorescens* which do not have any homolog within other *Pseudomonas* or representative genomes are undetectable by this method and will be missing from our results.

Gene insertions/deletions were mapped onto the phylogenies according to a parsimony principle. When there was ambiguity, a set of one or two outgroup genomes were chosen to help determine whether the genes near the base of the tree were inserted or deleted. If the homologs of genes are present in the outgroup genomes, the genes are referred to as deletions, and if not, they are noted as insertions.

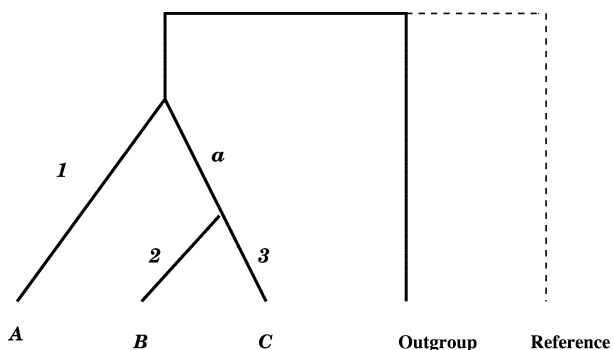


FIG. 1.—An example phylogeny to demonstrate the method to infer gene insertions, deletions, and duplications. All the possibilities are shown in table 3. Throughout figures 1–10, external branches are numbered and internal branches are labeled alphabetically.

The species analyzed and their outgroups are shown in table 1. They cover a broad spectrum of bacterial diversity and have several different levels of within-group divergences. There are nine groups of taxa with a total of 50 species analyzed. The groups analyzed included species from the *Chlamydiaceae* group and the *Mycoplasmataceae*, *Streptococcaceae*, *Rhizobiales*, *Actinomycetales*, *Bacillaceae*, *Staphylococcaceae*, and *Campylobacterales* groups.

The 16s rRNA sequence from each species was extracted from the genomes or from the Ribosomal Database Project at <http://rdp.cme.msu.edu/html/download>.

Table 3
The Method Used to Infer Gene Number Changes Illustrated with the Phylogeny Shown in figure 1

Gene in Species A	Gene in Species B	Gene in Species C	Gene in Outgroup	Gene in Representatives	Conclusion
+	–	–	–	–	∅
–	+	–	–	–	∅
–	–	+	–	–	∅
+	–	–	+	–	del a
+	–	–	–	+	ins 1
+	–	–	+	+	del a
–	+	–	+	–	ins 2
–	+	–	–	+	ins 2
–	+	–	+	+	ins 2
–	–	+	+	–	ins 3
–	–	+	–	+	ins 3
–	–	+	+	+	ins 3
+	+	–	–	–	del 3
+	+	–	+	–	del 3
+	+	–	–	+	del 3
+	+	–	+	+	del 3
+	–	+	–	–	del 2
+	–	+	+	–	del 2
+	–	+	–	+	del 2
+	–	+	+	+	del 2
–	+	+	–	–	ins a
–	+	+	+	–	del 1
–	+	+	–	+	ins a
–	+	+	+	+	del 1
+	+	+	–	–	∅
+	+	+	+	–	∅
+	+	+	–	+	∅
++	+	+	+	+	dup 1
+	++	+	+	+	dup 2
+	+	++	+	+	dup 3
+	++	++	+	+	dup a

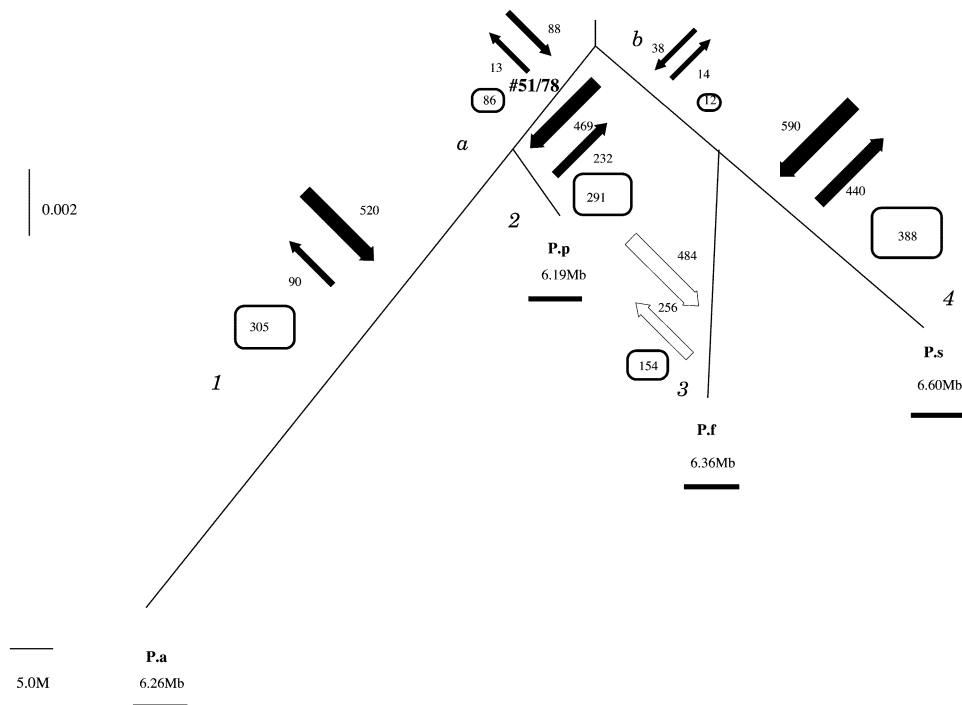


FIG. 2.—The presence of genes within *Pseudomonas* species. Arrows pointed toward or away from a branch indicate the estimated number of insertions or deletions, respectively. Rectangles mirror the inferred number of gene duplications. Branch lengths (with vertical scale bar) indicate phylogenetic distance based on rRNA. Horizontal scale bar indicates genome size. External branches are numbered and internal branches are labeled with lowercase letters. The bootstraps from Puzzle (top) and support from Mr.Bayes (bottom) are shown, if a branch was not supported by values of at least 80% by both methods. Open arrows indicate that the genome was not annotated at the time of this study.

html. These sequences were aligned using the Clustalw algorithm (Thompson, Higgins, and Gibson 1994) and phylogenetic trees were generated using Puzzle (Strimmer and von Haeseler 1996; 1,000 puzzling steps with the HKY substitution model) and using Mr.Bayes (Huelsenbeck and Ronquist 2001; 10,000 generations with the gamma distribution model). An example of the method used is shown in figure 1 and table 3.

Because parsimony involves an inference about the insertions/deletions, we used simulations to further explore the process of insertion/deletion. These simulations were conducted to gain an expectation of the patterns of gene presence/absence among the observed species without further inference. The simulations assume that there is a constant and equal insertion and deletion rate. These events affect only a single gene in each case and are appropriate to measure the number of genes inserted/deleted but are not a measure of the number of events expected in nature. All genes are assumed to have been potentially inserted or deleted. In the simulations the rate of insertion/deletion was chosen to give the same number of genes present in all species as that actually observed. The differences between simulation and observed results are listed in Appendix tables A.1 to A.3. For illustration purposes, the average number of genes in each genome was normalized to 1,000.

Results

The number of genes inferred to have been inserted, deleted, or duplicated is shown in figures 2 to 10. The

arrows directed toward a branch indicate insertions, arrows directed away from a branch indicate deletions, and numbers within a rectangular box indicate the inferred number of duplications. The length and width of the arrows are proportional to the number of insertions or deletions within each branch. Similarly, the length and width of the rectangular boxes are proportional to the number of duplications within each branch. The scale bar below each species name indicates the overall genome size.

In general there is no obvious correlation between the number of ins/del/dup versus branch length. In the *Pseudomonas* group, the correlation coefficients between ins/del/dup and branch length are 0.49, 0.00 and 0.36, respectively (table 4). Other R^2 values for the *Pseudomonadaceae* are 0.82 between insertions and duplications, 0.65 between insertions and deletions, 0.57 between deletions and duplications. However, these values are quite variable among different groups (data not shown). For example, *P. putida* KT 2440 has evolved comparatively recently within the *Pseudomonas* group. If the incidence of indels were proportional to evolutionary branch length, then the number of genes inserted into or deleted from the external branch leading to *P. putida* should be a fraction (0.14) as much as those in the branch leading to *P. aeruginosa*. Instead they are 0.90 and 2.58 times larger for insertions and deletions, respectively. In most of the groups, the largest insertion or deletion numbers were not on the longest branch. This is true for the number of duplications as well. For example, within *Mycoplasma* species (fig. 4), the external branch (branch

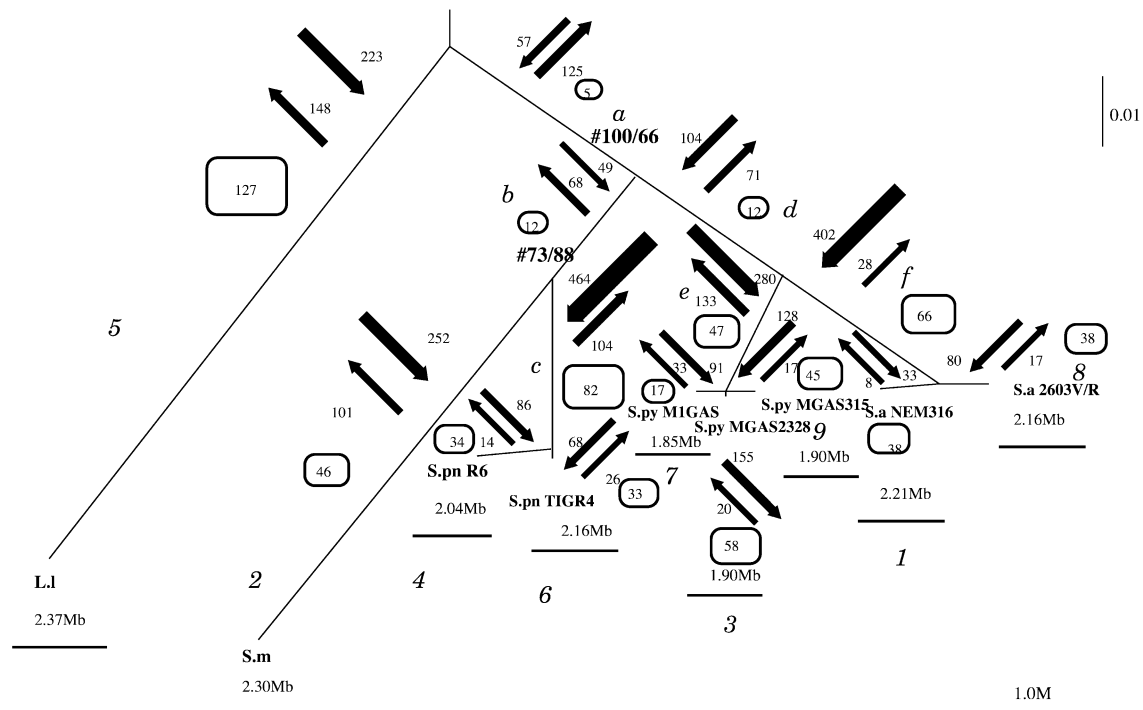


FIG. 5.—The presence of genes within the *Streptococcaceae* group (table 1). Symbols as in figure 2. Taxa 3, 7, and 9 form a multifurcation according to Puzzle (shown) but branch ((7,9),3) via Mr.Bayes.

(707) and duplications (981) in the external branch leading to *S. coelicolor* (branch 1). There are only 52 genes inferred to have been deleted during this branch. To gain insight into the function of these mobile genes a further analysis of the two largest genomes, *M. loti* and *S. coelicolor*, was done. Of the genes inserted into the external branch leading to *S. coelicolor* (branch 1 in fig. 9), there are 414 new genes involved in biosynthesis function (table 5). Some of these might be important in mycelium formation (Redenbach, Scheel, and Schmidt 2000). Some of them may be essential for antibiotic production. In *M. loti*, 9.61% of the inserted proteins are transport related, and some of them are essential for adhesion to their host plant (Hacker and Kaper 2000).

Two *Brucella* species were analyzed and each has two chromosomes. Based on the parsimony method, it is likely that the two chromosomes in each *Brucella* split before the two *Brucella* speciated. There are fewer inferred gene events if the two chromosomes split earlier than the two species diverged (data not shown). At present, there are seven genomes with one large and one smaller chromosome available. They are *Vibrio vulnificus*, *V. parahaemolyticus* RIMD 2210633, *V. cholerae*, *Leptospira interrogans*, *Deinococcus radiodurans* R1, *Brucella suis*, and *B. melitensis*. Usually ribosomal RNAs are present only on the large chromosome, but *rrn* operons were found on both the large and small chromosomes in *V. parahaemolyticus* (Tagomori, Iida, and Honda 2002) and *B. suis* 1330 (Paulsen et al. 2002). It is possible that the *rrn* operon in the small chromosome of *B. suis* was duplicated from the large one.

From the diagrams in figures 2 to 10, it is apparent that there are more gene insertions in external branches

than the internal branches. And there are fewer gene deletions/duplications in the internal branches. This suggests that most of the genes inserted in one species are genes unique to that group and are likely to be deleted before a new species evolves.

There are two paraphyletic groups found with the phylogenetic analysis. *Ureaplasma uralyticum* branches within the *Mycoplasma* genus (fig. 4), and *Oceanobacillus iheyensis* branches within *Bacillus* (fig. 10). *Ureaplasma* and *Mycoplasma* are members of the class *Mollicutes*. Based on the reclassification of *Mollicutes* (Weisburg et al. 1989), *M. genitalium*, *M. pneumoniae*, and *U. uralyticum* belong to the *Pneumoniae* group, and *M. pulmonis* belongs to the *Hominis* group. That *Ureaplasma* is within one of the *Mycoplasma* clusters was shown by Yoshida et al. (2002). In the *Bacillus* group, *B. halodurans* is distantly related to the other three *Bacillus* compared with *Oceanobacillus iheyensis* (Lu, Nogi, and Takam 2001).

Within the *Bacillaceae* genus, *B. cereus*, *B. anthracis*, and *B. thuringiensis* have been assigned as varieties of the same species (Helgason et al. 2000) and alternatively as different species (Ivanova et al. 2003). In this study, we found that about 200 unique genes separate *B. cereus* and *B. anthracis*. This large difference in genes perhaps leads to their remarkably different phenotypes, even though the strains may be closely related.

Simulations were conducted to determine an expectation for patterns of gene presence/absence in the observed phylogenies. In these simulations, 1,000 genes were modeled with a constant and equal rate of insertion and deletion. Each insertion was considered to be a unique gene. The phylogeny branch lengths as estimated

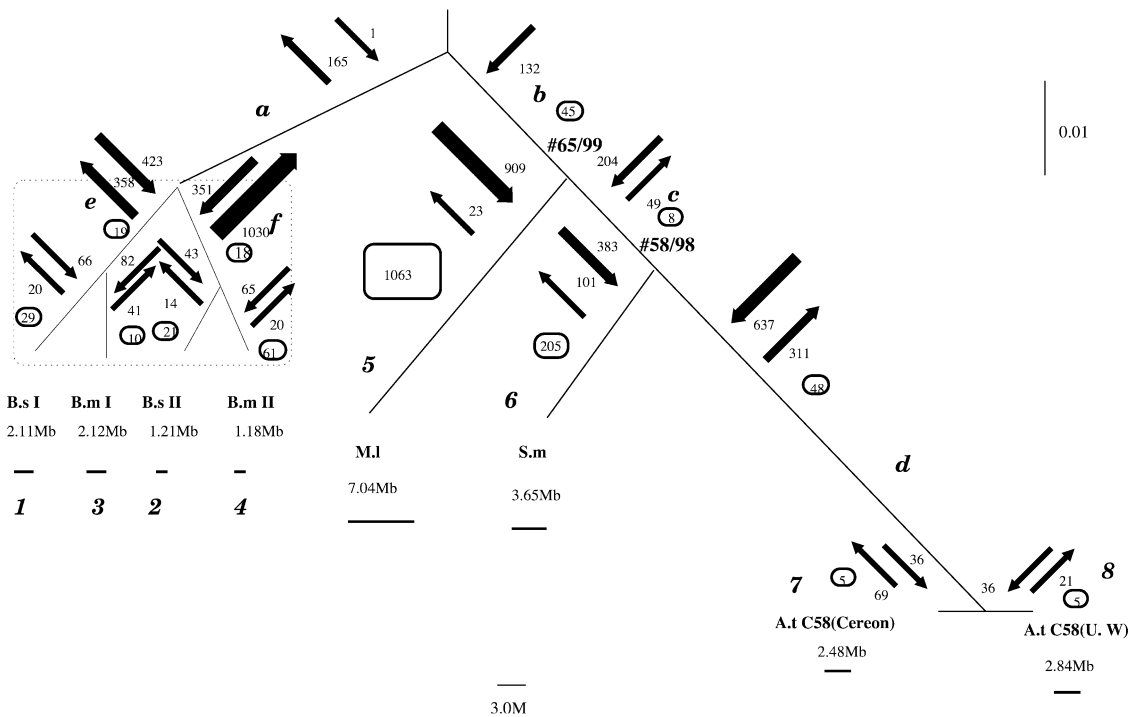


FIG. 6.—The presence of genes within the *Rhizobiales* group (table 1). Symbols as in figure 2. In the dashed box, the branch lengths are not shown to scale. The branch lengths leading to *Brucella suis* and *B. melitensis* are 0.00001 each. Both species have two chromosomes.

from the rRNA genes were used in the simulation. The overall rate of gene ins/del was adjusted to give the observed numbers of genes that are conserved across all species. The simulations confirm the above results (Appendix table A.1–A.3). They show that while there is a bias such that insertions are expected to exceed deletions,

the observed data deviate beyond what is expected. The insertions exceed deletions and external branches have an excess of inserted/deleted genes. The identity of the inserted genes in each branch in the phylogeny is given as supplementary information at <http://life.biology.mcmaster.ca/~weilong/research.html>.

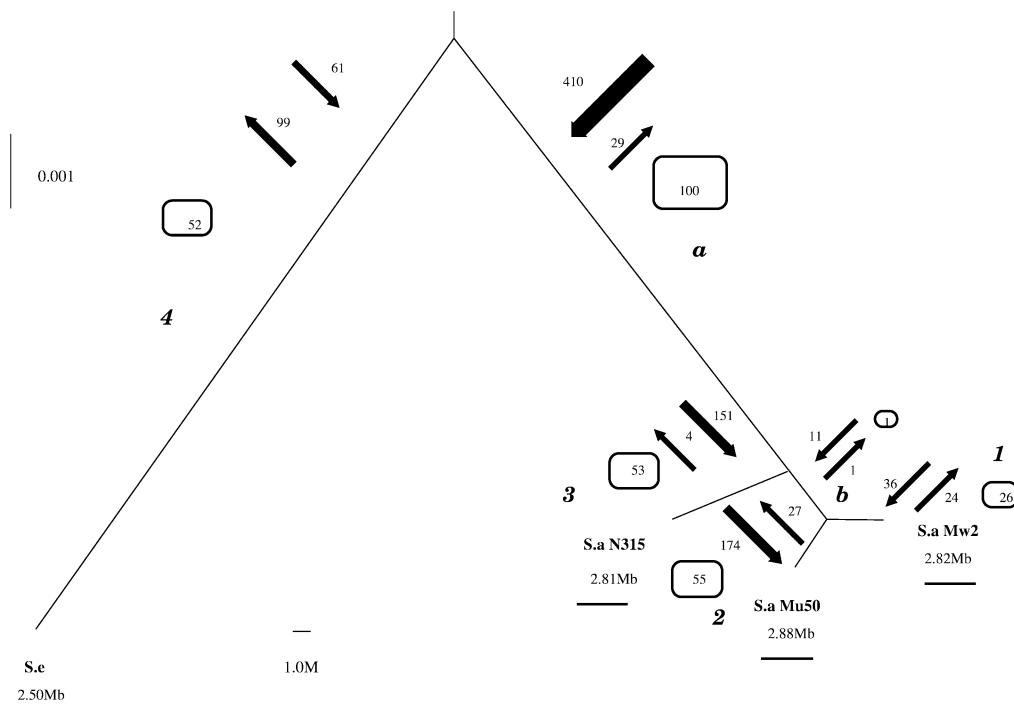


FIG. 7.—The presence of genes within the *Staphylococcus* group (table 1). Symbols as in figure 2.

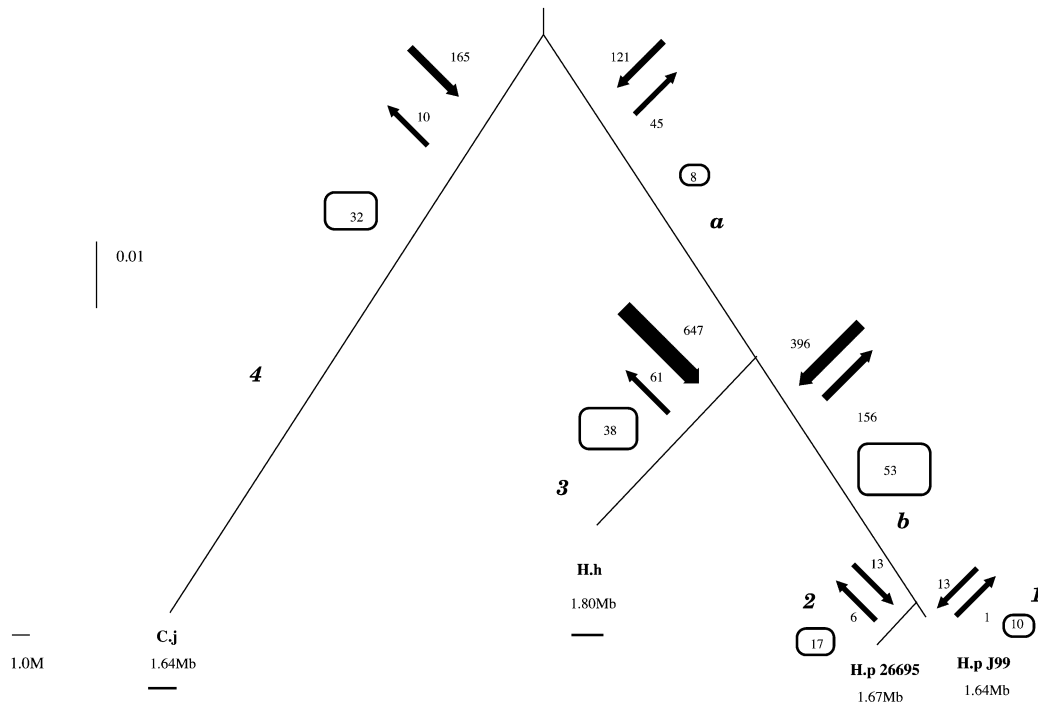


FIG. 8.—The presence of genes within the *Campylobacteriales* group (table 1). Symbols as in figure 2.

Discussion

The revelation that many genes are laterally transferred in bacteria has changed our concept of bacterial genome evolution (for a review, see Doolittle 2000). But with the exception of viral genomes (e.g., see McLysaght, Baldi, and Gaut 2003), there have still been comparatively

few analyses of entire genomes (but see Martin et al. 2002; Daubin, Laurat, and Perriere 2003; Daubin, Moran, and Ochman 2003). In general a complete analysis of genomic changes in related species is quite difficult (notwithstanding several excellent studies of *E. coli* strains; e.g., Perna et al. 2001). One major reason for this is that most sequencing projects do not examine very closely related genomes. Even

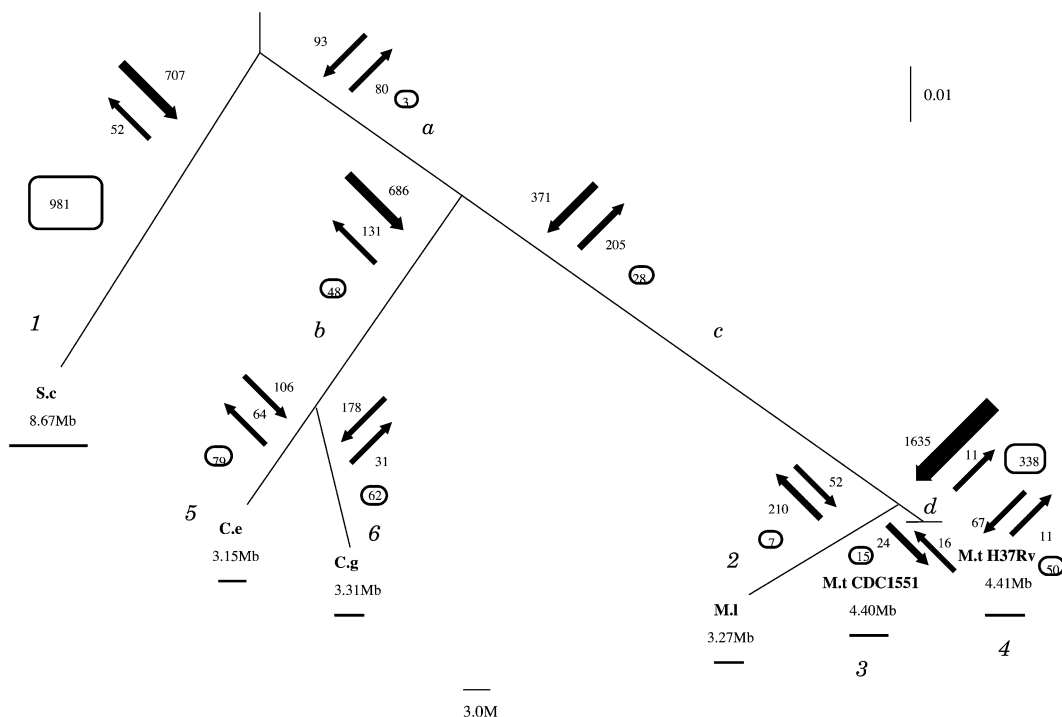


FIG. 9.—The presence of genes within the *Actinomycetales* group (table 1). Symbols as in figure 2.

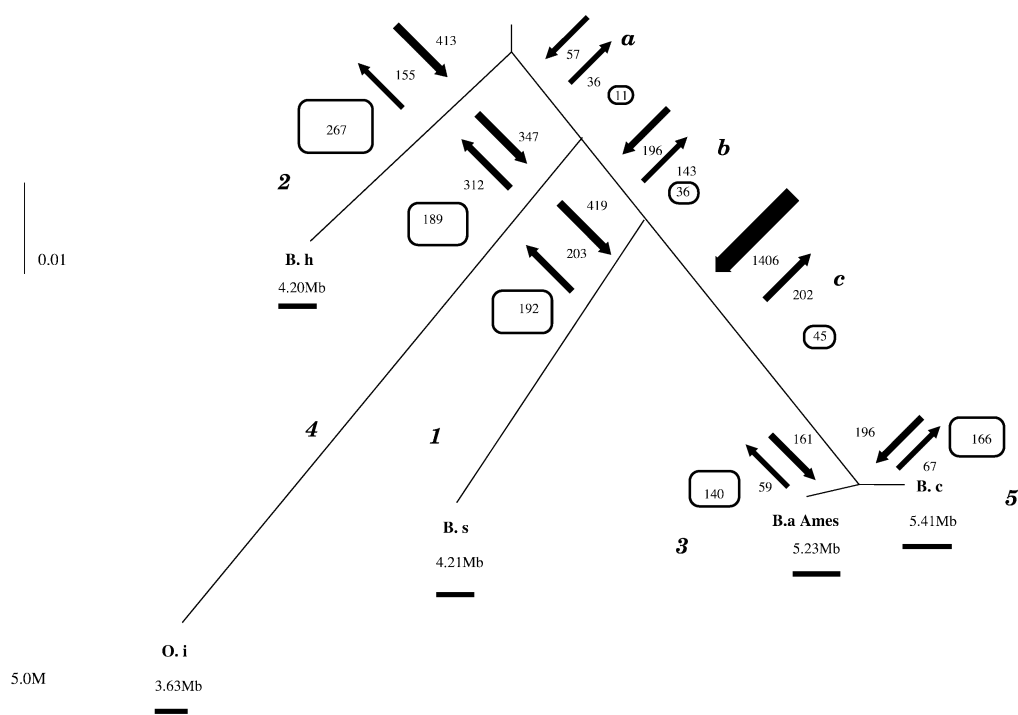


FIG. 10.—The presence of genes within the *Bacillaceae* group (table 1). Symbols as in figure 2.

for those projects that have determined the genome from congeneric species, the divergence of the species can still be quite extensive. For example, the four *Pseudomonas* species examined here are sufficiently diverged to have effectively masked most of their genome rearrangements. In addition, when given good data, the correct inference of genome rearrangements is computationally difficult (Sankoff 2001; Nadeau and Sankoff 1998). Finally, the inference of lateral transfer can itself be difficult. Therefore, we have taken a simpler approach here and have asked only if the same gene exists in related bacteria. The presence or absence data for each gene have then been mapped onto the phylogeny of the species (as inferred from rRNA sequences).

By the “same gene” we mean a gene that shares significant similarity but not necessarily an actual homolog. If the correct homolog were replaced via a lateral transfer with a similar gene, the methods employed here would fail to detect these transfers. If a gene belonged to a family of similar genes, transfers would not be detected, and only a duplication or deletion would be inferred.

A gene is inferred to exist only if a similar gene already exists in some other bacterial species. The standards used to identify a similar gene are comparatively stringent, so some homologs may have been missed because of extensive sequence divergence. In addition, genes that might have been laterally transferred might have distant similarity to existing genes and these would be classified as duplicated family members. Some false positives and a larger number of false negatives are to be expected (Geneux and Logsdon 2003). Nevertheless, many lateral transfers will be missed by these methods and hence these results should be considered conservative.

This method is also biased to ignore any unique LT genes and to underestimate changes at terminal branches. For example, any gene uniquely present on any external branch and not shared among the group or within the outgroups will be ignored. The parsimony method used to infer the branch upon which an insertion/deletion occurred is also known to be biased; it is well known to underestimate the number of events (Galtier and Boursot 2000; Dean et al. 2002; Felsenstein 2004). Taken together, these biases will cause the number of genes inserted/deleted/duplicated at the termini of the tree to be underestimated. Despite these biases, however, the majority of gene insertions occur on external branches. Because genome sizes are not radically changed, this must imply that the majority of gene insertions tend to occur uniquely within the external branches and will be deleted (or replaced) before new species diverge. This supports the hypothesis that these lateral transfers help the bacteria to occupy a species specific niche.

Table 4
The ins/del/dup Genes Are Not Correlated with the Evolutionary History Within *Pseudomonas* Species

Branch	Length	Insertions	Deletions	Duplications
1	0.03465	520	90	305
2	0.00501	469	232	291
3	0.01878	484	256	154
4	0.01348	590	440	388
a	0.00784	88	13	86
b	0.00783	38	14	12
Correlation coefficient		0.49	0.00	0.36

Table 5
The function of inserted genes in *Mesorhizobium loti* MAFF303099 and *Streptomyces coelicolor* A3(2)

	Unknown Function	Transport Related	Transcriptional Regulator	Biosynthesis Enzyme	Other
<i>M. loti</i>	429(40.02%)	103(9.61%)	73(6.81%)	338(31.53%)	129(12.03%)
<i>S. coelicolor</i>	207(21.10%)	99(10.09%)	89(9.07%)	414(42.20%)	172(17.53%)

From the diagrams (figs. 2–10), it is plain that there are generally more inserted genes than deleted genes. This result is consistent with previous research which reported that there are more insertions than deletions (Rode et al. 1999; Ochman and Jones 2000). All other factors kept equal, in bacteria a small genome has an evolutionary advantage (Moran 2002) and hence, the fate of nonfunctional sequence is to be eliminated especially in obligate parasites. But in this study, even in the obligate parasite *Mycoplasma*, there are more insertions than deletions.

A recurrent theme in figures 2–10 is the lack of correlation of insertions and length of evolutionary history. As measured by rRNA substitutions, *Streptococcus agalactiae* strain NEM316 has had a long evolutionary history compared with strain 2603V/R. But NEM316 has fewer inserted genes than 2603V/R. In addition, there are more insertions than deletions in most branches. In some species, for example *Buchnera aphidicola*, there is a stable genome with a conserved gene order (van Ham et al. 2003) and there are only a few deletions from the branches of this group (Silva, Latorre, and Moya 2003). Again, gene rearrangement in the genome at this level of resolution is not dependent on the length of evolutionary history.

The results of Mirkin et al. (2003) suggest equal numbers of insertions and deletions, whereas Daubin, Lerat, and Perriere (2003) and the results presented here suggest more insertions than deletions. The main difference between these studies is the distance between the species examined. The phylogeny used by Mirkin et al. (2003) covers the complete scope of life—bacteria, archaea, and eukaryota. Their main purpose was to determine the subset of genes present in the last common ancestor of all life. Here genomes were specifically selected from related groups. Although the genes of bacterial genomes are in a constant state of flux (Snel, Bork, and Huynen 2002), not all insertion/deletion events are easily detected. With large divergences between species, this inference becomes more and more difficult, but the inference of the gene set in the last common ancestor requires this level of divergence. Consistent with this, we have detected a larger number of ins/del events than did Mirkin et al. (2003), and we are probably still missing a large number of ins/del events because the species chosen are still too distantly related. As more genomes become available in the future a more accurate estimate will become possible.

Nor are duplicated genes correlated with the length of evolutionary history or consistent across taxa. In these species, there is no evidence that repeated events of genome doubling are a major force for evolution (Riley and Anilionis 1978; Mira, Ochman, and Moran 2001).

Although it was reported that there is an almost linear relationship between divergence of sequence and gene order degradation in closely related bacterial genomes (Suyama and Bork 2001), our study looks at a deeper phylogenetic level where there are no longer remnants of any correlation between ins/del/dup and evolutionary branch length. Sometimes the largest ins/del/dup number occurred along a short branch. A similar result was found in the evolutionary history of viral genomes with independent duplication and deletion events in lineages of viruses (Herniou et al. 2001; Hughes and Friedman 2003). The correlation coefficients between duplications and insertions are variable. The R^2 values are 0.02 of *Bacillaceae*, 0.24 of *Actinomycetales*, 0.39 of *Mycoplasmataceae*, 0.40 of *Streptococcaceae*, 0.56 of *Campylobacteriales*, 0.66 of *Chlamydiaceae*, 0.82 of *Pseudomonadaceae*, and 0.85 of *Staphylococcaceae*. Some duplications have been shown to be due to LGT (Gogarten, Doolittle, and Lawrence 2002; Snel, Bork, and Huynen 2002). Duplications are more common among laterally transferred genes than among native genes (Hooper and Berg 2003). Therefore, many of the genes inferred to have been duplicated may also have been laterally transferred. These would inflate the number of insertions beyond what is inferred here.

The rRNA genes have been used here to provide a measure of the evolutionary history. Daubin, Moran, and Ochman (2003) suggest that LGT has not eliminated the phylogenetic signals. Nevertheless, it is possible that this reflection of the species history given by the rRNA gene sequence is inaccurate but it is still apparent that no other evolutionary history could eliminate the patterns observed. While the numbers of indels are not correlated with the rRNA evolutionary history, nor are they strongly correlated with each other. Some evolutionary branches have high numbers of inserted genes and yet low numbers of deleted genes.

Lateral gene transfer can occur via acquisition and uptake from the environment, so genes involved in DNA uptake and recombination play an important role in LGT (Silva, et al. 2003). It is well known that the loss of many genes that encode proteins associated with recombination and DNA uptake maintains the stability of *B. aphidicola* genomes (Silva, Latorre, and Moya 2003). The functional loss or enhancement of genes involved in recombination and DNA uptake is more likely based on selection or global deletion-insertion biases (Petrov 2001). Many other forces might affect the number of deletions/insertions, but these are not clear. Which force is important and how selection works to enhance or curtail gene insertions and deletions will be very important in future studies.

Table A.1
The Difference Between Observed Patterns and Simulated Patterns

Patterns ^a	<i>Campylobacterales</i>		<i>Staphylococcaceae</i>		<i>Pseudomonadaceae</i>	
	Observed	Simulated	Observed	Simulated	Observed	Simulated
0001	7	12	11	0	92	80
0010	8	24	15	12	104	104
0011	323	145	6	15	17	73
0100	524	128	1	15	69	43
0101	2	7	4	2	97	13
0110	2	2	75	0	7	9
0111	106	294	245	223	17	132
1000	162	378	34	235	93	178
1001	2	1	7	0	28	1
1010	0	0	0	0	68	1
1011	47	43	0	9	62	19
1100	112	67	0	8	34	76
1101	4	10	8	9	59	69
1110	1	5	11	0	95	48
1111	486	487	741	741	611	610
GN ^b	1,226		1,943		2,859	

^a In this column a '1' indicates a gene present in the species and a '0' indicates absence. The order of the numbers is matched with the species order from left to right in each phylogeny in figures 2 to 10.

^b Observed values are normalized to 1,000 genes to facilitate comparison between groups. To recover the actual number observed in each pattern, multiply by the average number of genes analyzed for each group (GN).

Our results suggest that an amazingly large number of genes have been laterally transferred even within comparatively closely related bacteria. Lateral transfers are thought to be influenced more by physical proximity than by phylogenetic proximity of the organisms (Matte-Tailliez et al. 2002) and are seen to share similar genomic properties such as genome size, genome G/C composition, and carbon utilization (Jain et al. 2003). The functions of most laterally transferred genes are still unknown. The most studied lateral transfers belong to pathogenic islands, but pathogenicity islands are only a small part of genes laterally transferred. The results presented here suggest that many of the LT genes may be species-specific adaptations.

Acknowledgments

This work was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) grant to G.B.G. The authors wish to thank Prof. R.A. Morton, the editor, and two anonymous reviewers for their helpful comments on previous drafts.

Appendix: Simulations of Gene Patterns

The patterns of gene presence for each species are listed in tables A.1 to A.3. In these tables, a '1' indicates a gene present in the species and a '0' indicates absence. The order of these numbers is matched to the species order from left to right within each phylogeny. Simulations were conducted with the assumption of constant and equal insertion/deletion rates.

The expected number of each gene pattern can be calculated easily if one assumes a constant and equal rate of gene insertion/deletion. Deleted genes cannot be observed, but if one conceptually retains a placeholder for a gene then the number of empty placeholders can be counted. If the rate of gene insertion is ν and the rate of gene deletion is μ (both events involving only a single

gene and all events independent), then the equilibrium frequency for a gene to be present is $\nu/(\mu + \nu)$. As evolution proceeds, genes will be deleted or inserted at these rates, and the probability at time t that any one placeholder will be occupied by a gene (p^t) or not can be found as follows,

$$p^t = \nu/(\mu + \nu) + (1 - \nu - \mu)^t [p^0 - \nu/(\mu + \nu)]$$

and if $\mu = \nu$ then

$$p^t = \frac{1}{2} + (1 - 2\mu)^t \left[p^0 - \frac{1}{2} \right].$$

For two taxa separated by t^1 and t^2 generations and starting initially with all gene placeholders occupied ($p^0 = 1$), the gene patterns will be observed according to

11	$p^{t^1} \times p^{t^2}$
10	$p^{t^1} \times (1 - p^{t^2})$
01	$(1 - p^{t^1}) \times p^{t^2}$
00	$(1 - p^{t^1}) \times (1 - p^{t^2})$

More complicated patterns and phylogenies can be calculated using the appropriate initial conditions for each ancestral node.

More realistically, placeholders cannot be observed and when an insertion occurs in any branch it would not have any homology to any other gene present in the group (otherwise it would be considered a deletion). Any insertion is therefore treated as a new, unique gene. This assumption complicates the above results, and a simulation, which incorporates this, was done to compare observed and expected gene patterns of presence/absence (tables A.1 and A.2). In the simulation study, the simulation process was started with a total of 1,000 genes, and this number was kept relatively constant due to an equal rate of deletions and insertions. Genes were picked at random for deletion. When an insertion occurred in any branch, it

Table A.2
The Difference Between Observed Patterns and Simulated Patterns

Patterns ^a	<i>Chlamydiaceae</i>		<i>Mycoplasmataceae</i>		<i>Actinomycetales</i>		<i>Bacillaceae</i>	
	Observed	Simulated	Observed	Simulated	Observed	Simulated	Observed	Simulated
000001	2	64	8	2	10	0	56	2
000010	2	11	21	1	6	0	44	7
000011	77	67	300	92	699	51	491	167
000100	7	45	384	157	3	43	100	185
000101	5	1	5	1	14	0	10	0
000110	6	3	3	1	2	0	8	0
000111	14	70	63	71	180	302	78	71
001000	0	0	234	181	56	146	76	283
001001	0	0	0	0	0	0	6	0
001010	0	0	5	0	0	0	5	0
001011	0	0	26	3	6	1	44	16
001100	2	0	53	3	0	0	26	15
001101	0	0	0	0	0	0	3	1
001110	0	0	5	0	0	0	2	0
001111	1	0	39	27	4	13	45	80
010000	1	3	55	183	38	121	94	195
010001	0	0	0	0	0	0	7	0
010010	0	0	0	0	0	0	7	0
010011	0	0	8	2	3	2	40	31
010100	0	0	8	3	0	0	36	26
010101	0	0	0	1	0	0	3	2
010110	0	0	0	0	0	0	4	0
010111	0	0	13	20	1	21	69	145
011000	1	0	45	51	316	177	43	48
011001	0	0	0	1	1	0	1	1
011010	0	0	0	0	0	0	4	0
011011	0	0	5	15	23	14	39	82
011100	0	0	18	16	3	2	49	71
011101	0	0	0	3	1	0	8	4
011110	0	0	5	2	1	0	10	0
011111	1	1	84	172	66	145	396 ^c	395
100000	1	1	153	315	248	428	— ^c	—
100001	0	0	0	0	5	0	—	—
100010	0	0	18	0	1	0	—	—
100011	0	0	8	5	90	12	—	—
100100	0	0	11	7	2	3	—	—
100101	0	0	0	1	1	0	—	—
100110	0	0	0	1	0	0	—	—
100111	0	0	5	61	49	133	—	—
101000	1	0	13	5	22	10	—	—
101001	0	0	0	0	0	0	—	—
101010	0	0	8	0	0	0	—	—
101011	0	0	8	5	5	2	—	—
101100	0	0	5	7	0	0	—	—
101101	0	0	0	2	0	0	—	—
101110	0	0	5	1	0	0	—	—
101111	5	3	76	57	13	23	—	—
110000	24	0	21	4	9	11	—	—
110001	0	0	0	0	0	0	—	—
110010	0	0	0	0	0	0	—	—
110011	0	0	8	4	4	2	—	—
110100	1	0	11	5	0	1	—	—
110101	0	0	0	1	0	0	—	—
110110	0	0	0	1	0	0	—	—
110111	1	0	29	46	1	36	—	—
111000	222	86	18	30	63	85	—	—
111001	5	0	0	1	1	0	—	—
111010	0	1	0	0	2	0	—	—
111011	9	31	18	29	52	19	—	—
111100	51	62	13	36	4	4	—	—
111101	8	7	5	8	2	0	—	—
111110	8	30	5	3	1	0	—	—
111111	776	777	361	362	226	225	—	—
GN ^b	870		380		2,004		2,511	

^a Patterns in this column indicate gene presence/absence as described in table A.1.

^b Observed values are normalized to 1,000 genes to facilitate comparison between groups. To recover the actual number observed in each pattern, multiply by the average number of genes analyzed for each group (GN).

^c There are only five *Bacillaceae* species. To make the table concise, the first letter of the pattern was set to 0.

Table A.3
The Comparison of Inferred ins/del Between Observed and Simulated Results for the *Pseudomonas* Species

	Branch	1	2	3	4	a	b
Observed	Insertions	182	164	169	206	31	13
	Deletions	26	69	69	100	4	3
Simulation	Insertions	179	55	110	87	37	36
	Deletions	142	20	77	53	38	38

became an instance of a new gene. The phylogenies given in figures 2 to 10 were used. The overall rate of insertion/deletion was chosen to give the number of genes that are observed to be conserved across all species (e.g., pattern 1111). Each simulation was repeated 10 times and the number for each presence/absence pattern counted. The average of these numbers is shown in tables A.1 and A.2.

To facilitate comparison between the different bacterial groups, the observed results have been normalized to a total of 1,000 genes. The average number of genes analyzed for each group is given in tables A.1 and A.2, and the actual number in each pattern can be found by multiplying the number of given in the table by the average and dividing by 1,000.

The patterns from the simulations and those observed are quite different. The number of events inferred on ancestral branches should generally be much larger. For example, in the *Pseudomonadaceae* the patterns '0011' and '1100' are strongly underrepresented in the observed data versus the simulated data (17 and 34 vs. 73 and 76, respectively). In figure 2 this translates to a deficiency of events on the ancestral branches *a* and *b*. Similarly, there should be more events on the ancestral branch *a* in figures 4 and 8 (or alternatively branch 4 depending on the state of the outgroup) and more deletions in branch *a* or more insertions in branch 4 in figure 7. Reinforcing this, the numbers of genes which have a dispersed and discontinuous distribution within the phylogeny are generally much larger in the observed results than in the simulations. This may be the result of lateral transfer occurring more readily within a group than to more taxonomically distant groups. For example, there are 97 genes (normalized to a total of 1,000) present in *P. putida*, *P. syringae* and absent from *P. aeruginosa*, *P. fluorescens*, but only 13 in the simulation results. Similar results hold for genes in *P. aeruginosa*, *P. syringae* that are absent from *P. putida*, *P. fluorescens*. Using the *Pseudomonas* group as an example, the inferred number of insertions/deletions was compared with the simulation results (table A.3; in this table deletions of duplicated genes were removed from consideration because the simulation did not include the possibility of duplicated genes). The simulation results indicate that even when insertion/deletion rates are equal, there is an observational bias toward more insertions than deletions. But the observed gene numbers deviate beyond this bias and even more strongly favor insertions over deletions.

Literature Cited

- Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389–3402.
- Bansal, A. K., and T. E. Meyer. 2002. Evolutionary analysis by whole-genome comparisons. *J. Bacteriol.* **184**:2260–2272.
- Daubin, V., E. Lerat, and G. Perriere. 2003. The source of laterally transferred genes in bacterial genomes. *Genome Biol.* **4**:R57.
- Daubin, V., N. A. Moran, and H. Ochman. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science* **301**:829–832.
- Dean, A. M., C. Neuhauser, E. Grenier, and G. B. Golding. 2002. The pattern of amino acid replacements in alpha/beta-barrels. *Mol. Biol. Evol.* **19**:1846–1864.
- Doolittle, W. F. 2000. Uprooting the tree of life. *Sci. Am.* **282**:90–95.
- Dutta, C., and A. Ran. 2002. Horizontal gene transfer and bacterial diversity. *J. Biosci.* **27**:27–33.
- Felsenstein, J. 2004. *Inferring phylogenies*. Sinauer Associates, Sunderland, Mass.
- Friedman, R., and A. L. Hughes. 2003. The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol. Biol. Evol.* **20**:154–161.
- Galtier, N., and P. Boursot. 2000. A new method for locating changes in a tree reveals distinct nucleotide polymorphism vs divergence patterns in mouse mitochondrial control region. *J. Mol. Evol.* **50**:224–231.
- Garcia-Vallvé, S., A. Romeu, and J. Palau. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.* **10**:1719–1725.
- Genereux, D. P., and J. M. Logsdon Jr. 2003. Much ado about bacteria-to-vertebrate lateral gene transfer. *Trends Genet.* **19**:191–195.
- Gogarten, J. P. 2003. Gene transfer: gene swapping craze reaches eukaryotes. *Curr. Biol.* **13**:R53–R54.
- Gogarten, J. P., W. F. Doolittle, and J. G. Lawrence. 2002. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19**:2226–2238.
- Hacker, J., L. Bender, M. Ott, J. Wingender, B. Lund, R. Marre, and W. Goebel. 1990. Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal *Escherichia coli* isolates. *Microbial Pathogenet.* **8**:213–225.
- Hacker, J., and J. B. Kaper. 2000. Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* **40**:169–189.
- Helgason, E., O. A. Økstad, D. A. Caugant, H. A. Johansen, A. Fouet, M. Mock, I. Hegna, and A.-B. Kolstø. 2000. *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*—one species on the basis of genetic evidence. *Appl. Environ. Microbiol.* **66**:2627–2630.
- Herniou, E. A., T. Luque, X. Chen, J. M. Vlak, D. Winstanley, J. S. Cory, and D. R. O'Reilly. 2001. Use of whole genome sequence data to infer baculovirus phylogeny. *J. Virol.* **75**:8117–8126.
- Hooper, S. D., and O. G. Berg. 2003. Duplication is more common among laterally transferred genes than among indigenous genes. *Genome Biol.* **4**:R48.
- Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**:754–755.
- Hughes, A. L., and R. Friedman. 2003. Genome-wide survey for genes horizontally transferred from cellular organisms to baculoviruses. *Mol. Biol. Evol.* **20**:979–987.
- Ivanova, N., A. Sorokin, I. Anderson et al. (23 co-authors) 2003. Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. *Nature* **423**:87–91.
- Jain, R., M. C. Rivera, J. E. Moore, and J. A. Lake. 2002. Horizontal gene transfer in microbial genome evolution. *Theor. Popul. Biol.* **61**:489–495.

- . 2003. Horizontal gene transfer accelerates genome innovation and evolution. *Mol. Biol. Evol.* **20**:1598–1602.
- Keeling, P. J., and W. F. Doolittle. 1997. Evidence that eukaryotic triosephosphate isomerase is of alpha-proteobacterial origin. *Proc. Natl. Acad. Sci. USA.* **94**:1270–1275.
- Klotz, M. G., and P. C. Loewen. 2003. The molecular evolution of catalytic hydroperoxidases: evidence for multiple lateral transfer of genes between prokaryota and from bacteria into eukaryota. *Mol. Biol. Evol.* **20**:1098–1112.
- Koonin, E. V., N. D. Fedorova, J. D. Jackson et al. (18 co-authors) 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**:R7.
- Koonin, E. V., A. R. Mushegian, and P. Bork. 1996. Non-orthologous gene displacement. *Trends Genet.* **12**:334–336.
- Koski, L. B., and G. B. Golding. 2001. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **52**:540–542.
- Koski, L. B., R. A. Morton, and G. B. Golding. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.* **18**:404–412.
- Lawrence, J. G., and H. Ochman. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA* **95**:9413–9417.
- Lu, J., Y. Nogi, and H. Takami. 2001. *Oceanobacillus iheyensis* gen. nov., sp. nov., a deep-sea extremely halotolerant and alkaliphilic species isolated from a depth of 1050 m on the Iheya Ridge. *FEMS Microbiol. Lett.* **205**:291–297.
- Martin, W., T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa, and D. Penny. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. USA* **99**:12246–12251.
- Matte-Tailliez, O., C. Brochier, P. Forterre, and H. Philippe. 2002. Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* **19**:631–639.
- McLysaght, A., P. F. Baldi, and B. S. Gaut. 2003. Extensive gene gain associated with adaptive evolution of poxviruses. *Proc. Natl. Acad. Sci. USA* **100**:15655–15660.
- Mira, A., H. Ochman, and N. A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**:589–596.
- Mirkin, B. G., T. I. Fenner, M. Y. Galperin, and E. V. Koonin. 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.
- Moran, N. A. 2002. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**:583–586.
- Nadeau, J. H., and D. Sankoff. 1998. Counting on comparative maps. *Trends Genet.* **14**:495–501.
- Ochman, H., and I. B. Jones. 2000. Evolutionary dynamics of full genome content in *Escherichia coli*. *EMBO J.* **19**:6637–6643.
- Ochman, H., F. C. Soncini, F. Solomon, and E. A. Groisman. 1996. Identification of a pathogenicity island required for *Salmonella* survival in host cells. *Proc. Natl. Acad. Sci. USA.* **93**:7800–7804.
- Paulsen, I. T., R. Seshadri, K. E. Nelson et al. (31 co-authors) 2002. The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. *Proc. Natl. Acad. Sci. USA.* **99**:13148–13153.
- Perna, N. T., G. Plunkett, V. Burland et al. (28 co-authors) 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**:529–533.
- Pesole, G., C. Gissi, A. De Chirico, and C. Saccone. 1999. Nucleotide substitution rate of mammalian mitochondrial genomes. *J. Mol. Evol.* **48**:427–434.
- Petrov, D. A. 2001. Evolution of genome size: new approaches to an old problem. *Trends Genet.* **17**:23–28.
- Redenbach, M., J. Scheel, and U. Schmidt. 2000. Chromosome topology and genome size of selected actinomycetes species. *Antonie Van Leeuwenhoek* **78**:227–235.
- Riley, M., and A. Anilionis. 1978. Evolution of the bacterial genome. *Annu. Rev. Microbiol.* **32**:519–560.
- Rode, C. K., L. J. Melkerson-Watson, A. T. Johnson, and C. A. Bloch. 1999. Type-specific contributions to chromosome size differences in *Escherichia coli*. *Infect. Immun.* **19**:230–236.
- Sankoff, D. 2001. Gene and genome duplication. *Curr. Opin. Genet. Dev.* **11**:681–684.
- Silva, F. J., A. Latorre, and A. Moya. 2003. Why are the genomes of endosymbiotic bacteria so stable? *Trends Genet.* **19**:176–180.
- Snel, B., P. Bork, and M. A. Huynen. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* **12**:17–25.
- Strimmer, K., and A. von Haeseler, 1996. Quartet puzzling: a quartet maximum-likelihood for reconstructing tree topologies. *Mol. Biol. Evol.* **13**:964–969.
- Suyama, M., and P. Bork. 2001. Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet.* **17**:10–13.
- Tagomori, K., T. Iida, and T. Honda. 2002. Comparison of genome structures of vibrios, bacteria possessing two chromosomes. *J. Bacteriol.* **184**:4351–4358.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- van Ham, R. C., J. Kamerbeek, C. Palacios et al. (16 co-authors) 2003. Reductive genome evolution in *Buchnera aphidicola*. *Proc. Natl. Acad. Sci. USA* **100**:581–586.
- Weisburg, W. G., J. G. Tully, D. L. Rose et al. (12 co-authors) 1989. A phylogenetic analysis of the mycoplasmas: basis for their classification. *J. Bacteriol.* **171**:6455–6467.
- Yoshida, T., S.-I. Maeda, T. Deguchi, and H. Ishiko. 2002. Phylogeny-based rapid identification of mycoplasmas and ureaplasmas from urethritis patients. *J. Clin. Microbiol.* **40**: 105–110.

William Martin, Associate Editor

Accepted February 17, 2004