

Gene Gain and Gene Loss in *Streptococcus*: Is It Driven by Habitat?

Pradeep Reddy Marri,¹ Weilong Hao,¹ and G. Brian Golding

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

Bacterial genomes can evolve either by gene gain, gene loss, mutating existing genes, and/or by duplication of existing genes. Recent studies have clearly demonstrated that the acquisition of new genes by lateral gene transfer (LGT) is a predominant force in bacterial evolution. To better understand the significance of LGT, we employed a comparative genomics approach to model species-specific and intraspecies gene insertions/deletions (ins/del among 12 sequenced streptococcal genomes using a maximum likelihood method. This study indicates that the rate of gene ins/del is higher on the external branches and varies dramatically for each species. We have analyzed here some of the experimentally characterized species-specific genes that have been acquired by LGT and conclude that at least a portion of these genes have a role in adaptation.

Introduction

With the availability of a large number of bacterial genomes, it has become evident that the gene set of an individual bacterial organism may not be completely representative of the other organisms belonging to the same genus/species. Even though the organisms belonging to the same genus/species share a common gene set (core), individual organisms differ in the subset of genes that are representative of the physiological and virulent properties of an organism (Wren 2000; Dobrindt and Hacker 2001). This subset of genes is thought to be responsible for the survival of an organism in its chosen niche. This adaptation-specific variation can be due to gene gain by lateral gene transfer (LGT) (Lawrence 1997, 1999; Lawrence and Ochman 1998; de Koning et al. 2000; Ochman et al. 2000; Lawrence and Ochman 2002; Springael and Top 2004; Hughes and Friedman 2005), gene loss (Cole et al. 2001; Ogata et al. 2001; Foster et al. 2005), or modification of some of the existing genes (Sokurenko et al. 1998; Feldgarden et al. 2003). Acquisition of new genes by LGT is a predominant force in bacterial evolution (Lan and Reeves 1996; Ochman et al. 2000; Gogarten et al. 2002; Jain et al. 2003; Kunin and Ouzounis 2003; Mirkin et al. 2003; Snel et al. 2005). Laterally acquired genes provide an organism with a readily available novel gene pool that provides additional physiological properties that are helpful for exploiting a new niche (Pal et al. 2005; Ricard et al. 2006).

In the study of bacterial genome evolution, lateral transfers together with deletions can be inferred with parsimony (Daubin, Lerat, et al. 2003; Daubin, Moran, et al. 2003; Mirkin et al. 2003; Hao and Golding 2004) and maximum likelihood methods (Hao and Golding 2006). However, the presence of genome sequences of congeneric species is the prerequisite for such a kind of analysis. The presence of multiple sequenced closely related genomes that inhabit different niches will not only enable us to get an understanding of the pattern of gene movement but also will provide us an excellent data set to gain insights into the role of species-specific genes. The *Streptococcus* genus, with 12

congeneric genomes sequenced, serves as an excellent group to model gene insertions and deletions. Moreover, LGT has been widely recognized among *Streptococcus* species (Balsalobre et al. 2003; Broker and Spellerberg 2004; Franken et al. 2004; Towers et al. 2004; Green et al. 2005; Sumbly et al. 2005).

The genus *Streptococcus* comprises a wide variety of pathogenic and commensal gram-positive bacteria. They inhabit a wide range of hosts, including humans, horses, pigs, and cows. Within each host, streptococci are found to colonize diverse habitats including mucosal surfaces, pharynx and the respiratory, intestinal, and urinogenital tracts. The 12 genomes of *Streptococcus* belong to 5 species: *Streptococcus pyogenes*, *Streptococcus agalactiae*, *Streptococcus pneumoniae*, *Streptococcus mutans*, and *Streptococcus thermophilus*.

Streptococcus pyogenes belongs to group A *Streptococcus* (GAS) and is a common cause of severe invasive infections resulting in high rates of morbidity and mortality. The common diseases caused by various strains of *S. pyogenes* include pharyngitis, cellulitis, sepsis, and acute rheumatic fever (Smoot et al. 2002).

Streptococcus agalactiae is a Lancefield's group B *Streptococcus* (GBS). It is a commensal bacterium inhabiting the intestinal tract of a significant proportion of the human population. Also present in the urinogenital tract, it is the leading cause of septicemia and meningitis in neonates (Glaser et al. 2002).

Streptococcus pneumoniae is a major cause of pneumonia and otitis media, and this species is one of the top 10 causes of death in the United States (Klein 1999). A readily transformed organism, *S. pneumoniae* is a transient commensal inhabiting the throat and upper respiratory tract of humans (Hoskins et al. 2001; Tettelin et al. 2002).

Streptococcus mutans is the leading cause of tooth decay in humans (Ajdic et al. 2002).

Streptococcus thermophilus is a food microorganism with major economic importance. *Streptococcus thermophilus* is a "Generally recognized as safe" species widely used in the manufacture of dairy products (Bolotin et al. 2004; Hols et al. 2005).

In the present study, we 1) have employed a comparative genomics approach to model the species-specific and intraspecies gene transfers among the streptococcal genomes using a maximum likelihood method and 2) studied the species-specific genes to better understand the role of LGT in adaptation.

¹ Both authors contributed equally to this work.

Key words: adaptive evolution, *Streptococcus*, maximum likelihood analysis, adaptation, lateral gene transfer, gene gain, gene loss, phylogeny.

E-mail: golding@mcmaster.ca.

Mol. Biol. Evol. 23(12):2379–2391. 2006

doi:10.1093/molbev/msl115

Advance Access publication September 11, 2006

Methods

Sequences Used

The study involved 12 streptococcal genome sequences: *S. agalactiae* NEM316 (Sag1; NC_004368; Glaser et al. 2002), *S. agalactiae* 2603V/R (Sag2; NC_004116; Tettelin et al. 2002), *S. mutans* UA159 (Smu; NC_004350; Ajdic et al. 2002), *S. pneumoniae* TIGR4 (Spn1; NC_003028; Tettelin et al. 2001), *S. pneumoniae* R6 (Spn2; NC_003098; Hoskins et al. 2001), *S. pyogenes* MGAS10394 (Spy1; NC_006086; Banks et al. 2004), *S. pyogenes* MGAS8232 (Spy2; NC_003485; Smoot et al. 2002), *S. pyogenes* MGAS315 (Spy3; NC_004070; Beres et al. 2002), *S. pyogenes* SSI-1 (Spy4; NC_004606; Nakagawa et al. 2003), *S. pyogenes* M1 GAS (Spy5; NC_002737; Ferretti et al. 2001), *S. thermophilus* CNRZ1066 (Sth1; NC_006449; Bolotin et al. 2004), and *S. thermophilus* LMG18311 (Sth2; NC_006448; Bolotin et al. 2004). The genome of *Lactococcus lactis* (NC_002662; Bolotin et al. 2001) was used as an outgroup. All the sequences were downloaded from the National Center for Biotechnology Information (NCBI) database (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria>).

Generation of Species Tree

A concatenated DNA sequence, obtained by joining the individual DNA sequences of *rpoB*, *gltX*, *pheS*, *purD*, *recA*, *ynaE*, and *yjjG*, was used to reconstruct the phylogeny of streptococcal taxa. The phylogeny was reconstructed using MrBayes (Huelsenbeck and Ronquist 2001; 200,000 generations, sampled every 100 generations with a gamma-distribution model and invariant class). The robustness of the obtained phylogeny was tested using single-copy genes present in all 13 genomes (including *L. lactis*). A phylogeny for each gene was constructed by PROTML, and the consensus tree was generated by the CONSENSE program in the PHYLIP package version 3.6 (Felsenstein 1989). To avoid the confounding effects of duplication during evolution (Gu et al. 2002; Zhang et al. 2003), paralogs of gene families were excluded from the phylogeny construction.

Maximum Likelihood Analysis

The genes were first classified into gene families. The method to identify members of a gene family has been described in Hao and Golding (2004). In short, potential homologs were identified according to sequence similarities, and all paralogs in each genome were clustered as a single gene family. The phyletic patterns (gene presence or absence in each genome) of all genes were used for a maximum likelihood analysis. The method to estimate the maximum likelihood value has been described earlier (Hao and Golding 2006). In brief, gene presence or gene absence was treated as a binary character (0,1) state. If the evolutionary history is known, the probability of gene gain/loss can be computed from insertion and deletion rates. Like the maximum likelihood estimation of phylogeny using DNA sequence, the likelihood of a character state at any node on a given phylogeny can be calculated from the character states in the immediate descendant nodes. On each branch, the insertion rate was assumed equal to the deletion rate.

Several different models were used to estimate the maximum likelihood in this study. The branch-specific rates were optimized to find those rates that maximized the likelihood of observing the gene patterns.

Identification of Species-Specific Genes

The protein sequences from all the genomes were compared with each other using BlastP (Altschul et al. 1997), with an *E*-value cutoff set at 1.0×10^{-20} and an additional criteria of match length set at 85% of the query sequence. A set of genes present in all the strains belonging to a single species but not in any other streptococcal genomes under study were considered specific to that species. For example, the genes present in both strains of *S. thermophilus* (CNRZ1066 and LMG18311) but absent from the other streptococcal genomes were considered specific to *S. thermophilus*. It should be noted that all the species-specific genes of *S. mutans* that did not have a hit in NCBI nr database were treated as ORFans (an annotated gene that is uniquely present in one genome; Daubin and Ochman 2004) and not used for further analysis. As more data are accumulated, we are aware that what is considered unique in each species will be altered.

Lateral Gene Transfer

The unique proteins of each species were compared with the NCBI nr database using BlastP with the expect value cutoff set at 1.0×10^{-10} to identify homologs in other organisms. For each protein, the first 50 hits with an expect value less than 1.0×10^{-10} were chosen for further analysis regarding possible LGTs. The complete protein sequences of these 50 hits were extracted from the GenBank database, and a multiple alignment was performed using ClustalW (Thompson et al. 1994). The multiple alignment was used to generate a phylogenetic tree using the Neighbor-Joining method (Saitou and Nei 1987) using a distance matrix obtained from "protdist" (JTT model of amino acid change by Jones, Taylor, and Thornton 1992), as implemented in PHYLIP (Felsenstein 2004). Distances were calculated with Γ -distributed rates (α calculated using Tree-Puzzle; Strimmer and von Haeseler 1996).

Calculation of K_a/K_s Ratio

The genes present in the *S. pyogenes* taxa were used to calculate the K_a/K_s ratio (ω) and tree length using the PAML package (Yang 1997). The *S. pyogenes* species-specific genes were compared with the genes present in all 13 genomes. The tree length was calculated as the sum of the branch lengths for the taxa only within the *S. pyogenes* group using the maximum likelihood method from the PAML package. The tree length gives the expected number of substitutions per site along all branches in the phylogeny. A single K_a/K_s ratio was assumed throughout the length of each gene sequence in this study. To avoid the effects of duplication during evolution (Gu et al. 2002; Zhang et al. 2003), paralogs of gene families were also excluded from the K_a/K_s ratio and tree-length analyses. Similarly, the K_a and K_s values were estimated between the 2 *S. pneumoniae* strains, between the 2 *S. agalactiae* strains, and between the 2 *S. thermophilus* strains. Protein sequences

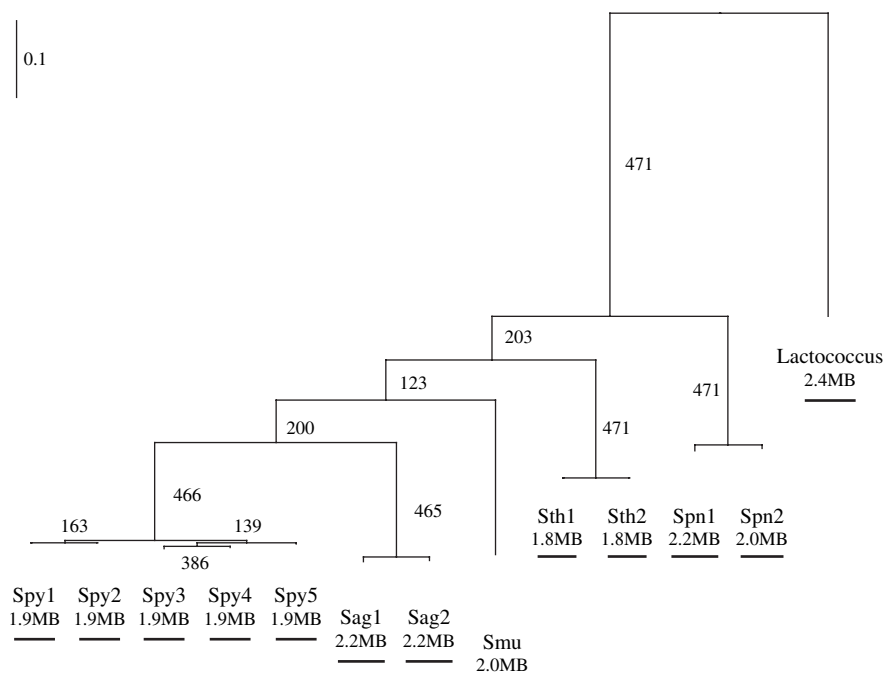


FIG. 1.—Robustness of the phylogeny based on concatenation of *rpoB*, *gltX*, *pheS*, *purD*, *recA*, *ynaE*, and *yjjG* DNA sequences. In all, 471 single-copy genes, which are present in all 13 strains, were used to test the robustness of the tree. The topology of the consensus tree is exactly same as the concatenated DNA phylogeny. All branches are supported by posterior probability values of >90%. The number of genes supporting each branch is labeled next to that branch. The abbreviations are Spy1 (*Streptococcus pyogenes* MGAS10394), Spy2 (*S. pyogenes* MGAS8232), Spy3 (*S. pyogenes* MGAS315), Spy4 (*S. pyogenes* SSI-1), Spy5 (*S. pyogenes* M1 GAS), Sag1 (*Streptococcus agalactiae* NEM316), Sag2 (*S. agalactiae* 2603V/R), Smu (*Streptococcus mutans* UA159), Sth1 (*Streptococcus thermophilus* CNRZ1066), Sth2 (*S. thermophilus* LMG 18311), Spn1 (*Streptococcus pneumoniae* TIGR4), and Spn2 (*S. pneumoniae* R6).

and their corresponding DNA sequences were extracted from the genome. Protein sequences were aligned using ClustalW (Thompson et al. 1994), and nucleotide sequence alignments were created from the protein alignments by replacing each amino acid with its corresponding codon.

Testing for Positively Selected Sites

The genes present in the *S. pyogenes* taxa were used to test for sites under positive selection using the PAML package (Yang 1997). Selective constraints across sites were estimated using different models of codon substitution (Yang et al. 2000). Model M0 assumes that all amino acid sites have the same value of ω . Model M3 assumes 3 classes of sites with ω values. Model M7 (beta) assumes 8 classes of sites with ω values, which are limited to the interval (0,1) and follow a beta distribution. Model M8 (beta + ω) is similar to M7, but with an additional ω category that can exceed 1. Nested models (M0 vs. M3 and M7 vs. M8) were compared using a likelihood ratio test. It has been shown that M7 versus M8 comparison is the most stringent test of positive selection (Anisimova et al. 2001). Only the genes that were detected to have sites under positive selection from an M7 versus M8 comparison were used for further comparison.

Structural Analysis

Threaded protein structures for 3 proteins SpyM18_0133, SpyM18_0491, and SpyM18_0617 were

created using Wurst (Torda et al. 2004). Protein structure images were produced using RasMol 2.7.1 (Sayle and Milner-White 1995).

Results

Reconstruction of Phylogeny

Because single-gene phylogenies might not always reflect the evolutionary history of species due to the high degree of LGT (Ochman et al. 2000; Kunin and Ouzounis 2003; Mirkin et al. 2003; Baptiste et al. 2005), a concatenated DNA sequence obtained by joining 7 DNA sequences was used to generate a species tree. The robustness of the tree was tested with 471 single-copy genes present in all 13 genomes. The consensus phylogeny obtained from the 471 single-copy genes was the same as obtained from the concatenated sequence indicating that it was a reasonable choice for phylogenetic construction. However, the number of genes which support the topology is low on some branches such as among the Spy strains, on the branch leading to Spy and Sag, on the branch leading to Spy, Sag, and Smu, and on the branch leading to Sag, Smu, and Sth (fig. 1). The low support on these branches is almost certainly due to the lack of informative signals to distinguish the very short branches.

Insertion and Deletion Rates

The maximum likelihood analysis used the phylogeny obtained by concatenation of protein sequences to infer the

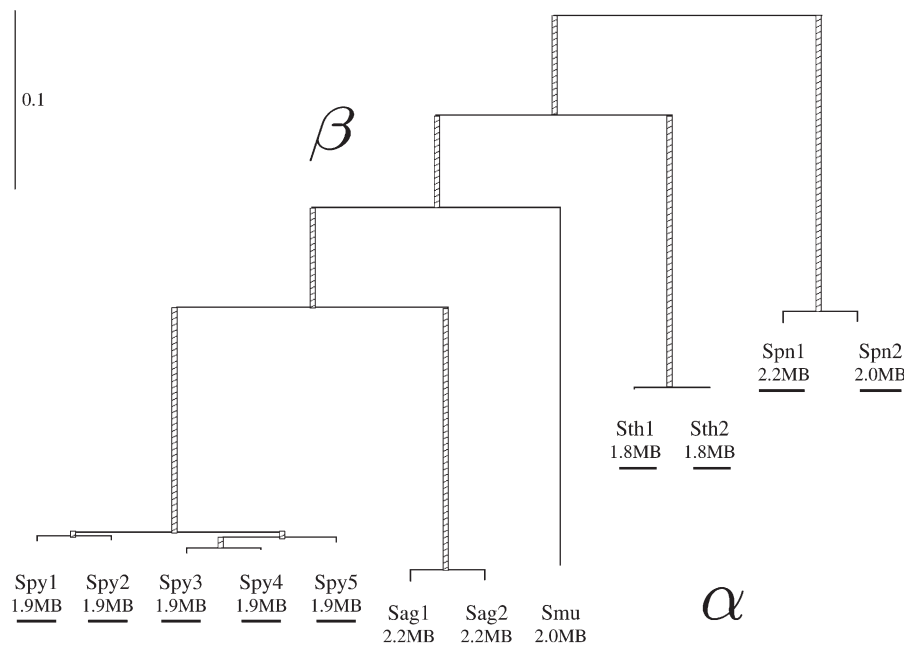


FIG. 2.—Ins/Del rates were tested separately between external branches (α) and internal branches (β). Internal branches are hatched.

relative insertion/deletion (ins/del) rates by assuming that individual insertion and deletion events occur independently. Assuming a single constant ins/del rate ($\alpha = \beta$ in fig. 2) on the phylogeny resulted in an ins/del rate of 1.17 ($\text{Ln}L = -16,148$). This is larger than 1, which is the evolutionary time period required to observe one substitution per site, indicating a higher rate of gene ins/dels. It is greater than the rate inferred from a study of *Bacillus* strains (that rate was 0.51; Hao and Golding 2006). When the ins/del rates were calculated separately for internal and external branches, the rates on external branches were higher than those on internal branches (2.14 vs. 1.01; $\text{Ln}L = -16,001$). These results suggest that more gene ins/dels take place at the tips of the phylogeny. Additionally, when the ins/del rates were tested separately on species branches and within-species branches (as shown in fig. 3), it was found that within-species branches had a higher ins/del rate than the branches leading to a species and that ins/del rates varied among different species (table 1).

Faster Evolution of Recently Transferred Genes

The analysis of the tree length of the species-specific genes of *S. pyogenes* showed that the recently transferred genes have larger tree lengths indicating a faster rate of evolution (fig. 4A vs. B; $P < 0.001$ in a Wilcoxon rank test). The analysis of the nonsynonymous/synonymous substitution rate (K_a/K_s) ratio among the species-specific genes indicated that the genes transferred recently have a higher K_a/K_s ratio compared with the genes that have been resident within the species longer. The K_a/K_s ratio of the species-specific genes of *S. pyogenes* was higher compared with the genes that were present in all 13 genomes (fig. 5A vs. B; a Wilcoxon rank test of the difference in these distributions is significant with $P < 0.001$). Similar results were obtained by comparing the 2 *S. pneumoniae* strains. The K_a/K_s

ratio of the Spn-specific genes was higher than the genes present in all the 13 genomes. Moreover, both synonymous and nonsynonymous changes in Spn-specific genes are greater than those in the genes present in all 13 strains (each with $P < 0.001$ in a Wilcoxon rank test; fig. 6). Similarly, Sth- and Sag-specific genes evolve faster than their counterparts that are present in all the 13 genomes (data not shown).

Species-Specific Genes and LGTs

The species-specific genes of each species were divided into 3 categories: 1) LGTs (genes that cluster outside Firmicutes and are supported by a bootstrap value of 70 and above) and 2) possible multiple deletion or lateral transfers (MDLT; genes that cluster within Firmicutes) and 3) no hits (genes that have no hits in the NCBI database other than the species under study). Some of the species-specific genes for each species are discussed below. For the identification of LGT, the species-specific genes initially obtained were used for Blast analysis with a relaxed E value (1.0×10^{-10}) and no length constraint. As a result of this relaxation in the search criteria, the species-specific genes reported here are slightly lower than those used for the likelihood analysis. An entire list of species-specific genes can be found at http://evol.mcmaster.ca/Streptococcus_specific.html.

Streptococcus mutans

The study identified a set of 151 genes that are specific to *S. mutans* of which LGT and MDLT together constitute 106 genes. Of the 106 laterally transferred genes' LGT/MDLT, 78 are functionally annotated and the remaining encode hypothetical proteins. Twenty of the 78 functionally annotated laterally transferred genes (LTGs) encode transport-related proteins. Some of the specific transport-related LTGs include SMU.1961 and SMU.1963 encoding a

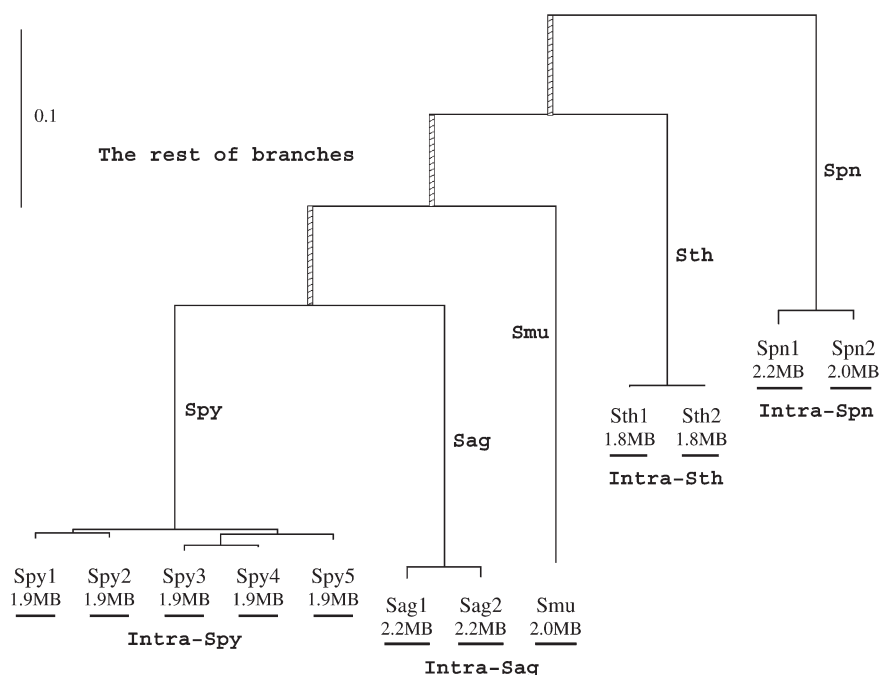


FIG. 3.—Ins/Del rates were tested separately between species branches (Spy, Sag, Smu, Sth, and Spn) and within-species branches. The rest of the branches are hatched.

sugar-specific phosphotransferase system (PTS) and sugar-binding periplasmic protein, respectively, and possibly acquired from *Lactobacillus johnsonii* and SMU.1960 encoding a mannose-specific PTS that is possibly acquired from *Leuconostoc mesenteroides*. The acquisition of these genes might aid in the transport of sugars to and from the cells and help in adjusting to the high and low concentrations of sugars in the oral cavity (Vadeboncoeur and Pelletier 1997). However, functional characterization of the remaining proteins will enable a better understanding about the adaptive nature of these proteins. Additionally, the acquisition of the alcohol dehydrogenase gene (*adh*) and an upstream gene SMU.118, encoding a protease from *Actinobacillus pleuropneumoniae* (see fig. 1 at http://evol.mcmaster.ca/Streptococcus_specific.html) might enhance its ability to survive under low pH conditions (Korithoski et al. 2005). The presence of the *fru B* gene encoding fructan hydrolase, the malolactic enzymes' *mleR* and *mleS*, and the genes *gltA* and *gltB* encoding glutamate synthase subunits might increase its ability to utilize a diverse array of carbohydrates (Bowden and Hamilton 1998; Ajdic et al. 2002). The gene *fic* encodes a filamentous protein and might possibly aid in adhesion. The LTGs also included 3 bacteriocin-related proteins that are specific to *S. mutans*. Bacteriocins have antimicrobial effect on most of the gram-positive bacteria (Fabio et al. 1987; Ajdic et al. 2002) and might help in maintaining its niche. The bacitracin-encoding gene *bacA1* is possibly acquired from *Escherichia coli*, whereas the genes SMU.1149 and SMU.1150 encoding bacteriocin immunity proteins are acquired from *Staphylococcus warnerii*. Another important feature of *S. mutans* is the presence of the genes belonging to histidine metabolism pathway (see fig. 2 at http://evol.mcmaster.ca/Streptococcus_specific.html). These results indicate that some of the *S. mutans*—

specific genes confer on it abilities to survive under low pH conditions, metabolize a large variety of carbohydrates, and survive the competition from other microorganisms living in its habitat.

Streptococcus pneumoniae

The study revealed the presence of 288 species-specific genes for *S. pneumoniae* of which 117 were identified as LGT/MDLT. Unlike in *S. mutans*, where a majority of the LGT are functionally annotated, only half of the LGT have been assigned a putative function in *S. pneumoniae*. The LTGs/MDLT include a higher number of genes encoding restriction endonucleases agreeing with an earlier report of Martin-Galiano et al. (2004), which suggested that a higher proportion of restriction endonuclease genes are acquired by LGT. The presence of these genes might help to keep a check on the foreign DNA that comes into the genome from its surroundings by transformation. Some of the functionally annotated LGT include the 3 genes encoding proteins related to spermidine synthesis and transport—spermidine synthetase (*speE*), carboxynorspermidine decarboxylase

Table 1
Ins/Del Rates Inferred on Species Branches and Within-Species Branches as Shown in Figure 3

	Maximum Likelihood Estimation						
	Branches	Spy	Sag	Smu ^a	Sth	Spn	The Remainder ^a
LnL = -15,121	Species	1.21	0.93	0.68	1.11	1.04	0.32
	Within species	5.84	4.83		21.70	2.64	

^a No within-species branches are available.

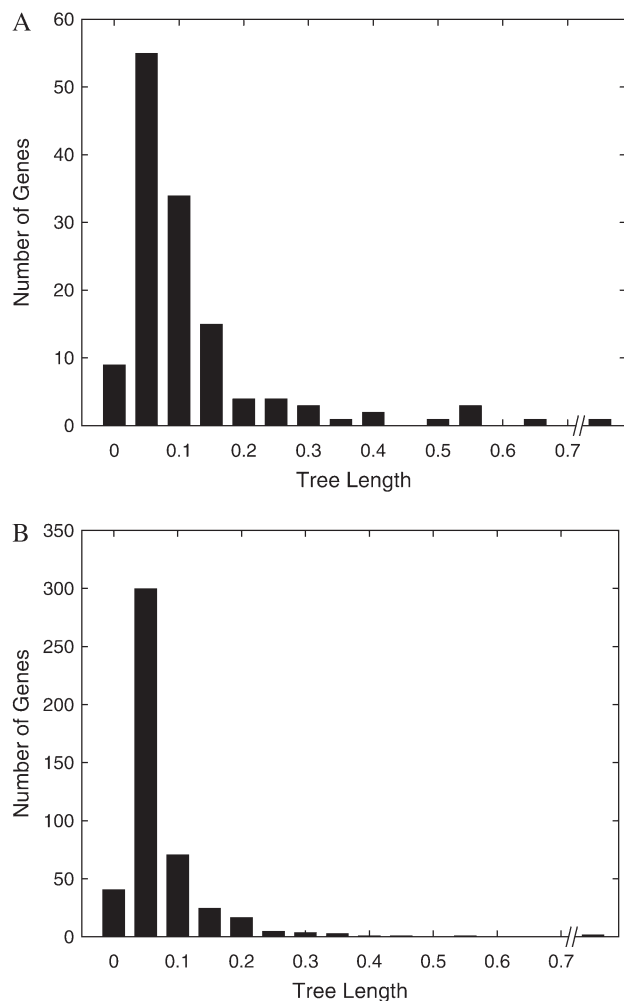


FIG. 4.—Fast evolution in the *Streptococcus pyogenes* group-specific genes. (A) Tree length for the Spy taxa as indicated by genes that are present only within this group of taxa; (B) tree length for the Spy taxa as indicated by genes that are present in all 13 species. The tree lengths are calculated as the sum of the branch lengths for each gene using the maximum likelihood estimate from PAML. The breaks on the x axis indicate the points after which all the values are pooled.

(*nspC*), and spermidine acetyltransferase (*bltD*). These together with the lysine decarboxylase gene, *cad* (spr0816), are acquired from *Selenomonas ruminatum* (see fig. 3 at http://evol.mcmaster.ca/Streptococcus_specific.html). The teichoplanin resistance protein-encoding gene *vanZ* that is possibly acquired from *Clostridium tetani* might lead to an increased resistance of *S. pneumoniae* against the antibiotics vancomycin and teichoplanin (Arthur et al. 1995). *Streptococcus pneumoniae* has 2 copies of the gene (*nanA* and *nanB*) encoding neuraminidase (Camara et al. 1994; Berry et al. 1996). The gene *nanB* appears to have been acquired from *Macrobacteria* by LGT agreeing with an earlier study that also demonstrated the acquisition of *nanB* by *S. pneumoniae* via LGT (Martin-Galiano et al. 2004). As neuraminidase causes damage to the cell wall of *Neisseria meningitidis* and *Haemophilus influenzae*, coinhabitants of the respiratory tract, the acquisition of a second copy of the neuraminidase gene might promote the colonization of *S. pneumoniae* to nasopharynx region (Tong et al. 2000;

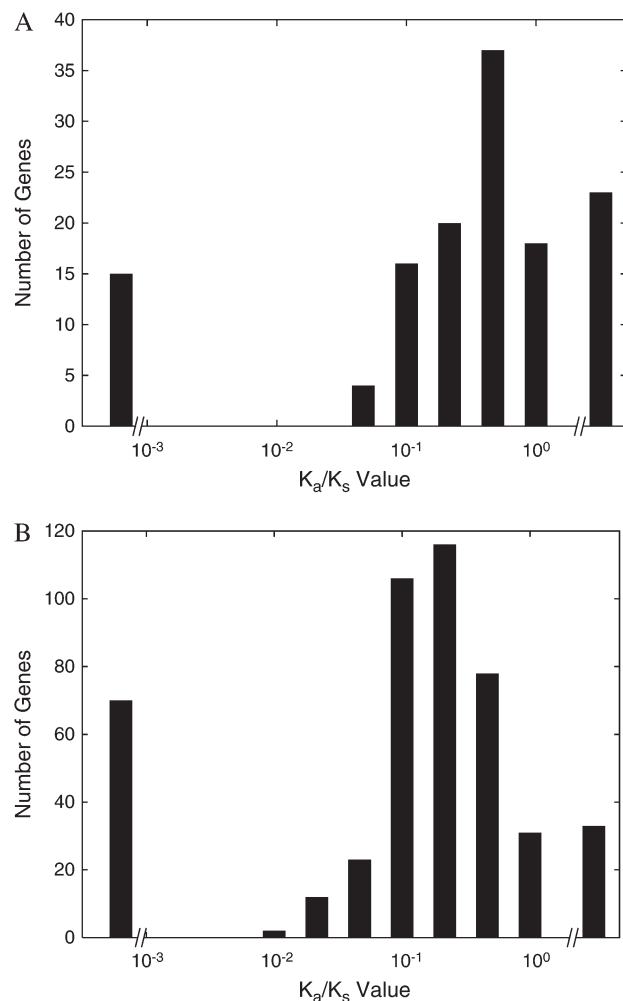


FIG. 5.—Relaxed or positive evolution in the *Streptococcus pyogenes* group-specific genes. (A) K_a/K_s ratio for the Spy taxa as indicated by genes that are present only within this group of taxa; (B) K_a/K_s ratio for the Spy taxa as indicated by genes that are present in all 13 species. The breaks on the x axis indicate the points before and after which all the values are pooled.

Shakhnovich et al. 2002). The pyridoxine biosynthesis gene (*pdxI*) and a gene spr1321 (upstream of *pdxI*) had similarity to *H. influenzae*, another inhabitant of respiratory tract, suggesting that *S. pneumoniae* may have acquired these genes from *H. influenzae* by LGT, and they possibly have a role in colonization of *S. pneumoniae* in the respiratory tract. The species-specific genes also include the genes *fcsK*, *fucU*, *fucA*, and *fcsR* involved in fucose metabolism and possibly acquired from *Clostridium perfringens*. However, the significance of these genes is not clearly understood as *S. pneumoniae* cannot utilize fucose as a solitary energy source (Hoskins et al. 2001).

Streptococcus agalactiae

The study identified a set of 222 genes that are specific to *S. agalactiae* of which 85 are LGT/MDLT. A majority (61/85) of the LTGs/MDLT encode hypothetical proteins. The prominent species-specific genes include gbs0671 encoding beta-glucuronidase, gbs1919 encoding a neuraminidase,

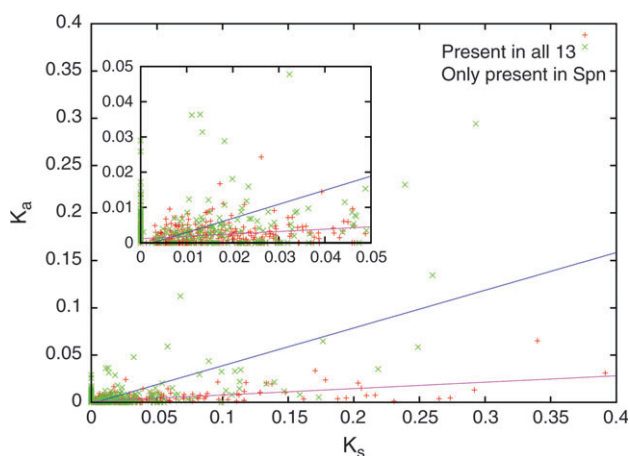


FIG. 6.—Fast and relaxed evolution in the *Streptococcus pneumoniae*-specific genes. The regression line for genes present in all 13 strains is $y = 0.067x + 0.001$ ($R^2 = 0.386$), whereas the regression line for *S. pneumoniae*-specific genes is $y = 0.399x - 0.001$ ($R^2 = 0.466$).

and the gene cluster consisting of genes *cylX*–*cylE* that are responsible for the production of hemolysin. The presence of a gene encoding beta-glucuronidase (*gbs0671*) that catalyzes a wide variety of water soluble carbohydrates might help in a better survival of *S. agalactiae* in the gastrointestinal (GI) tract. A high similarity (74% identity based on BlastP search) of this gene to a gene in *Haemophilus somnus*, an inhabitant of intestinal and urinogenital tract of cattle, might suggest that it might have acquired this gene from *H. somnus* by LGT. The presence of a gene encoding neuraminidase, an antibacterial agent, might suggest a mechanism to survive the competition from other microflora inhabiting the GI tract, whereas the hemolysin produced by the *cyl* genes has been shown to have a role in pathogenesis (Spellerberg et al. 1999, 2000). Hemolysin causes injury to various epithelial and endothelial cells and macrophages and was found to be responsible for the activation of endothelial cell genes that are implicated in inflammatory response of *S. agalactiae*-induced meningitis (Yother et al. 2002).

Streptococcus pyogenes

The analysis identified a set of 169 genes that are specific to *S. pyogenes*. Of the 169 species-specific genes, only 47 were identified as LGT/MDLT. The prominent species-specific LTGs include *salR*, *salY*, and *salA*, belonging to an operon that imparts resistance against salivaricin. This operon is present in another species of *Streptococcus*, *Streptococcus salivarius* but absent in *S. agalactiae*, *S. pneumoniae*, *S. mutans*, and *S. thermophilus* indicating that *S. pyogenes* might have acquired these genes from *S. salivarius* by LGT. In light of a recent study indicating that a salivaricin-related gene, *salA*, can modulate lantibiotic production and possibly influence the population ecology in the oral cavity (Upton et al. 2001), the acquisition of salivaricin-related genes could possibly help in a better survival of *S. pyogenes* in its niche. The other set of *S. pyogenes*-specific genes include the gene cluster consisting of genes *sagA*–*sagI* that encode streptolysin S, an oxygen-stable cytolytic toxin (Nizet et al. 2000). This gene cluster

might help in pathogenesis as streptolysin elicits a cytotoxic effect and inhibits neutrophil phagocytosis in vitro and has also been found to cause soft-tissue damage in a murine model (Yother et al. 2002). Interestingly, *S. pyogenes* has the enzymes *hutH*, *hutU*, *hutI*, and *hutG* that are responsible for the degradation of amino acid histidine to L-glutamate (see fig. 4 at http://evol.mcmaster.ca/Streptococcus_specific.html). This is surprising because *S. pyogenes* does not have the enzymes related to the histidine metabolism pathway, probably suggesting that *S. pyogenes* may be able to absorb histidine from its host and convert it to glutamate to use as an energy resource. However, *S. thermophilus* has the first 2 enzymes (*hutH* and *hutU*) in this pathway, and they are not homologous to those present in *S. pyogenes*.

Streptococcus thermophilus

The study identified 267 species-specific genes for *S. thermophilus* of which only 68 were classified as LGT/MDLT. Of the 68 LTGs/MDLT, 26 encode hypothetical proteins, whereas the remaining are functionally annotated. One of the important abilities of *S. thermophilus* is to hydrolyze urea present in milk. The urease gene cluster absent in other studied taxa is responsible for the hydrolysis of urea (Mora et al. 2004, 2005) and has also been identified to have a role in stress response that helps in the survival of *S. thermophilus* under low pH conditions (Mora et al. 2005). The regulatory mechanism of the urease gene cluster in *S. thermophilus* is also modified to be independent of carbohydrate concentration, unlike the closely related *S. salivarius*, where the regulation is driven by carbohydrate concentration (Chen et al. 1998). This could be due to the fact that carbohydrate is never a limiting factor for *S. thermophilus* (due to its utilization of milk) unlike for *S. salivarius*, which inhabits the oral cavity. Some of the other functionally characterized *S. thermophilus*-specific genes include the genes *tatA* and *tatC* that are a part of twin arginine translocase system and the genes *labB*, *labC*, and *labT* responsible for the synthesis and transport of a lantibiotic. Though most of the genes related to virulence factors have been lost or reduced to pseudogenes in *S. thermophilus* (Bolotin et al. 2004), we found that it has the genes *labB*, *labC*, and *labT*, probably encoding bacteriocins. The role of these genes is not clearly understood, but there have been reports suggesting the production of bacteriocins by *S. thermophilus*, which have antibacterial effects (Ward and Somkuti 1995; Ivanova et al. 1998). This may be one of the mechanisms to survive competition, using milk as a primary food source.

How Adaptive Are LGTs?

The above analysis of the species-specific genes indicates that a portion of the genes acquired by LGT are probably adaptive and might help in better survival of the organism in its chosen niche. To get further insights into the adaptive role of LGT, we compared the recently transferred genes of *S. pyogenes* with the genes that have had a longer residence time to get a measure of the proportion of LGT evolving under positive selection. The results indicate that the proportion of genes under positive selection (according to PAML) is 3-fold higher for the recently

Table 2
Putative Genes under Positive Selection in Recently Transferred and Ancient Genes of *Streptococcus pyogenes*

	Recent Transfers	Ancient Genes
Number of single-copy genes	134	471
Average gene length	260 ± 21	332 ± 9
Number of genes with nonsynonymous changes	115	386
Number of genes under positive selection	26	31

transferred genes compared with the ancient genes. Only 6% of the ancient genes showed traces of positive selection compared with 19% of the genes that indicated positive selection in case of the recently transferred genes (table 2).

Structure Analysis

We were able to predict the threaded structures for only 3 of the 26 proteins that were identified to be under positive selection. Mapping of positively selected sites on the corresponding protein structures revealed that the positively selected sites were mostly on the surface of the protein molecules (fig. 7).

Functional Classes of Genes in Gene Patterns

We have classified the genes in each pattern (patterns as listed in table A.1, Supplementary Material online) into functional categories based on COG database (Tatusov et al. 2000) to understand the role of gene gains/gene losses in these genomes. As expected, the highest percentage of genes present in all the taxa (pattern 11111111111) belonged to information processing representing the core genes that are less amenable to lateral transfer (Jain et al. 2003). Interestingly, the second most abundant class of genes in this category represented the genes whose functions are uncharacterized to date. The presence of a large number of hypothetical genes conserved across all the taxa suggests important roles for these genes. Analysis of the genes belonging to other patterns revealed that about 35–60% of these genes encode hypothetical proteins. It is clearly evident from figure 8 that the genomes also have a tendency to lose specific sets of genes. *Streptococcus pyogenes* appears to have specifically lost many genes that are involved in amino acid metabolism and transport (yellow vertical bars in fig. 8), whereas *S. mutans* and *S. thermophilus* appears to have lost some of the genes responsible for carbohydrate metabolism and transport that are present in the other *Streptococcus* genomes (dark green vertical bars in fig. 8).

Discussion

In addition to classical evolutionary change via single-site substitutions, bacterial genomes also evolve rapidly by gaining a new set of genes, by losing an existing set of genes, and/or by duplication of its genes. Although evolution by duplication seems to be less prevalent among bacteria, acquiring a new set of genes that give an adaptive advantage is one of the prominent features of bacterial evolution (Lan and Reeves 1996; Gogarten et al. 2002; Jain et al. 2003). Bacteria modify their genomes in such

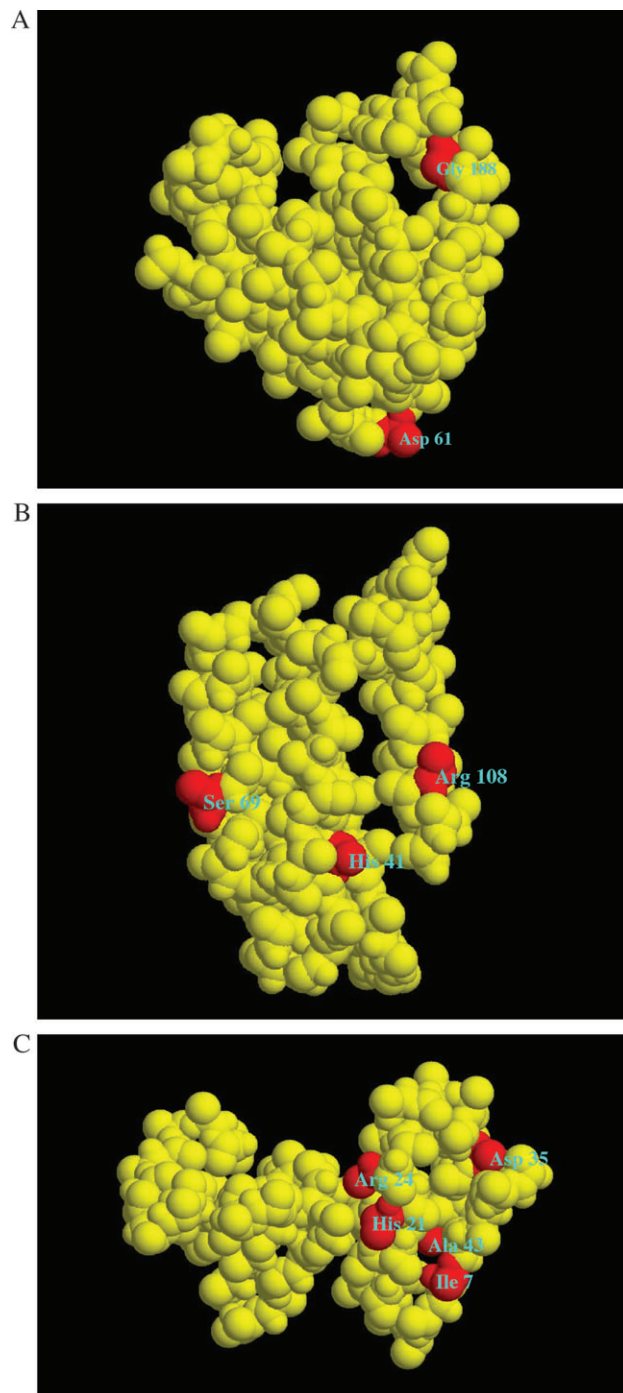


FIG. 7.—Amino acids under positive selection (red) in SpyM18_0133 (A), SpyM18_0491 (B), and SpyM18_0617 (C).

a way that it gives them evolutionary advantage and helps survival in their chosen niche (Sokurenko et al. 1998; de Koning et al. 2000; Wren 2000; Feldgarden et al. 2003; Read et al. 2003).

A comparison of the closely related bacterial species that live in different habitats will help to identify species-specific genes that are responsible for its survival in that particular environment. The current study involved the comparison of streptococcal genomes belonging to 5 different species that live in different habitats. The insertion and

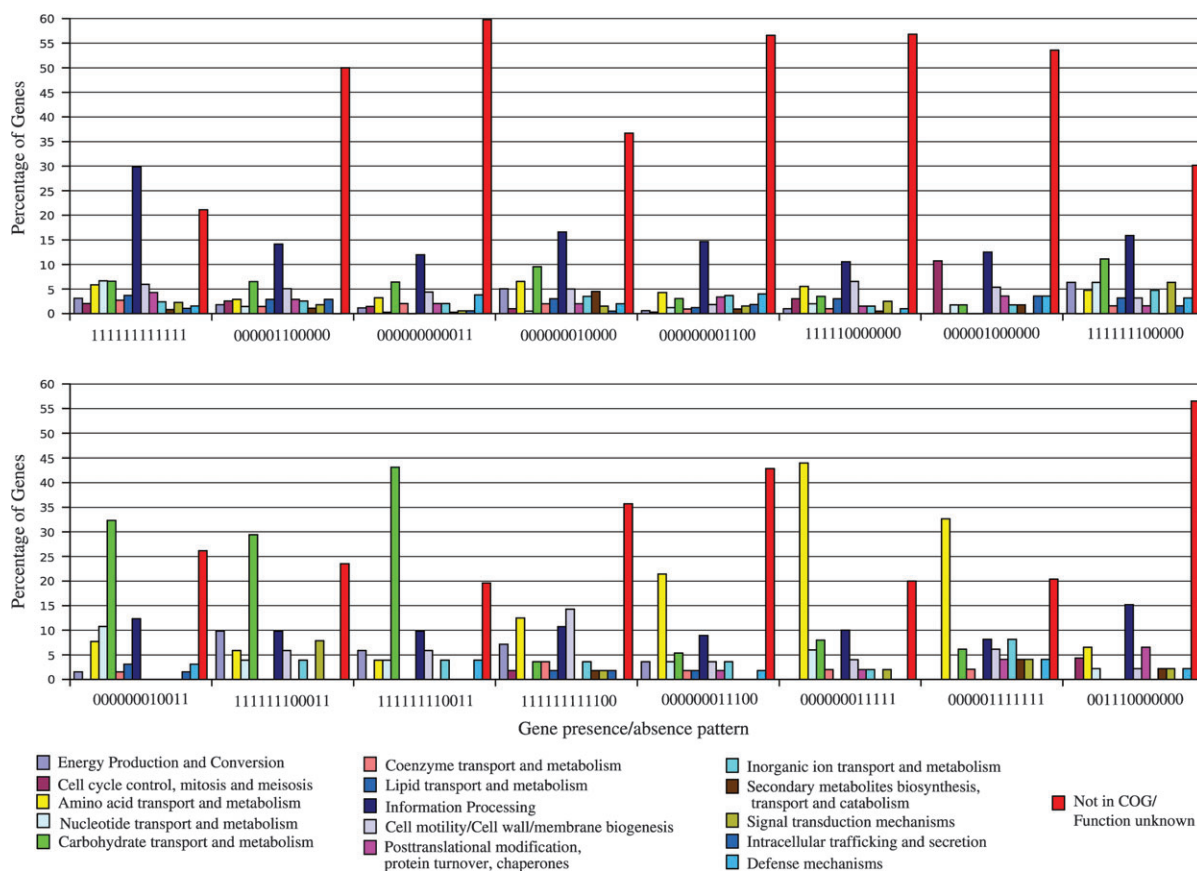


FIG. 8.—Functional classification of various categories of *Streptococcus* genes. Genes in each category, represented by the pattern of gene presence/absence in each of the genome studied, were divided into functional classes based on the COG database. For the details of the various categories of genes refer table A.1, Supplementary Material online.

deletion rates were modeled using maximum likelihood to understand the rate of ins/del at different levels of the phylogeny.

The primary requirement for the maximum likelihood modeling is the presence of a robust phylogeny for the genomes under study. Phylogenies obtained from single genes can sometimes be distorted due to rampant LGT (Ochman et al. 2000; Kunin and Ouzounis 2003; Mirkin et al. 2003; Baptiste et al. 2005), and rRNA sequences are not useful due to the lack of informative characters differentiating closely related species and varying functional constraints over the molecule (Fox et al. 1992; Santos and Ochman 2004). We used a concatenated DNA sequence obtained by joining the sequences of 7 genes to reconstruct the phylogeny of streptococcal species. The 7 genes were chosen such that they were present in all taxa in this study. To reduce computational burden, only 7 genes rather than a larger number of genes were chosen. The robustness of the topology was tested using the phylogenies of 471 single-copy genes present in all taxa. The topology of the concatenated tree was identical with the consensus tree generated using 471 genes present in all the 13 genomes.

Assumption of a constant ins/del rate on internal and external branches resulted in a single constant ins/del rate of 1.17 gene gain/loss per rate of base substitution. Branch lengths are measured relative to the estimated number of

base substitutions, suggesting that there is 1 gene gained/lost for every nucleotide substitution. The rate of ins/del on external branches (2.14) is higher than that on internal branches (1.01). Because the branch lengths leading to the *Spy* strains are small and yet they vary in gene content, the raw data also suggest that the rate of ins/del is larger than that of base substitution. The high ins/del rate indicates that the gain/loss of genes plays an important role in the evolution of streptococcal genomes. These results are supported by a recent study by Ochman and Davalos (2006), which shows a high gene turnover rate in the evolution of *E. coli*. Moreover, the presence of about 10–20% of the genes that are specific to each species and only 838 genes that are common to all the taxa studied (see table A.1, Supplementary Material online) indicates that these genomes evolve largely by gene gain/loss. The rate is twice the rate obtained for the *Bacillus* group (0.51) in another study (Hao and Golding 2006). The higher rate in *Streptococcus*, which forms a more closely related group than the *Bacillus* strains according to the rate of base substitutions, confirms the earlier study in *Bacillus* that inferred a higher rate of gene transfer at the tips of the phylogeny (Hao and Golding 2006). The higher ins/del rate at the tips of the phylogeny is further confirmed by an almost 2-fold higher rate on external branches compared with the internal branches. Similar results were observed by comparing the ins/del rates on the species-specific branches with those on the within-species

Table 3
Distribution of LTGs

Group	<i>Streptococcus mutans</i>	<i>Streptococcus pneumoniae</i>	<i>Streptococcus agalactiae</i>	<i>Streptococcus pyogenes</i> ^a	<i>Streptococcus thermophilus</i>
Actinobacteria	5	5	2	1	3
Archaea	1	3	2	1	4
Cyanobacteria	1	1	2	0	0
Eukaryotes	1	2	3	1	0
Firmicutes	73	70	55	31	43
Proteobacteria	14	27	13	3	13
Others ^b	11	19	8	3	5
Total	106	117	85	40	68

^a *Streptococcus pyogenes* has an additional 7 genes acquired from bacteriophages.

^b Others include *Treponema denticola*, *Thermotoga maritima*, *Paenobacillus durus*, *Fusobacterium nucleatum*, and *Chlorobium tepidum*.

branches. The different inferred rates for different species could be due to slightly different genome sizes of different species. However, genome-size change is mainly due to the expansion of gene family size in large genomes, and LGT plays little role (Pushker et al. 2004; Wiezer and Merkl 2005). The rate estimation of ins/del is robust for different cutoffs used for determining gene homologues. Different cutoffs (expect value $< 10^{-20}$ with match length $> 85\%$, expect value $< 10^{-10}$ with match length $> 70\%$, and expect value $< 10^{-5}$ with match length $> 50\%$) show the same essential trend that external branches tend to have higher rates of ins/dels than internal branches (Supplementary Material online).

In this study, the phyletic patterns were derived primarily from genome annotation. As described previously in Hao and Golding (2006), nonannotated genes were picked up via a TBlastN search, and genes that are uniquely present in only one studied genome were removed from the study. Furthermore, a comparison was made by removing the ORFs only present in *Spy* but not present in any other complete bacterial genomes (Supplementary Material online). The rates do not change remarkably after removing the *Spy* group unique genes. This suggests that the fast rate of evolution of recently acquired genes is not an artifact of fast-evolving but erroneously annotated ORFs in *Spy*.

It is clearly evident from the results of this study that about half of the species-specific genes are possibly acquired by LGT. Most of these genes come from diverse backgrounds (table 3), and many of these genes encode hypothetical proteins. However, it appears that the laterally acquired genes fall into 2 broad categories: genes that help to sustain the bacteria in a niche and survive the competition from other inhabitants of its niche and genes that help in virulence and pathogenesis. The acquisition of large number of transport and carbohydrate metabolism-related genes by *S. mutans* appears to be in accord with its ability to survive under low pH conditions in the oral cavity. On the other hand, the acquisition of many genes that encode restriction endonucleases probably helps *S. pneumoniae* keep a check on the foreign DNA that comes in by transformation. The presence of many genes that encode hypothetical proteins currently prevents us from a detailed understanding of the significance of these genes. Considering that 28% of the essential genes of a minimal bacterial genome encode

hypothetical proteins (Glass et al. 2006), these hypothetical proteins in streptococcal genomes could have a larger role.

The K_a/K_s ratio and the tree-length analysis of the species-specific genes suggest that the recently transferred genes tend to evolve faster and have more relaxed constraints than ancient genes. These results agree with the earlier results in *E. coli* and *Bacillus*, which reported that the recent transfers have relatively higher K_a/K_s ratios (Daubin and Ochman 2004; Hao and Golding 2006). Moreover, recently transferred genes also show greater evolutionary changes (both synonymous and nonsynonymous) than those of ancient genes. Perhaps many of these genes might be evolving faster as they have no functional constraint and are on their way out. An argument for some beneficial genes that have a role in adaptation can also be made. A comparison of the recently transferred genes of *S. pyogenes* with those of the ancient genes indicates that the recently transferred genes have a higher proportion (18% vs. 6%) of genes that evolve under positive selection, which suggests that at least some of these laterally acquired genes have an adaptive role. The number of genes having positively selected sites could be an underestimate as the method used to detect positive selection in this study has been shown to be conservative (Anisimova et al. 2001). The high proportion of genes under positive selection in recently transferred genes is consistent with their higher K_a/K_s ratios (fig. 5). From the results of the *S. pyogenes* study and some of the genes discussed herein, it can be speculated that at least 15–20% of the LGT have an adaptive role. However, as a majority of these LGT encode hypothetical proteins, the functional characterization of these hypothetical proteins will be critical for a complete understanding of the role of LGT in adaptation.

Supplementary Material

A supplementary table A.1 is available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

This work was supported by a Natural Sciences and Engineering Research Council of Canada grant to G.B.G.

Literature Cited

- Ajdic D, McShan WM, McLaughlin RE, et al. (19 co-authors). 2002. Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. *Proc Natl Acad Sci USA*. 99:14434–14439.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389–3402.
- Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol*. 18:1585–1592.
- Arthur M, Depardieu F, Molinas C, Reynolds P, Courvalin P. 1995. The vanZ gene of Tn1546 from *Enterococcus faecium* BM4147 confers resistance to teicoplanin. *Gene*. 154:87–92.
- Balsalobre L, Ferrandiz MJ, Linares J, Tubau F, de la Campa AG. 2003. Viridans group streptococci are donors in horizontal transfer of topoisomerase IV genes to *Streptococcus pneumoniae*. *Antimicrob Agents Chemother*. 47:2072–2081.
- Banks DJ, Porcella SF, Barbian KD, Beres SB, Phillips LE, Voyich JM, DeLeo FR, Martin JM, Somerville GA, Musser JM. 2004. Progress toward characterization of the group A *Streptococcus* metagenome: complete genome sequence of a macrolide-resistant serotype M6 strain. *J Infect Dis*. 190:727–738.
- Baptiste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF. 2005. Do orthologous gene phylogenies really support tree-thinking? *BMC Evol Biol*. 5:33.
- Beres SB, Sylva GL, Barbian KD, et al. (16 co-authors). 2002. Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc Natl Acad Sci USA*. 99:10078–10083.
- Berry AM, Lock RA, Paton JC. 1996. Cloning and characterization of nanB, a second *Streptococcus pneumoniae* neuraminidase gene, and purification of the NanB enzyme from recombinant *Escherichia coli*. *J Bacteriol*. 178:4854–4860.
- Bolotin A, Quinquis B, Renault P, et al. (23 co-authors). 2004. Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. *Nat Biotechnol*. 22:1554–1558.
- Bolotin A, Wincker P, Mauger S, Jaillon O, Malarme K, Weissenbach J, Ehrlich SD, Sorokin A. 2001. The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. *Genome Res*. 11:731–753.
- Bowden GH, Hamilton IR. 1998. Survival of oral bacteria. *Crit Rev Oral Biol Med*. 9:54–85.
- Broker G, Spellerberg B. 2004. Surface proteins of *Streptococcus agalactiae* and horizontal gene transfer. *Int J Med Microbiol*. 294:169–175.
- Camara M, Boulnois GJ, Andrew PW, Mitchell TJ. 1994. A neuraminidase from *Streptococcus pneumoniae* has the features of a surface protein. *Infect Immun*. 62:3688–3695.
- Chen YY, Weaver CA, Mendelsohn DR, Burne RA. 1998. Transcriptional regulation of the *Streptococcus salivarius* 57.I urease operon. *J Bacteriol*. 180:5769–5775.
- Cole ST, Eiglmeier K, Parkhill J, et al. (44 co-authors). 2001. Massive gene decay in leprosy bacillus. *Nature*. 409:1007–1011.
- Daubin V, Lerat E, Perriere G. 2003. The source of laterally transferred genes in bacterial genomes. *Genome Biol*. 4:R57.
- Daubin V, Moran NA, Ochman H. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science*. 301:829–832.
- Daubin V, Ochman H. 2004. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res*. 14:1036–1042.
- de Koning AP, Brinkman FS, Jones SJ, Keeling PJ. 2000. Lateral gene transfer and metabolic adaptation in the human parasite *Trichomonas vaginalis*. *Mol Biol Evol*. 17:1769–1773.
- Dobrindt U, Hacker J. 2001. Whole genome plasticity in pathogenic bacteria. *Curr Opin Microbiol*. 4:550–557.
- Fabio U, Bondi M, Manicardi G, Messi P, Neglia R. 1987. Production of bacteriocin-like substances by human oral streptococci. *Microbiologica*. 10:363–370.
- Feldgarden M, Byrd N, Cohan FM. 2003. Gradual evolution in bacteria: evidence from *Bacillus* systematics. *Microbiology*. 149:3565–3573.
- Felsenstein J. 1989. PHYLIP (phylogeny inference package). Version 3.2. *Cladistics*. 5:164–166.
- Felsenstein J. 2004. *Inferring phylogenies*. Sunderland (MA): Sinauer Associates, Inc.
- Ferretti JJ, McShan WM, Ajdic D, et al. (23 co-authors). 2001. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc Natl Acad Sci USA*. 98:4658–4663.
- Foster J, Ganatra M, Kamal I, et al. (26 co-authors). 2005. The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol*. 3:e121.
- Fox GE, Wisotzkey JD, Jurtschuk P Jr. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol*. 42:166–170.
- Franken C, Brandt C, Broker G, Spellerberg B. 2004. ISSAg1 in streptococcal strains of human and animal origin. *Int J Med Microbiol*. 294:247–254.
- Glaser P, Rusniok C, Buchrieser C, et al. (12 co-authors). 2002. Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. *Mol Microbiol*. 45:1499–1513.
- Glass JI, Assad-Garcia N, Alperovich N, Yooshep S, Lewis MR, Maruf M, Hutchison CA 3rd, Smith HO, Venter JC. 2006. Essential genes of a minimal bacterium. *Proc Natl Acad Sci USA*. 103:425–430.
- Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol*. 19:2226–2238.
- Green NM, Zhang S, Porcella SF, Nagiec MJ, Barbian KD, Beres SB, LeFebvre RB, Musser JM. 2005. Genome sequence of a serotype M28 strain of group A *Streptococcus*: potential new insights into puerperal sepsis and bacterial disease specificity. *J Infect Dis*. 192:760–770.
- Gu Z, Nicolae D, Lu HH, Li WH. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet*. 18:609–613.
- Hao W, Golding GB. 2004. Patterns of bacterial gene movement. *Mol Biol Evol*. 21:1294–1307.
- 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.
- Hols P, Hancy F, Fontaine L, et al. (14 co-authors). 2005. New insights in the molecular biology and physiology of *Streptococcus thermophilus* revealed by comparative genomics. *FEMS Microbiol Rev*. 29:435–463.
- Hoskins J, Alborn WE Jr, Arnold J, et al. (42 co-authors). 2001. Genome of the bacterium *Streptococcus pneumoniae* strain R6. *J Bacteriol*. 183:5709–5717.
- Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. 17:754–755.
- Hughes AL, Friedman R. 2005. Poxvirus genome evolution by gene gain and loss. *Mol Phylogenet Evol*. 35:186–195.
- Ivanova I, Miteva V, Stefanova TS, Pantev A, Budakov I, Danova S, Moncheva P, Nikolova I, Dousset X, Boyaval P. 1998. Characterization of a bacteriocin produced by *Streptococcus thermophilus* 81. *Int J Food Microbiol*. 42:147–158.
- Jain R, Rivera MC, Moore JE, Lake JA. 2003. Horizontal gene transfer accelerates genome innovation and evolution. *Mol Biol Evol*. 20:1598–1602.

- Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8:275–282.
- Klein DL. 1999. Pneumococcal disease and the role of conjugate vaccines. *Microb Drug Resist.* 5:147–157.
- Korithoski B, Krastel K, Cvitkovitch DG. 2005. Transport and metabolism of citrate by *Streptococcus mutans*. *J Bacteriol.* 187:4451–4456.
- Kunin V, Ouzounis CA. 2003. The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* 13:1589–1594.
- Lan R, Reeves PR. 1996. Gene transfer is a major factor in bacterial evolution. *Mol Biol Evol.* 13:47–55.
- Lawrence JG. 1997. Selfish operons and speciation by gene transfer. *Trends Microbiol.* 5:355–359.
- Lawrence JG. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. *Curr Opin Microbiol.* 2:519–523.
- Lawrence JG, Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA.* 95:9413–9417.
- Lawrence JG, Ochman H. 2002. Reconciling the many faces of lateral gene transfer. *Trends Microbiol.* 10:1–4.
- Martin-Galiano AJ, Wells JM, de la Campa AG. 2004. Relationship between codon biased genes, microarray expression values and physiological characteristics of *Streptococcus pneumoniae*. *Microbiology.* 150:2313–2325.
- 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.
- Mora D, Maguin E, Masiero M, Parini C, Ricci G, Manachini PL, Daffonchio D. 2004. Characterization of urease genes cluster of *Streptococcus thermophilus*. *J Appl Microbiol.* 96:209–219.
- Mora D, Monnet C, Parini C, Guglielmetti S, Mariani A, Pintus P, Molinari F, Daffonchio D, Manachini PL. 2005. Urease biogenesis in *Streptococcus thermophilus*. *Res Microbiol.* 156:897–903.
- Nakagawa I, Kurokawa K, Yamashita A, et al. (13 co-authors). 2003. Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res.* 13:1042–1055.
- Nizet V, Beall B, Bast DJ, Datta V, Kilburn L, Low DE, De Azavedo JC. 2000. Genetic locus for streptolysin S production by group A *Streptococcus*. *Infect Immun.* 68:4245–4254.
- Ochman H, Davalos LM. 2006. The nature and dynamics of bacterial genomes. *Science.* 311:1730–1733.
- Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature.* 405:299–304.
- Ogata H, Audic S, Renesto-Audiffren P, et al. (11 co-authors). 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science.* 293:2093–2098.
- Pal C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet.* 37:1372–1375.
- Pushker R, Mira A, Rodriguez-Valera F. 2004. Comparative genomics of gene-family size in closely related bacteria. *Genome Biol.* 5:R27.
- Read TD, Myers GSA, Brunham RC, et al. (21 co-authors). 2003. Genome sequence of *Chlamydomophila caviae* (*Chlamydia psittaci* GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. *Nucleic Acids Res.* 31:2134–2147.
- Ricard G, McEwan NR, Dutilh BE, et al. (17 co-authors). 2006. Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic carbohydrates rich environment. *BMC Genomics.* 7:22.
- Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–422.
- Santos SR, Ochman H. 2004. Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environ Microbiol.* 6:754–759.
- Sayle RA, Milner-White EJ. 1995. RASMOL: biomolecular graphics for all. *Trends Biochem Sci.* 20:374.
- Shakhnovich EA, King SJ, Weiser JN. 2002. Neuraminidase expressed by *Streptococcus pneumoniae* desialylates the lipopolysaccharide of *Neisseria meningitidis* and *Haemophilus influenzae*: a paradigm for interbacterial competition among pathogens of the human respiratory tract. *Infect Immun.* 70:7161–7164.
- Smoot JC, Barbian KD, Van Gompel JJ, et al. (18 co-authors). 2002. Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci USA.* 99:4668–4673.
- Snel B, Huynen MA, Dutilh BE. 2005. Genome trees and the nature of genome evolution. *Annu Rev Microbiol.* 59:191–209.
- Sokurenko EV, Chesnokova V, Dykhuizen DE, Ofek I, Wu XR, Krogfelt KA, Struve C, Schembri MA, Hasty DL. 1998. Pathogenic adaptation of *Escherichia coli* by natural variation of the FimH adhesin. *Proc Natl Acad Sci USA.* 95:8922–8926.
- Spellerberg B, Martin S, Brandt C, Luticken R. 2000. The cyl genes of *Streptococcus agalactiae* are involved in the production of pigment. *FEMS Microbiol Lett.* 188:125–128.
- Spellerberg B, Pohl B, Haase G, Martin S, Weber-Heynemann J, Luticken R. 1999. Identification of genetic determinants for the hemolytic activity of *Streptococcus agalactiae* by ISS1 transposition. *J Bacteriol.* 181:3212–3219.
- Springael D, Top EM. 2004. Horizontal gene transfer and microbial adaptation to xenobiotics: new types of mobile genetic elements and lessons from ecological studies. *Trends Microbiol.* 12:53–58.
- Strimmer K, von Haeseler A. 1996. Quartet puzzling: a quartet maximum-likelihood for reconstructing tree topologies. *Mol Biol Evol.* 13:964–969.
- Sumbly P, Porcella SF, Madrigal AG, et al. (11 co-authors). 2005. Evolutionary origin and emergence of a highly successful clone of serotype M1 group A *Streptococcus* involved multiple horizontal gene transfer events. *J Infect Dis.* 192:771–782.
- Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28:33–36.
- Tettelin H, Massignani V, Cieslewicz MJ, et al. (43 co-authors). 2002. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proc Natl Acad Sci USA.* 99:12391–12396.
- Tettelin H, Nelson KE, Paulsen IT, et al. (39 co-authors). 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science.* 293:498–506.
- Thompson JD, Higgins DG, Gibson TJ. 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.
- Tong HH, Blue LE, James MA, DeMaria TF. 2000. Evaluation of the virulence of a *Streptococcus pneumoniae* neuraminidase-deficient mutant in nasopharyngeal colonization and development of otitis media in the chinchilla model. *Infect Immun.* 68:921–924.

- Torda AE, Procter JB, Huber T. 2004. Wurst: a protein threading server with a structural scoring function, sequence profiles and optimized substitution matrices. *Nucleic Acids Res.* 32:W532–W535.
- Towers RJ, Gal D, McMillan D, Sriprakash KS, Currie BJ, Walker MJ, Chhatwal GS, Fagan PK. 2004. Fibronectin-binding protein gene recombination and horizontal transfer between group A and G streptococci. *J Clin Microbiol.* 42:5357–5361.
- Upton M, Tagg JR, Wescombe P, Jenkinson HF. 2001. Intra- and interspecies signaling between *Streptococcus salivarius* and *Streptococcus pyogenes* mediated by SalA and SalA1 lantibiotic peptides. *J Bacteriol.* 183:3931–3938.
- Vadeboncoeur C, Pelletier M. 1997. The phosphoenolpyruvate:sugar phosphotransferase system of oral streptococci and its role in the control of sugar metabolism. *FEMS Microbiol Rev.* 19:187–207.
- Ward DJ, Somkuti GA. 1995. Characterization of a bacteriocin produced by *Streptococcus thermophilus* ST134. *Appl Microbiol Biotechnol.* 43:330–335.
- Wiezer A, Merkl R. 2005. A comparative categorization of gene flux in diverse microbial species. *Genomics.* 86:462–475.
- Wren BW. 2000. Microbial genome analysis: insights into virulence, host adaptation and evolution. *Nat Rev Genet.* 1:30–39.
- Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13:555–556.
- Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics.* 155:431–449.
- Yother J, Trieu-Cuot P, Klaenhammer TR, De Vos WM. 2002. Genetics of streptococci, lactococci, and enterococci: review of the sixth international conference. *J Bacteriol.* 184:6085–6092.
- Zhang P, Gu Z, Li WH. 2003. Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol.* 4:R56.

Laura Katz, Associate Editor

Accepted September 1, 2006