The fate of laterally transferred genes: Life in the fast lane to adaptation or death

Weilong Hao and G. Brian Golding[1]

Department of Biology, McMaster University, Hamilton, Ontario, Canada L8S 4K1

Large-scale genome arrangement plays an important role in bacterial genome evolution. A substantial number of genes can be inserted into, deleted from, or rearranged within genomes during evolution. Detecting or inferring gene insertions/deletions is of interest because such information provides insights into bacterial genome evolution and speciation. However, efficient inference of genome events is difficult because genome comparisons alone do not generally supply enough information to distinguish insertions, deletions, and other rearrangements. In this study, homologous genes from the complete genomes of 13 closely related bacteria were examined. The presence or absence of genes from each genome was cataloged, and a maximum likelihood method was used to infer insertion/deletion rates according to the phylogenetic history of the taxa. It was found that whole gene insertions/deletions in genomes occur at rates comparable to or greater than the rate of nucleotide substitution and that higher insertion/deletion rates are often inferred to be present at the tips of the phylogeny with lower rates on more ancient interior branches. Recently transferred genes are under faster and relaxed evolution compared with more ancient genes. Together, this implies that many of the lineage-specific insertions are lost quickly during evolution and that perhaps a few of the genes inserted by lateral transfer are niche specific.

[Supplemental material is available online at www.genome.org.]

Gene insertions and deletions, together with gene inversions and translocations, play important roles in shaping bacterial genomes (Itaya 1997; Brunder and Karch 2000; Tillier and Collins 2000; Liu et al. 2002; Kuwahara et al. 2004; Cerdeno-Tarraga et al. 2005), and gene insertions and deletions, in particular, are essential driving forces that influence gene content (Ochman and Jones 2000; Kunin and Ouzounis 2003; Mirkin et al. 2003). It is clear that a large number of insertions/deletions can be observed in many bacterial species (Mirkin et al. 2003; Hao and Golding 2004) even though they may be comparatively rare in some endosymbiotic bacteria (Silva et al. 2003). The portion of insertions and deletions in a genome, therefore, varies among different species (Garcia-Vallvé et al. 2000).

Gene insertions and deletions can be inferred by examining the presence or absence of a gene (or a gene family) on a phylogenetic tree. In some recent studies, the parsimony method has been used to infer insertions/deletions (Daubin et al. 2003a,b; Mirkin et al. 2003; Hao and Golding 2004). Gene insertions have been distinguished as gene genesis (birth) or lateral gene transfers (LGT), and insertions/deletions have been tested with varying penalties for LGTs in different methodologies (Snel et al. 2002; Kunin and Ouzounis 2003; McLysaght et al. 2003). However, the inference of insertions/deletions is difficult because of the possibility of parallel deletions and insertions on multiple branches (Copley and Dhillon 2002; Snel et al. 2002; Stoebel 2005) and because of variable evolutionary rates of change on different branches (Hao and Golding 2004). Furthermore, the parsimony method is well known to underestimate the number of events in phylogeny reconstruction (Galtier and Boursot 2000; Dean et al. 2002; Felsenstein 2004).

Likelihood analysis has been successfully used to reconstruct phylogenies using sequence data since its first application by Neyman (Neyman 1971; Felsenstein 1988, 1989, 2004; Gu 2001). Maximum likelihood analyses have also been applied to the study of genome content (Gu and Zhang 2004; Huson and Steel 2004), and the phyletic pattern of gene presence/absence has been used to reconstruct evolutionary history in a Markov analysis (Lake and Rivera 2004). In this study, a maximum likelihood method is used to infer insertion/deletion rates on the phylogeny of the *Bacillaceae* group of Gram-positive bacteria. An advantage of this group is the large number of genomes that have been completely sequenced.

For the likelihood analysis, the insertion rate was assumed to be equal to the deletion rate on each branch, but insertion/deletion rates could vary among different branches or in different parts of the phylogeny. These results suggest that recently transferred genes are more common. If this is to be an evolutionarily stable situation, it suggests that many laterally transferred genes are more likely to have a high propensity of being deleted quickly after transfer. The rates of insertion/deletion from the maximum likelihood analysis were compared to observed nucleotide substitution rates and found to be of a comparable or larger rate; the rates inferred increase at the tips of the phylogeny.

Results

The maximum likelihood analysis used the phylogeny of concatenated DNA sequences from the genes *gmk*, *glpF*, and *pycA* (Fig. 1) and inferred the relative insertion/deletion rates by assuming that individual insertion and deletion events occur independently (this model also assumes that genes can be regained multiple times after having been deleted). Initially, a single constant insertion/deletion rate was assumed on the phylogeny (Case 1 in Fig. 2) using the observed gene presence/absence patterns (Table

[1] **Corresponding author.**
E-mail Golding@McMaster.CA; fax (905) 522-6066.

Article is online at http://www.genome.org/cgi/doi/10.1101/gr.4746406. Freely available through the *Genome Research* Open Access option.

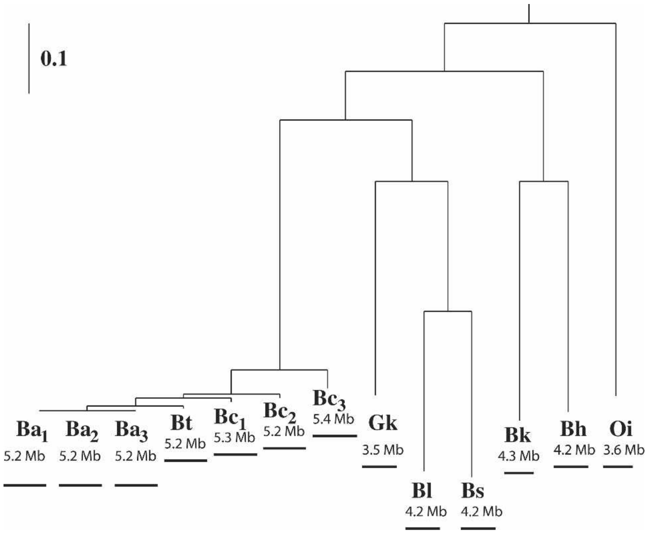

Figure 1. Bayesian tree obtained for the concatenated sequence of *gmk*, *glpF*, and *pycA* genes. The abbreviations are (Ba$_1$) *Bacillus anthracis* Ames; (Ba$_2$) *Bacillus anthracis* "Ames Ancestor"; (Ba$_3$) *Bacillus anthracis* Sterne; (Bt) *Bacillus thuringiensis*; (Bc$_1$) *Bacillus cereus* ZK; (Bc$_2$) *Bacillus cereus* ATCC 10,987; (Bc$_3$) *Bacillus cereus* ATCC 14,579; (Gk) *Geobacillus kaustophilus*; (Bl) *Bacillus licheniformis*; (Bs) *Bacillus subtilis*; (Bk) *Bacillus clausii*; (Bh) *Bacillus halodurans*; and (Oi) *Oceanobacillus iheyensis*.

Table 1. The phyletic patterns with most genes in all members of the *Bacillaceae* group

Number of genes	Ba$_1$	Ba$_2$	Ba$_3$	Bt	Bc$_1$	Bc$_2$	Bc$_3$	Gk	Bl	Bs	Bk	Bh	Oi
1139	1	1	1	1	1	1	1	1	1	1	1	1	1
1024	1	1	1	1	1	1	1	0	0	0	0	0	0
285	0	0	0	0	0	0	0	0	1	1	0	0	0
194	1	1	1	0	0	0	0	0	0	0	0	0	0
156	0	0	0	0	0	0	0	1	0	0	0	0	0
148	0	0	0	0	0	0	0	0	0	0	0	1	0
132	1	1	1	1	1	1	0	0	0	0	0	0	0
128	0	0	0	0	0	0	0	0	0	0	1	0	0
119	0	0	0	0	0	0	0	0	0	1	0	0	0
118	0	0	0	0	0	0	0	0	0	0	0	0	1
109	1	1	1	1	1	0	1	0	0	0	0	0	0
103	1	1	1	0	1	0	0	0	0	0	0	0	0
99	1	1	1	1	1	0	0	0	0	0	0	0	0
96	1	1	1	1	1	1	1	1	1	1	1	1	0
90	1	1	1	1	1	1	1	0	1	1	1	1	1
85	0	0	0	0	0	0	0	0	1	0	0	0	0
3203	Other patterns												

The taxa abbreviations are shown in Figure 1.

1). The likelihood function for this model is a simple continuous function for the chosen parameter ranges, and there is a smooth continuous change of the likelihood as the insertion/deletion rates change (Fig. 3). The insertion/deletion rate that gives the maximum likelihood value in this case is 0.51.

The strains from *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis* are closely related and have been suggested to form the *B. cereus* group (the Bc group) (Priest et al. 2004). Two different insertion/deletion rates were assumed on the two parts of the phylogeny (Case 2 in Fig. 2). The branches among the Bc group have an insertion/deletion rate α, while the remaining branches have an insertion/deletion rate β. Fitting the maximum likelihood model to this scenario suggests that the rate α is 4.42, while the rate β is 0.35 (Fig. 3). It is striking that the rate α is much greater than the rate β and also much larger than 1, which is the evolutionary time period required to observe one substitution per nucleotide site. Hence, during the evolutionary time period required for one substitution per site, an entire gene could possibly have been inserted/deleted about five times in the Bc group. There are more gene movements observed relative to evolutionary branch length by comparing more closely related strains, and hence, there are more gene movements inferred at the tips of phylogeny.

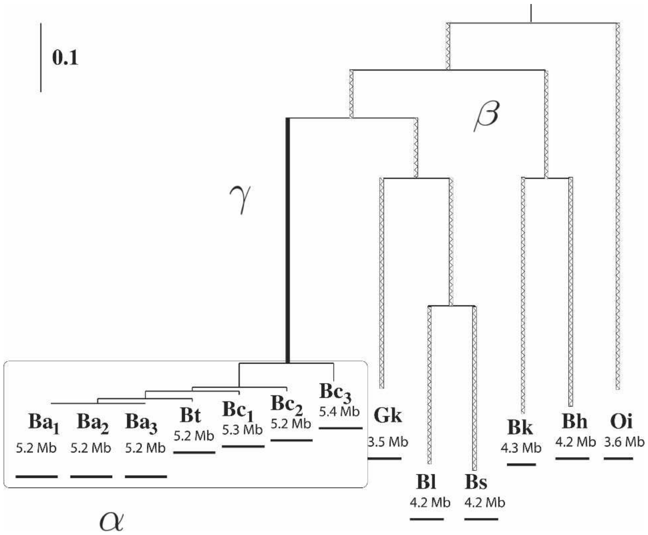

Figure 2. Different insertion/deletion rates were assumed on the phylogeny. The rate on the branches of the Bc group is α (boxed), the rate on the branch leading to the Bc group is γ (in black), the rate on the remaining branches is β (hatched). Case 1: a single constant rate throughout the phylogeny (α = β = γ). Case 2: two rates differentiate the Bc group (α, β = γ). Case 3: three rates differentiate the Bc group and the branch leading to this group (α, β, γ).

As can be observed in Figure 2, there is a divergence in the genome size between the Bc group and the genome size of the other *Bacillaceae*. The branch leading to the Bc group was therefore separated from the other branches with a distinct rate γ (Case 3 in Fig. 2). The application of the maximum likelihood analysis to this model indicates that the rate α is 3.92, the rate β is 0.28, and the rate γ is 1.23. The rate γ is approximately five times greater than the rate β. However, the rate α among the Bc group measuring rates closer to the tips of the phylogeny is still much greater than either the rate β or γ.

In addition to the above likelihood analysis, an analysis was performed with the assumption that genes cannot be regained after having been deleted. These results yield insertion/deletion rates that are similar to those under the assumption that genes can be regained after deletion (Table 2). However, in every case, the likelihood value is much lower. With a single constant insertion/deletion rate, the MLE is 0.48. In the case of two separate rates, the rate α is 4.48 among the Bc group and the rate β is 0.33. When the rate on the branch leading to the Bc group was also considered distinct, the rate γ was estimated at 1.08 and the rate β as 0.28. Both rates β and γ are much smaller than the rate α at 3.90.

Finally, the rate on internal branches in the Bc group was separated from that on external branches in the likelihood estimation (boxed portion of the phylogeny in Fig. 2). This shows

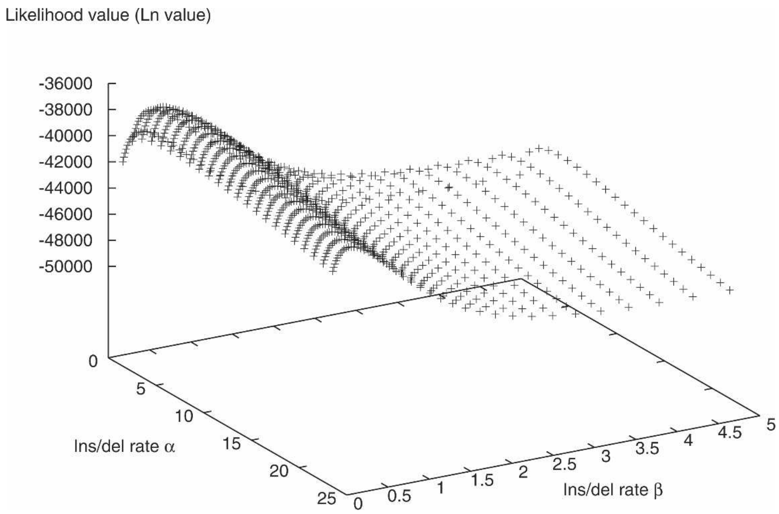

Figure 3. An example of the likelihood surface with different rates of insertion and deletion. The likelihood value shows a continuous pattern of change as the insertion/deletion rate changes.

that the insertion/deletion rate on external branches is higher than that on internal branches under both models of gene transfer (4.42 vs. 3.62 and 5.34 vs. 2.78, respectively) (Table 3). If the branch leading to the three *B. anthracis* strains is treated as an external branch in the likelihood estimation because of the close evolutionary relationship of the *B. anthracis* strains, the rate difference between external branches and internal branches becomes more dramatic (7.19 vs. 0.91 and 7.63 vs. 0.81, respectively) (Table 3). This again confirms that more gene insertions/deletions take place at the tips of the phylogeny.

To explore among the most closely related taxa, five strains from the Bc group; *B. anthracis* Ames (Ba_1), *B. anthracis* Ames "ancestor" (Ba_2), *B. anthracis* Sterne (Ba_3), *B. thuringiensis* (Bt), and *B. cereus* ZK (Bc_1) were analyzed separately. The comparison of homologs shows that >96% of the genes present in all five strains share at least 90% sequence identity with each other in their protein sequences. Therefore, the substitutions between homologs among these five strains should be considered as relatively limited. All phyletic patterns of these five strains are shown in Table 4. Of the 5076 gene families, there are only 3956 present in all five strains. Hence, 22.1% of the genes are not shared by all five strains, even though these five strains are believed to represent one species (Helgason et al. 2000) and to have diverged very recently.

To determine the rates of evolution in the recently transferred genes, the tree lengths for the Bc-group-specific genes were measured (Fig. 4A) and compared with the tree lengths of genes present in all 13 strains (Fig. 4B). In both cases, only the branch lengths within the Bc group of taxa were measured. This comparison indicates that the genes that are strain specific within the Bc group have much faster rates of evolution than do more ancestral genes present in the other taxa.

The rates of nonsynonymous (K_a) and synonymous (K_s) substitutions were estimated for the genes present only within the Bc group and compared to genes that are more broadly distributed within the *Bacillaceae* group. Again, only changes that have occurred within the Bc group are measured with the genes categorized by their breadth of distribution. Both the K_s and K_a rates are elevated in the Bc group (Supplemental material), but the K_a values are most strongly affected. The K_a/K_s ratios for genes limited to the Bc group are shown in Figure 5A. It is clear that the Bc-group-specific genes have elevated K_a/K_s ratios. Genes present only within the Bc group have larger K_a/K_s ratios than genes present within the Bc group, *Geobacillus kaustophilus*, *Bacillus licheniformis*, and *Bacillus subtilis* or within the Bc group, *G. kaustophilus*, *B. licheniformis*, *B. subtilis*, *Bacillus clausii*, and *Bacillus halodurans*, or in all taxa (Fig. 5, A vs. B,C,D). The genes more recently transferred appear to contribute to a higher K_a/K_s ratio.

Discussion

To determine the patterns of LGT, it is useful to examine closely related but fully sequenced genomes. A complete genome sequence is necessary to eliminate the possibility of a hidden paralog or of a genome rearrangement masking a homolog. Closely related taxa help to determine the number of genes that might have been laterally transferred. To this end, we have examined the gene content from 13 completely sequenced genomes from the *Bacillaceae* group.

The results demonstrate that LGT occurs rapidly and extensively between strains of the same species. A phylogeny was constructed to measure the rate of LGT relative to nucleotide substitutions. The concatenated DNA sequences of *gmk*, *glpF*, and *pycA* genes rather than ribosomal RNA sequences were used to reconstruct the phylogeny in this study. It is difficult to reconstruct the phylogenetic relationship within the Bc group owing to their

Table 2. Insertion/deletion rates inferred from the maximum likelihood analysis using different models as shown in Figure 2

	Reversible		Deleted once[a]	
Rate	MLE	LnL	MLE	LnL
$\alpha = \beta = \gamma$	0.51	−40277	0.48	−40716
α	4.42		4.48	
$\beta = \gamma$	0.35	−36902	0.33	−37281
α	3.92		3.90	
β	0.28	−36128	0.28	−36592
γ	1.23		1.08	

[a] Genes cannot be regained after deletion.

Table 3. Insertion/deletion rates inferred on external branches (α_1) and internal branches (α_2) in the Bc group (boxed in Fig. 2)

	Reversible		Deleted once[a]	
Rate	MLE	LnL	MLE	LnL
$\alpha_1 = \alpha_2$	4.10	−11350	3.80	−11600
α_1	4.42		5.34	
α_2	3.62	−11342	2.78	−11525
α_1[a]	7.19		7.63	
α_2	0.91	−10958	0.81	−11013

[a] The branch leading to Ba_1, Ba_2, and Ba_3 is treated as an external branch.

Table 4. **The phyletic patterns in the five most closely related strains of the Bc group**

Number of genes	Ba_1	Ba_2	Ba_3	Bt	Bc_1
3956	1	1	1	1	1
332	1	1	1	0	0
200	0	0	0	1	1
199	1	1	1	0	1
136	1	1	1	1	0
125	0	0	0	1	0
98	0	0	0	0	1
7	1	1	0	1	1
7	1	1	0	0	0
3	0	1	1	1	1
3	0	0	1	1	1
3	1	0	1	1	0
3	0	0	1	0	0
1	1	1	0	1	0
1	1	0	1	0	0
1	0	1	1	1	0
1	0	1	0	1	0
5076			Total		

remarkably similar rRNA sequences (Ash et al. 1991) and the divergence of rRNA sequences in genomes with multiple *rrn* operons (Klappenbach et al. 2001; Acinas et al. 2004). The *gmk*, *glpF*, *pycA*, *tpi*, *ilvD*, *pta*, and *pur* genes have been studied in the past as tools for reconstructing the evolutionary history of the *B. cereus* group. It was found that *gmk*, *glpF*, *pycA*, and *tpi* genes strongly conform to the concatenated tree of all seven genes (Priest et al. 2004). (Note that the topological position of *Oceanobacillus iheyensis* in the concatenated phylogeny is different from that in a 16S rRNA-based phylogeny [Hao and Golding 2004].)

When maximum likelihood estimates of the rates of insertion/deletion are mapped onto this phylogeny, it suggests that there are more genes coming in and going out at the tips of phylogeny. This is clear even if one looks at the table of gene presence/absence (Table 1) and observes that differences in gene content between taxa that are considered a single species are among the most common patterns observed. If this is an evolutionarily stable situation, then most of the laterally transferred genes must be lost shortly after their insertion during evolution. Genome annotation can be an error-prone task (Kyrpides and Ouzounis 1999). As a result, all of the predicted ORFs that are present in only one genome and that do not have homologs detectable by BLAST were removed from this study. Since many of these may be proper and functional genes (Siew and Fischer 2003, 2004), this method tends to further underestimate the events on external branches. Not surprisingly, a maximum likelihood estimation including the uniquely present ORFs further inflates the rates at the tips of phylogeny (data not shown).

It has been suggested that *B. anthracis*, *B. cereus*, and *B. thuringiensis* are one species (Helgason et al. 2000). A close evolutionary relationship among these strains is inferred by comparing substitutions in the concatenated sequences (Fig. 1) and from the protein similarity among five of the most closely related members from the Bc group. On the other hand, >22% of the genes are not present in all five strains (Table 4). Together this shows that gene insertions/deletions have a role in the evolution of these strains equal to or greater than the role of changes at the sequence level. This compares favorably to the results of Lynch and Conery (2003) that the rate of gene duplications is also of the same order of magnitude as base substitutions. The maximum likelihood analysis within the Bc group was performed by optimizing separate insertion/deletion rates on internal branches and external branches. It shows that the tips of the phylogeny have greater insertion/deletion rates (Table 3). Nor can the high insertion/deletion rates in these genomes be simply explained by a highly mobile feature of genomes such as pathogenicity islands. Although three pathogenicity/genomic islands have been reported in *B. cereus* ATCC 14,579 (Bc_3) in a previous study (Zhang and Zhang 2003), most of the annotated ORFs in these islands have been excluded from this study since they do not have homologs in completed genomes from other bacteria. Additionally, while the Bc_3 strain-specific genes in this study are not located in these three islands, they too are excluded as they lack known homologs in other species. It is plausible that the ORFs in the strain-specific pathogenicity islands, if they are real genes and not annotation errors or pseudogenes, might contribute to an even higher turnover rate than that found in this study. Indeed, it has been shown that islands can have high numbers of insertions, deletions, and pseudogenes (Ullrich et al. 2005). High rates on external branches are observed in this study, and therefore, the addition of strain-specific genes or unique genes on

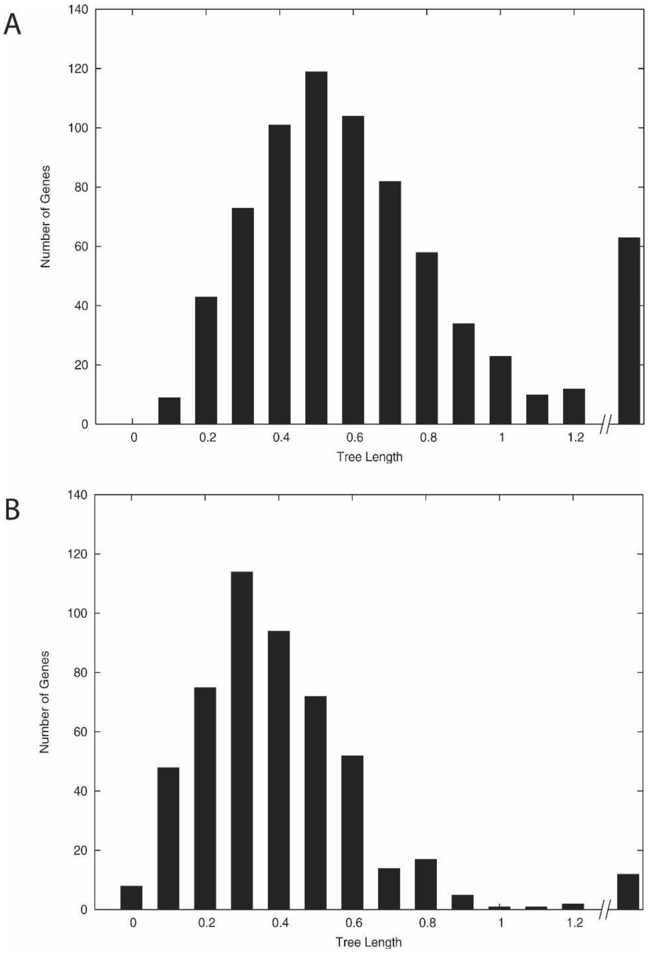

Figure 4. Rapid evolution in the Bc group specific genes. (*A*) Tree length for the Bc taxa as indicated by genes that are present only within this group of taxa; (*B*) tree length for the Bc taxa as indicated by genes that are present in all 13 strains.

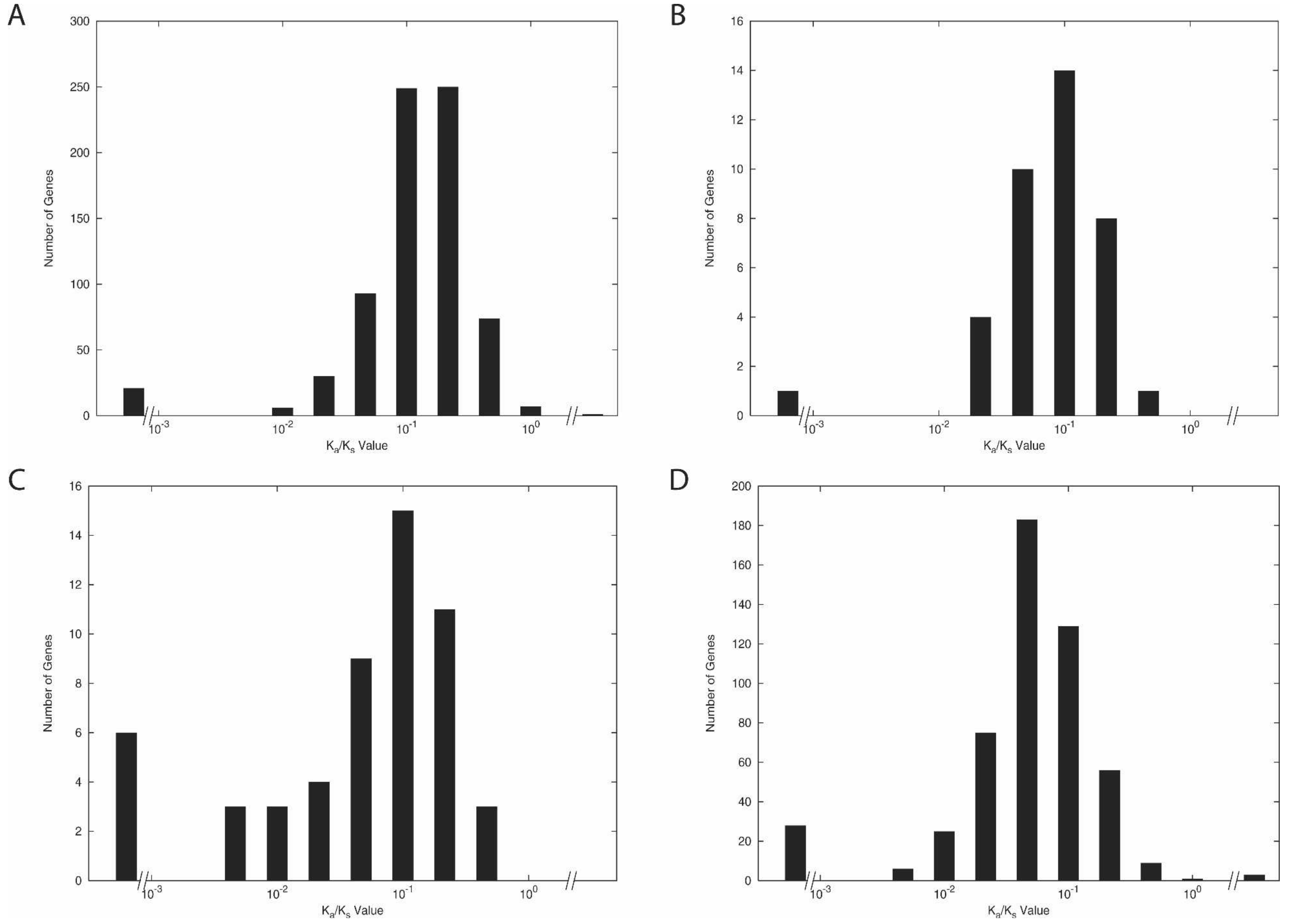

Figure 5. K_a/K_s ratio of different group-specific genes. Data are measured using the genes present in the Bc group (Ba_1, Bt, and Bc_1). (*A*) Genes present in the Bc group; (*B*) Genes present in Bc, Gk, Bl, and Bs; (*C*) genes present in Bc, Gk, Bl, Bs, Bk, and Bh; (*D*) genes present in all 13 strains.

pathogenicity islands would further inflate the rate at the tips of phylogeny.

In the maximum likelihood estimation, the insertion rate is assumed to be equal to the deletion rate. This assumption was made to ensure that in the long term, genome sizes would not tend to zero or infinity. In the short term, this assumption is unlikely to be correct, and Thompson et al. (2005) have shown that even within closely related bacterial populations, the genome content may change. The model of gene insertion/deletion can be improved by assuming unequal rates of insertion/deletion, fixed numbers of genes that are not deleted, variation in indel rate among genes, and so on. But the increased rate at the tips of the phylogeny is not likely to be an artifact of genome size variation or the limitations of the likelihood model. Firstly, the members of the Bc group have larger genome size than the non-Bc group taxa, but, while the indel rate on the branch leading to the Bc group shows a higher rate, it is still much smaller than the rate within the Bc group. Secondly, the seven members of the Bc group have similar genome sizes, and an estimation using only the members of the Bc group again shows (Table 3) higher indel rates at the tip of phylogeny. Genome size variation is well known as a problem in genome content phylogeny reconstruction, but it has been noted that phylogeny, rather than phenotype or LGT, is the major quantitative determinant of gene content (Snel et al. 1999).

The more recently transferred genes have longer tree length (with $P < 0.001$ in a Wilcoxon rank test) (Fig. 4), suggesting that the recently transferred genes are evolving faster than ancient genes. The K_a/K_s ratio study suggests that, in general, more recently transferred genes have less functional constraints (Fig. 5). The different subfigures show the K_a/K_s ratio in genes with increasing depth in the phylogeny. Those genes that are inferred to have arisen recently via LGT in the phylogeny have a higher ratio than those that were transferred somewhat more distantly (Fig. 5, A vs. B is $P \approx 0.003954$). As the length of time that the genes have been resident within the host increases, the K_a/K_s ratio continues to go down (Fig. 5, B vs. D is $P \approx 0.001618$). Hence, genes that have been recently transferred into these hosts have a higher percentage of nonsynonymous substitutions. Together, the recently transferred genes not only have faster evolutionary change but also have higher K_a/K_s ratios. Similarly, a previous study reported that orphans in *Escherichia coli* have relatively high K_a/K_s ratios (Daubin and Ochman 2004).

This study demonstrates that more recently transferred genes are under relaxed and faster evolution compared with the genes that have had a longer residence time. There are several possible reasons for this. It is possible that the laterally transferred genes with a higher rate are more prone to being laterally transferred. This is unlikely, as genes with a slightly longer residence time should not show the observed reduced rates of evolution. It has also been suggested that genes inserted into a new host will undergo amelioration of their sequence (Lawrence and Ochman 1997). In this process, the codon and base content bias of a new gene will mutate to more closely resemble the inherent bias within the new host. It is therefore plausible to explain the higher rates of evolution in recently transferred genes compared to the completely ameliorated/native genes (Fig. 4). It is more difficult to see how amelioration could also cause an increase in the nonsynonymous/synonymous ratio. Amelioration is unlikely to be the main force increasing the K_a/K_s ratio in recently transferred genes since nonsynonymous sites should be affected more slowly than synonymous sites. In addition, the ratio hides the fact that the K_a value itself is much larger in many recently transferred genes (Supplemental material), again suggesting that simple amelioration is unlikely to be the cause.

Alternatively, the genes that have been recently transferred might be adapting to a new and local environment found in the new host. In this regard, it should be noted that several genes have a very large K_a/K_s ratio, suggesting directional selection. The recently transferred genes might also be evolving quickly as they are not required in their new hosts, offer minimal selective advantage, and could be in the process of being lost. This is in accord with the observation that genes come and go rapidly within closely related genomes. It is also in concordance with the very high tree lengths in Figure 4 that suggest genes change very rapidly. A recent study concluded that transferred genes are adaptive to specific environments (Pal et al. 2005) and not deleterious or nonfunctional as suggested previously (Kurland et al. 2003; Taoka et al. 2004). The data presented here suggests that the true story is probably a mixture of directional selection to adapt in some genes and neutral mutations destroying function in others. Some recently transferred genes probably only function within the organism's current niche and will be deleted upon a change of the organism's niche. In this way, gene movement in bacteria should be considered as a result of the dynamic process of passive uptake and of quick selection on those genes appropriate for the current bacterial niche or their rapid elimination.

Methods

To gain a better concept of genome evolution in closely related bacteria, a group of bacteria with an abundance of completely sequenced congeneric species was selected. Thirteen complete *Bacillaceae* genome sequences were obtained from NCBI (http://www.ncbi.nlm.nih.gov/) to carry out the analysis. They are *B. anthracis* Ames, *B. anthracis* "Ames ancestor," *B. anthracis* Sterne, *B. thuringiensis*, *B. cereus* ZK, *B. cereus* ATCC 10,987, *B. cereus* ATCC 14,579, *Geobacillus kaustophilus*, *B. licheniformis*, *B. subtilis*, *B. clausii*, *B. halodurans*, and *Oceanobacillus iheyensis*. It has been argued that *B. anthracis*, *B. cereus*, and *B. thuringiensis* might be one species (Helgason et al. 2000) as the ribosomal RNA sequences from these strains are remarkably similar (Ash et al. 1991). Therefore, the seven strains from *B. anthracis*, *B. cereus*, and *B. thuringiensis* are hereafter referred to as the *B. cereus* group (the Bc group) as suggested by Priest et al. (2004).

The evolutionary history of the Bc group has been reconstructed using the nucleotide sequences of the *gmk*, *glpF*, and *pycA* genes (Priest et al. 2004) because the rRNA sequences are too similar to provide reliable evolutionary relationships. In this study, the concatenated DNA sequences of these three genes from each genome were used to reconstruct a phylogeny using Mr.Bayes (Huelsenbeck and Ronquist 2001) (200,000 generations sampled every 100 generations with a gamma distribution model and invariant class). The method to identify members of a gene family has been described in Hao and Golding (2004). In short, potential homologs were measured according to sequence similarities, and all paralogs in each genome were clustered as a single gene family and only one member was retained for further analysis. Non-annotated proteins in a genome were identified by carrying out a TBLASTN search against the DNA sequence of each genome and using all annotated proteins from other *Bacillaceae* genomes as query sequences. Potential taxon species genes are determined by searching against all completed bacterial genomes. (The GenBank accession numbers are given as Supplemental information at http://evol.biology.mcmaster.ca/~weilong/likelihood.) The phyletic patterns (gene presence or absence in each genome) of all genes were used for the maximum likelihood analysis.

The gene families present in the Bc group were used to conduct tree length and K_a/K_s ratio (ω) analyses using the PAML package (Yang 1997). The tree length was calculated as the sum of the branch lengths for the taxa only with the Bc 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. Genes were categorized into four groups based on their presence/absence in different taxa (and hence on the inferred time period when the genes were transferred). The four groups are characterized by genes present only in the Bc group; genes present in Bc, Gk, Bl, and Bs; genes present in Bc, Gk, Bl, Bs, Bk, and Bh; and genes present in all 13 taxa. A single K_a/K_s ratio was assumed throughout the length of each sequence in this study. To avoid the effects of duplication during evolution (Gu et al. 2002; Zhang et al. 2003), any paralogs of gene families were excluded from the tree length and K_a/K_s ratio analyses. Protein sequences and their corresponding DNA sequences were extracted from the annotated genomes. 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.

To evaluate the likelihood of the observed phyletic patterns, a simple model of gene evolution was chosen. This model assumes that individual genes are inserted or deleted at constant rates. This model does not include a consideration of increasing or decreasing numbers of genes but, rather, a constant number of gene places that may or may not be occupied at any one time. All events are assumed to be independent. Let ν be the rate of gene insertion, and let μ be the rate of gene deletion. Let t be the length of time that separates a taxon from its ancestor. Let P indicate the gene presence, and let A indicate its absence. Then the probability of gene presence in the descendant taxon (d) can be calculated given knowledge of the state in the ancestral taxon (a). Thus,

$$Prob(P_d|P_a,t) = \nu/(\mu + \nu) + e^{-(\mu+\nu)t}[1 - \nu/(\mu + \nu)],$$

$$Prob(A_d|P_a,t) = \mu/(\mu + \nu) - e^{-(\mu+\nu)t}[1 - \nu/(\mu + \nu)],$$

$$Prob(P_d|A_a,t) = \nu/(\mu + \nu) - e^{-(\mu+\nu)t}[1 - \mu/(\mu + \nu)],$$

$$Prob(A_d|A_a,t) = \mu/(\mu + \nu) + e^{-(\mu+\nu)t}[1 - \mu/(\mu + \nu)],$$

The likelihood that gene i is present at node x, with descendants y and z separated by t_1 and t_2 generations, is then calculated as

$$L_x^i(P) = (Prob(P_d|P_a,t_1) * L_y^i(P) + Prob(A_d|P_a,t_1) * L_y^i(A)) \\ \times (Prob(P_d|P_a,t_2) * L_z^i(P) + Prob(A_d|P_a,t_2) * L_z^i(A)).$$

For a given gene phyletic pattern *i*, we assume that ancestral gene presence and absence at the root of the phylogeny contribute equally to the probability of the observed pattern. Thus, the likelihood of the gene phyletic pattern *i* at the last common ancestral node *x* will be

$$Q^i = (L_x^i(P) + L_x^i(A))/2.$$

The results must also be corrected for missing data. Those genes absent in all of the taxa are unobservable. Following Felsenstein (1992), we have made a correction for the missing data in the same way as was used for missing restriction sites, and the results then are conditional on observing the gene present in at least one species. This is

$$Q_+^i = \frac{Q^i}{1 - Q_-^i}.$$

where Q^i_- is the likelihood of gene *i* being absent in all taxa. For a given phylogenetic relationship, Q^i_- will have the same value for all *i*.

All observed patterns were used to calculate the overall likelihood at the last common ancestral node by multiplying individual likelihoods together. The overall likelihood for a total of *n* patterns will be

$$Q = \prod_{i=1}^{n} Q_+^i,$$

and after taking logs,

$$\log Q = \sum_{i=1}^{n} \log Q_+^i.$$

The deletion rate μ and the insertion rate ν were assumed to be equal ($\mu = \nu$) on each individual branch, but the overall rate of indels could vary among branches of the phylogeny. Hence, while the insertion and deletion rates were always equal, they could vary in magnitude on each branch. To estimate the maximum likelihood, the branch specific rates were optimized to find those rates that maximized the likelihood of observing the gene patterns (Table 1).

Acknowledgments

This work was supported by an NSERC grant to G.B.G. The authors wish to thank R. Morton for his suggestions on earlier versions of this manuscript and to thank the reviewers for their helpful suggestions.

References

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., and Polz, M.F. 2004. Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *J. Bacteriol.* **186:** 2629–2635.

Ash, C., Farrow, J.A., Dorsch, M., Stackebrandt, E., and Collins, M.D. 1991. Comparative analysis of *Bacillus anthracis, Bacillus cereus*, and related species on the basis of reverse transcriptase sequencing of 16S rRNA. *Int. J. Syst. Bacteriol.* **41:** 343–346.

Brunder, W. and Karch, H. 2000. Genome plasticity in Enterobacteriaceae. *Int. J. Med. Microbiol.* **290:** 153–165.

Cerdeno-Tarraga, A.M., Patrick, S., Crossman, L.C., Blakely, G., Abratt, V., Lennard, N., Poxton, I., Duerden, B., Harris, B., Quail, M.A., et al. 2005. Extensive DNA inversions in the *B. fragilis* genome control variable gene expression. *Science* **307:** 1463–1465.

Copley, S.D. and Dhillon, J.K. 2002. Lateral gene transfer and parallel evolution in the history of glutathione biosynthesis genes. *Genome Biol.* **3:** 0025.1–0025.16.

Daubin, V. and Ochman, H. 2004. Bacterial genomes as new gene homes: The genealogy of ORFans in *E. coli*. *Genome Res.* **14:** 1036–1042.

Daubin, V., Lerat, E., and Perriere, G. 2003a. The source of laterally transferred genes in bacterial genomes. *Genome Biol.* **4:** R57.

Daubin, V., Moran, N.A., and Ochman, H. 2003b. Phylogenetics and the cohesion of bacterial genomes. *Science* **301:** 829–832.

Dean, A.M., Neuhauser, C., Grenier, E., and Golding, G.B. 2002. The pattern of amino acid replacements in α/β-barrels. *Mol. Biol. Evol.* **19:** 1846–1864.

Felsenstein, J. 1988. Phylogenies from molecular sequences: Inference and reliability. *Annu. Rev. Genet.* **22:** 521–565.

———. 1989. PHYLIP (phylogeny inference package). Version 3.2. *Cladistics* **5:** 164–166.

———. 1992. Phylogenies from restriction sites: A maximum- likelihood approach. *Evolution Int. J. Org. Evolution* **46:** 159–173.

———. 2004. *Inferring phylogenies*. Sinauer Associates, Inc., Sunderland, MA.

Galtier, N. and Boursot, P. 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., Romeu, A., and Palau, J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.* **10:** 1719–1725.

Gu, X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.* **18:** 453–464.

Gu, X. and Zhang, H. 2004. Genome phylogenetic analysis based on extended gene contents. *Mol. Biol. Evol.* **21:** 1401–1408.

Gu, Z., Nicolae, D., Lu, H.H., and Li, W.H. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* **18:** 609–613.

Hao, W. and Golding, G.B. 2004. Patterns of bacterial gene movement. *Mol. Biol. Evol.* **21:** 1294–1307.

Helgason, E., Økstad, O.A., Caugant, D.A., Johansen, H.A., Fouet, A., Mock, M., Hegna, I., and Kolstø, A.-B. 2000. *Bacillus anthracis, Bacillus cereus*, and *Bacillus thuringiensis*—One species on the basis of genetic evidence. *Appl. Environ. Microbiol.* **66:** 2627–2630.

Huelsenbeck, J.P. and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17:** 754–755.

Huson, D.H. and Steel, M. 2004. Phylogenetic trees based on gene content. *Bioinformatics* **20:** 2044–2049.

Itaya, M. 1997. Physical map of the *Bacillus subtilis* 166 genome: Evidence for the inversion of an approximately 1900 kb continuous DNA segment, the translocation of an approximately 100 kb segment and the duplication of a 5 kb segment. *Microbiology* **143:** 3723–3732.

Klappenbach, J.A., Saxman, P.R., Cole, J.R., and Schmidt, T.M. 2001. rrndb: The Ribosomal RNA operon copy Number Database. *Nucleic Acids Res.* **29:** 181–184.

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

Kurland, C.G., Canback, B., and Berg, O.G. 2003. Horizontal gene transfer: A critical view. *Proc. Natl. Acad. Sci.* **100:** 9658–9662.

Kuwahara, T., Yamashita, A., Hirakawa, H., Nakayama, H., Toh, H., Okada, N., Kuhara, S., Hattori, M., Hayashi, T., Ohnishi, Y., et al. 2004. Genomic analysis of *Bacteroides fragilis* reveals extensive DNA inversions regulating cell surface adaptation. *Proc. Natl. Acad. Sci.* **101:** 14919–14924.

Kyrpides, N.C. and Ouzounis, C.A. 1999. Whole-genome sequence annotation: 'Going wrong with confidence.' *Mol. Microbiol.* **32:** 886–887.

Lake, J.A. and Rivera, M.C. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: Conditioned reconstruction. *Mol. Biol. Evol.* **21:** 681–690.

Lawrence, J.G. and Ochman, H. 1997. Amelioration of bacterial genomes: Rates of change and exchange. *J. Mol. Evol.* **44:** 383–397.

Liu, G.R., Rahn, A., Liu, W.Q., Sanderson, K.E., Johnston, R.N., and Liu, S.L. 2002. The evolving genome of *Salmonella enterica serovar* Pullorum. *J. Bacteriol.* **184:** 2626–2633.

Lynch, M. and Conery, J.S. 2003. The evolutionary demography of duplicate genes. *J. Struct. Funct. Genomics* **3:** 35–44.

McLysaght, A., Baldi, P.F., and Gaut, B.S. 2003. Extensive gene gain associated with adaptive evolution of poxviruses. *Proc. Natl. Acad. Sci.* **100:** 15655–15660.

Mirkin, B.G., Fenner, T.I., Galperin, M.Y., and Koonin, E.V. 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.

Neyman, J. 1971. Molecular studies of evolution: A source of novel statistical problems. In *Statistical decision theory and related topics* (eds. S.S. Gupta and J. Yackel), pp. 1–27. Academic Press, New York.

Ochman, H. and Jones, I.B. 2000. Evolutionary dynamics of full genome content in *Escherichia coli*. *EMBO J.* **19:** 6637–6643.

Pal, C., Papp, B., and Lercher, M.J. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* **37:** 1372–1375.

Priest, F.G., Barker, M., Baillie, L.W., Holmes, E.C., and Maiden, M.C. 2004. Population structure and evolution of the *Bacillus cereus* group. *J. Bacteriol.* **186:** 7959–7970.

Siew, N. and Fischer, D. 2003. Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins* **53:** 241–251.

———. 2004. Structural biology sheds light on the puzzle of genomic ORFans. *J. Mol. Biol.* **342:** 369–373.

Silva, F.J., Latorre, A., and Moya, A. 2003. Why are the genomes of endosymbiotic bacteria so stable? *Trends Genet.* **19:** 176–180.

Snel, B., Bork, P., and Huynen, M.A. 1999. Genome phylogeny based on gene content. *Nat. Genet.* **21:** 108–110.

———. 2002. Genomes in flux: The evolution of archaeal and proteobacterial gene content. *Genome Res.* **12:** 17–25.

Stoebel, D.M. 2005. Lack of evidence for horizontal transfer of the lac operon into *Escherichia coli*. *Mol. Biol. Evol.* **22:** 683–690.

Taoka, M., Yamauchi, Y., Shinkawa, T., Kaji, H., Motohashi, W., Nakayama, H., Takahashi, N., and Isobe, T. 2004. Only a small subset of the horizontally transferred chromosomal genes in *Escherichia coli* are translated into proteins. *Mol. Cell. Proteomics* **3:** 780–787.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 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.

Thompson, J.R., Pacocha, S., Pharino, C., Klepac-Ceraj, V., Hunt, D.E., Benoit, J., Sarma-Rupavtarm, R., Distel, D.L., and Polz, M.F. 2005. Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307:** 1311–1313.

Tillier, E.R. and Collins, R.A. 2000. Genome rearrangement by replication-directed translocation. *Nat. Genet.* **26:** 195–197.

Ullrich, S., Kube, M., Schubbe, S., Reinhardt, R., and Schuler, D. 2005. A hypervariable 130-kilobase genomic region of *Magnetospirillum gryphiswaldense* comprises a magnetosome island which undergoes frequent rearrangements during stationary growth. *J. Bacteriol.* **187:** 7176–7184.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.

Zhang, R. and Zhang, C.T. 2003. Identification of genomic islands in the genome of *Bacillus cereus* by comparative analysis with *Bacillus anthracis*. *Physiol. Genomics* **16:** 19–23.

Zhang, P., Gu, Z., and Li, W.H. 2003. Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol.* **4:** R56.

Received September 28, 2005; accepted in revised form February 21, 2006.