DOI: 10.1534/genetics.110.118448

Inferring Bacterial Genome Flux While Considering Truncated Genes

Weilong Hao*,[1] and G. Brian Golding†

**Department of Biology, Indiana University, Bloomington, Indiana 47405 and †Department of Biology, McMaster University, Hamilton, Ontario L8S 4K1, Canada*

Manuscript received May 4, 2010
Accepted for publication June 4, 2010

ABSTRACT

Bacterial gene content variation during the course of evolution has been widely acknowledged and its pattern has been actively modeled in recent years. Gene truncation or gene pseudogenization also plays an important role in shaping bacterial genome content. Truncated genes could also arise from small-scale lateral gene transfer events. Unfortunately, the information of truncated genes has not been considered in any existing mathematical models on gene content variation. In this study, we developed a model to incorporate truncated genes. Maximum-likelihood estimates (MLEs) of the new model reveal fast rates of gene insertions/deletions on recent branches, suggesting a fast turnover of many recently transferred genes. The estimates also suggest that many truncated genes are in the process of being eliminated from the genome. Furthermore, we demonstrate that the ignorance of truncated genes in the estimation does not lead to a systematic bias but rather has a more complicated effect. Analysis using the new model not only provides more accurate estimates on gene gains/losses (or insertions/deletions), but also reduces any concern of a systematic bias from applying simplified models to bacterial genome evolution. Although not a primary purpose, the model incorporating truncated genes could be potentially used for phylogeny reconstruction using gene family content.

GENE content variation as a key feature of bacterial genome evolution has been well recognized (Garcia-Vallvé *et al.* 2000; Ochman and Jones 2000; Snel *et al.* 2002; Welch *et al.* 2002; Kunin and Ouzounis 2003; Fraser-Liggett 2005; Tettelin *et al.* 2005) and gained increasing attention in recent years. Various methods have been employed to study the variation of gene content in the form of gene insertions/deletions (or gene gains/losses); there are studies of population dynamics (Nielsen and Townsend 2004), birth-and-death evolutionary models (Berg and Kurland 2002; Novozhilov *et al.* 2005), phylogeny-dependent studies including parsimony methods (Mirkin *et al.* 2003; Daubin *et al.* 2003a,b; Hao and Golding 2004), and maximum-likelihood methods (Hao and Golding 2006, 2008b; Cohen *et al.* 2008; Cohen and Pupko 2010; Spencer and Sangaralingam 2009). The pattern of gene presence/absence also contains phylogenetic signals (Fitz-Gibbon and House 1999; Snel *et al.* 1999; Tekaia *et al.* 1999) and has been used for phylogenetic reconstruction (Dutilh *et al.* 2004; Gu and Zhang 2004; Huson and Steel 2004; Zhang and Gu 2004; Spencer *et al.* 2007a,b). All these studies make use of the binary information of gene presence or absence and neglect the existence of gene segments or truncated genes.

Bacterial genomes are known to harbor pseudogenes. An intracellular species *Mycobacterium leprae* is an extreme case for both the proportion and the number of pseudogenes: estimated as 40% of the 3.2-Mb genome and 1116 genes (Cole *et al.* 2001). In free-living bacteria, pseudogenes can make up to 8% of the annotated genes in the genome (Lerat and Ochman 2004). Many pseudogenes result from the degradation of native functional genes (Cole *et al.* 2001; Mira *et al.* 2001). Pseudogenes could also result from the degradation of transferred genes and might even be acquired directly via lateral gene transfer. For instance, in plant mitochondrial genomes, which have an α-proteobacterial ancestry, most, if not all, of the laterally transferred genes are pseudogenes (Richardson and Palmer 2007). Furthermore, evidence has been documented that gene transfer could take place at the subgenic level in a wide range of organisms, *e.g.*, among bacteria (Miller *et al.* 2005; Choi and Kim 2007; Chan *et al.* 2009), between ancient duplicates in archaea (Archibald and Roger 2002), between different organelles (Hao and Palmer 2009; Hao 2010), and between eukaryotes (Keeling and Palmer 2001). A large fraction of pseudogenes have been shown to arise from failed lateral transfer events (Liu *et al.* 2004) and most of them are transient in bacterial genomes (Lerat and Ochman 2005). Zhaxybayeva *et al.* (2007) reported that genomes with truncated homologs might erroneously lead to false inferences of "gene gain" rather than multiple instances of "gene loss." This raises the question of how a false

Supporting Information is available online at http://www.genetics.org/cgi/content/full/genetics.110.118448/DC1.

[1] *Corresponding author:* Department of Biology, Indiana University, Bloomington, IN 47405. E-mail: haoweilong@gmail.com

diagnosis of gene absence affects the estimation of insertion/deletion rates. Recently, we showed that the effect of a false diagnosis of gene absence on estimation of insertion/deletion rates is not systematic, but rather more complicated (Hao and Golding 2008a). To further address the problem, a study incorporating the information of truncated genes is highly desirable. This will not only yield more accurate estimates of the rates of gene insertions/deletions, but also provide a quantitative view of the effect of truncated genes on rate estimation, which has been understudied in bacterial genome evolution.

In this study, we developed a model that considers the information of truncated genes and makes use of a parameter-rich time-reversible rate matrix. Rate variation among genes is allowed in the model by incorporating a discrete Γ-distribution. We also allow rates to vary on different parts of the phylogeny (external branches *vs.* internal branches). Consistent with previous studies, the rates of gene insertions/deletions are comparable to or larger than the rates of nucleotide substitution and the rates of gene insertions/deletions are further inflated in closely related groups and on external branches, suggesting high rates of gene turnover of recently transferred genes. The results from the new model also suggest that many recently truncated genes are in the process of being rapidly deleted from the genome. Some other interesting estimates in the model are also presented and discussed. One implication of the study, though not primary, is that the state of truncated genes could serve as an additional phylogenetic character for phylogenetic reconstruction using gene family content.

METHODS

Phylogenetic analysis and genome comparison: Four bacterial groups with an abundance of completely sequenced congeneric species/strains and relatively large genome sizes were selected (listed in Table 1 with outgroup information shown in supporting information, Table S1). The four groups are Bacillaceae, Clostridium, Escherichia/Shigella (Escherichia, for simplicity), and Pseudomonas. Within each group, analyses were conducted separately in three clades with different levels of divergence (Figure 1). A large number of universally present nonduplicated genes from each clade were extracted and examined to obtain a robust phylogenetic tree. The numbers of concatenated genes (and characters) are 325 genes (335,380 characters) for clades B1 and B2; 329 genes (362,583 characters) for clade B3; 108 genes (130,531 characters) for clades C1, C2, and C3; 755 genes (809,248 characters) for clades E1, E2, and E3; and 434 genes (516,571 characters) for clades P1, P2, and P3. Alignment of each gene was generated individually using MUSCLE (Edgar 2004) followed by a concatenation of individual alignments into a single giant alignment for each clade. A maximum-likelihood tree was generated for each clade on the concatenated sequences using the PHYLIP package (Felsenstein 1989) version 3.67 and the rate variation parameter α in a Γ-distribution was estimated using the PUZZLE program (Strimmer and von Haeseler 1996). The sum of branch lengths for each tree was used as an indicator for the divergence of the clade.

In addition to the two states, "*p*" for gene presence and "*a*" for gene absence, a new state "*f*" for fragment (truncated genes) was introduced. The method to identify members of a gene family was modified from Hao and Golding (2004), and all paralogs in each genome were clustered as a single gene family and only one member was retained for further analysis. First, annotated protein sequences were clustered into gene families following a criterion of *E*-value $\leq 10^{-20}$ and match length ≥85% in a BLASTP search (Altschul *et al.* 1997). Gene families retained for further analysis were required to have >100 amino acids in at least one gene member, since similarity searches using BLAST have low power to detect homologs in short sequences (Altschul *et al.* 1997). Genes clustered in gene families were considered as gene presence or *p*. Second, we further analyzed the genomes that do not have annotated protein sequences for each gene family by conducting a TBLASTN search (Altschul *et al.* 1997) using an annotated protein sequence as the query sequence. When a gene family has more than one annotated member, the protein sequence with the median length (the shorter of the two median genes in the case of an even number of genes) of the family was chosen as the query sequence. When no annotated protein sequence for a gene family was found in a genome, there are three possible conditions of the gene: the gene could be present (but not annotated), truncated (short in length), or genuinely absent: (1) gene presence (*p*) was inferred, if the BLAST hit has an *E*-value $\leq 10^{-20}$ and match length ≥85%; (2) gene truncation (*f*) was inferred, if the BLAST hit has an *E*-value $\leq 10^{-20}$ but match length <85%; and (3) hits that have an *E*-value $>10^{-20}$ were considered as gene absence (*a*). The observed patterns of gene presence/absence/truncation are shown in Table S2, Table S3, Table S4, and Table S5. To access the robustness of the analysis, a different criterion of *E*-value $\leq 10^{-10}$ and match length ≥70% in both BLASTP and TBLASTN searches was used in gene family identification (Table S6, Table S7, Table S8, and Table S9 and Figure 2).

As in Hao and Golding (2004), the "single link" method (Friedman and Hughes 2003) was employed to define gene families (*e.g.*, if A and B are in a family and B and C are in a family, then A, B, and C are in a family). By doing this, there is an increased risk of a truncated gene being mistakenly identified as "present" (Hao and Golding 2008a). The risk would become higher, when more genomes are compared. To avoid such a problem as much as possible, we limit the number of taxa in each clade to five, which also makes the computation less demanding.

The mathematical model: The transitions among *p*'s, *f*'s, and *a*'s are defined by a 3 × 3 instantaneous rate matrix Q with stationary probabilities (π_a, π_f, π_p). Here $\pi_a + \pi_f + \pi_p = 1$, and the matrix Q is reversible,

$$Q = \left(\begin{array}{cccc} & a & f & p \\ a & -\pi_f\alpha - \pi_p\beta & \pi_f\alpha & \pi_p\beta \\ f & \pi_a\alpha & -\pi_a\alpha - \pi_p\gamma & \pi_p\gamma \\ p & \pi_a\beta & \pi_f\gamma & -\pi_a\beta - \pi_f\gamma \end{array}\right), \tag{1}$$

where α, β, and γ are the rate ratios between the state pairs *af*, *ap*, and *fp*, respectively. They are also known as the exchangeability terms. For instance, $\forall x, y \in \{a, f, p\}$, $x \neq y$, $Q(x, y)$ is the rate at which state x changes to state y, and all entries satisfy $\pi_x Q(x, y) = \pi_y Q(y, x)$. When gene truncation is not considered, there is no f state and the matrix would be reduced to

$$Q = \left(\begin{array}{ccc} & a & p \\ a & -\pi_p & \pi_p \\ p & \pi_a & -\pi_a \end{array}\right), \tag{2}$$

which has been used in previous studies (Cohen *et al.* 2008; Spencer and Sangaralingam 2009; Cohen and Pupko 2010). When $\pi_a = \pi_p = 0.5$, the matrix entries $Q(a, p)$ and $Q(p, a) = 1$ (see Equation 3 below for detail), or

$$Q = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix},$$

and the model is equivalent to that used in Hao and Golding (2006, 2008b) and Cohen *et al.* (2008). This model is labeled M_{00} in Table 3. To reduce the number of parameters to be optimized, the α-parameter in the $Q_{3\times 3}$ matrix was fixed to be 1. It is a standard practice to allow only calibrated rate matrices; *i.e.*, Q satisfies

$$-\sum_{a} \pi_a Q(a, a) = 1, \tag{3}$$

so that a rate parameter (shown as μ below) is the average number of transition events per gene family per evolutionary time. The transition probability matrix is

$$P = e^{Q\mu t}, \tag{4}$$

where t is the branch length based on nucleotide sequences, and μ is a rate parameter.

Since some patterns are not observable, we calculate the likelihood conditional on a pattern being observable, L^i_+ as suggested in Felsenstein (1992),

$$L^i_+ = \frac{L^i}{1 - L^i_-}, \tag{5}$$

where L^i_- is the likelihood of unobservable patterns for gene family i. For cases where a number of patterns are unobservable, each such pattern is a disjoint event, and summation is taken over all unobservable patterns to get L^i_-,

$$L^i_- = \sum_{j \in U} L_{j-}, \tag{6}$$

where U is the set of unobservable patterns (Table S10), and L_{j-} is the likelihood of the jth unobservable pattern. Here L^i_- has the same value for all i.

Rate variation was also considered in a similar manner to nucleotide rate heterogeneity in phylogeny reconstruction (Yang 1994; Felsenstein 2001). A discrete Γ-model with eight rate categories ($M = 8$ categories) was implemented in the maximum-likelihood estimation. The likelihood on gene family i is the sum of likelihoods for each rate category ν for that gene family weighted by the category probabilities p_ν,

$$L^i = \sum_{\nu=1}^{M} p_\nu L^i_\nu, \tag{7}$$

where $\sum_{\nu=1}^{M} p_\nu = 1$. After incorporating a discrete Γ-model as done in Spencer and Sangaralingam (2009), the likelihood of observing the pattern of gene family i will be

$$L^i_+ = \frac{L^i}{1 - L^i_-} = \frac{\sum_{\nu=1}^{M} p_\nu L^i_\nu}{1 - \sum_{\nu=1}^{M} p_\nu L^i_{\nu-}}. \tag{8}$$

First, parameters were estimated by assuming π_a, π_f, and π_p to be the frequencies of each character state in the data. This is called model M_0. In the case of only two characters (a, p), model M_0 is when π_a and π_p are the frequencies of each character state in the data, and another model, in which π_a and $\pi_p = 0.5$, was introduced (called model M_{00}) since it has been used in previous studies (Hao and Golding 2006, 2008b; Cohen *et al.* 2008). Then π_a, π_f, and π_p were treated as parameters to be optimized and hence called model $M_0 + \pi$. Finally, a discrete Γ-distribution was incorporated (model $M_0 + \Gamma + \pi$). All free parameters were estimated such that they maximize the likelihood of the data. This was achieved using the Nelder–Mead simplex method (Nelder and Mead 1965), which is slower than some gradient-based methods and the EM method but less likely to be misled to local maxima (Yap and Speed 2005; R Development Core Team 2008). To further reduce the chance of being trapped in local maxima, different initial values were used, and the final estimates with the highest likelihood were picked.

RESULTS

In this study, information on truncated genes was incorporated into the maximum-likelihood model. Analyses were conducted in four bacterial groups, and each group contains three clades with different levels of divergence (Table 1 and Figure 1). The results reveal that closely related clades have higher rates of gene insertions/deletions (μ) than distantly related clades (Figure 2 and Table 2). This trend holds throughout all four groups, and the use of different cutoff thresholds on identifying gene families yields remarkably similar results (Figure 2 and Table S11). This is consistent with previous findings that recently acquired genes have high rates of gene turnover (Daubin and Ochman 2004; Hao and Golding 2004, 2006). Under the $M_0 + \pi$ model (Figure 2), the stationary probability π_a is positively associated with the tree length of each clade. It is important to clarify that the tree length is *not* an estimate from the gene insertion/deletion model. Indeed, it is the sum of branch lengths based on nucleotide substitution and it was used as an indicator for the degree of divergence in the clade.

To further access the robustness of using two different cutoff thresholds, we plotted the distribution of length variation in reciprocal best BLASTP hits from 24 selected genome pairs (see Figure S1 for details). Genome pairs Cph–Cth in C3 and Sfl–Sdy in E1 show the highest level of length variation (Figure S1). Then we plotted the DNA distance and K_a/K_s ratio of these homologs (Figure S2). All examined genome pairs have a median DNA distance <1.0 and a median K_s/K_s ratio <0.2 (Figure S2). Among the examined genome pairs, Cph–Cth is the most diverse, while Sfl–Sdy and Eco5–Eco6 are the least diverse. After that, we simulated the expected distribution of match length at given sequence divergence and functional constraints with no indels allowed. It is shown that the number of imperfect matches increases when the homologous pairs are more diverse (Figure S3) and the query sequences are shorter (Figure S4). Compared with the simulated data, reciprocal best BLASTP hits show a significantly higher level of length variation than expected. The most extreme case is the Sfl–Sdy pair, which has a remarkably low level of sequence divergence (Figure S2) but a significantly high level of length variation (Figure S1). The K_s/K_s ratios in the Sfl–Sdy pair are significantly higher than those in either Eco5–Eco6 or Efe–Sen (each with $P < 0.0001$ in a Wilcoxon rank test), suggesting that many genes are under relaxed functional constraints in at least one of the two genomes in Sfl–Sdy. In fact, the *Shigella flexneri* genome (Sfl) has been documented to undergo genome reduction and gene pseudogenization (Wei *et al.* 2003; Lerat and Ochman 2004; Dagan *et al.* 2006). The high level of length variation in the Cph–Cth pair can be explained in part by the high degree of sequence divergence. Importantly, the observed level of length variation in all genome pairs is always higher than that of the simulated data at a similar level of sequence divergence. Some of the inflated length variation could have been introduced by problematic annotations. However, during the annotation process closely related genomes are routinely used as references for gene identification and the annotations are frequently updated. It is reasonable to believe that some of the annotated genes might indeed have been affected by gene truncation. Finally we plotted the observed distribution of match length of the TBLASTN hits in the examined genomes (Figure S5). It is shown that the imperfect TBLASTN hits are not uniformly distributed; instead, the number of imperfect TBLASTN hits increases as match length increases. Possible explanations would be that (1) truncated genes are selectively disadvantageous and shorter gene lengths would likely result in a greater disadvantage, and (2) some truncated genes failed to be detected and more such failures occur when longer stretches of gene sequences are missing. During the TBLASTN search, we used the longest, median, and shortest sequences from each gene family as query sequences (see Figure S5 for details). Furthermore, a smaller word size (−W 2) in the TBLASTN search was used in addition to the default word size (−W 3). It shows that using the longest, median, or shortest sequences as query sequences and using a smaller word size in the search led to remarkably similar results in our examined genomes (Figure S5).

When the clade is more diverse, more gene families that were once present in the ancestral genome are lost from some descendants. Figure S6 illustrates a decreasing trend of the number of commonly present gene families (with the exception of a slight increase from E1 to E2) and an increasing trend of the number of strain-specific gene families when clades become more diverse. The most parsimonious explanation for the decrease of commonly present gene families in more diverse clades is the loss of ancient gene families during evolution. The increase of strain-specific gene families also supports the loss of some ancient genes. If genome size stays relatively constant over time, the increase of recently acquired genes should be a reflection of the decrease of ancient genes. Furthermore, the stationary probability π_a appears to be greater than the observed frequency of *a* in each clade (Figure S7). This is expected since the genes that were once present but have been deleted from all the descendants are unobservable in the current data (Table S10), but have been taken into account in the maximum-likelihood estimation (Equation 5). The rate variation parameter α (shown as α_Γ in Table 2) has a positive association with tree length in the Pseudomonas and Bacillaceae groups (Figure 2). These data are in agreement with Hao and Golding (2008b) that closely related groups tend to have high degrees of rate variation for gene insertions/deletions among genes, while distantly related groups tend to have low degrees of rate variation for gene insertions/deletions. However, such a positive association was not found in the Escherichia and Clostridium groups. We suspect that the low divergence in the E1, E2, and C1 clades and the relatively low absolute

TABLE 1

Strain information from a variety of phylogenetic groups

Clade	Tree length	Species/strain name	Abbreviation	Accession
B1	0.366	*Bacillus anthracis* str. Ames	Ba	NC_003997
		B. cereus ATCC 10987	Bc$_1$	NC_003909
		B. cereus ATCC 14579	Bc$_2$	NC_004722
		B. weihenstephanensis	Bw	NC_010184
		B. cereus subsp. cytotoxis	Bc$_3$	NC_009674
B2	1.981	*B. amyloliquefaciens*	Bam	NC_009725
		B. subtilis	Bs	NC_000964
		B. licheniformis ATCC 14580	Bl	NC_006322
		B. pumilus	Bp	NC_009848
		Geobacillus kaustophilus	Gk	NC_006510
B3	3.274	*B. halodurans*	Bh	NC_002570
		B. clausii	Bcl	NC_006582
		Oceanobacillus iheyensis	Oi	NC_004193
		Exiguobacterium sibiricum	Es	NC_010556
		Anoxybacillus flavithermus	Af	NC_011567
C1	0.055	*Clostridium botulinum* A str. Hall	Cbo$_1$	NC_009698
		C. botulinum A2 str. Kyoto	Cbo$_2$	NC_012563
		C. botulinum B1 str. Okra	Cbo$_3$	NC_010516
		C. botulinum Ba4 str. 657	Cbo$_4$	NC_012658
		C. botulinum A3 str. Loch Maree	Cbo$_5$	NC_010520
C2	1.415	*C. botulinum* B str. Eklund 17B	Cbo$_6$	NC_010674
		C. botulinum E3 str. Alaska E43	Cbo$_7$	NC_010723
		C. beijerinckii NCIMB 8052	Cbe	NC_009617
		C. perfringens SM101	Cpe	NC_008262
		C. acetobutylicum ATCC 824	Cac	NC_003030
C3	2.958	*C. novyi* NT	Cno	NC_008593
		C. tetani E88	Cte	NC_004557
		C. difficile 630	Cdi	NC_009089
		C. phytofermentans ISDg	Cph	NC_010001
		C. thermocellum ATCC 27405	Cth	NC_009012
E1	0.039	*Shigella boydii* Sb227	Sbo	NC_007613
		S. sonnei Ss046	Sso	NC_007384
		Escherichia coli E24377A	Eco$_1$	NC_009801
		S. flexneri 5 str. 8401	Sfl	NC_008258
		S. dysenteriae Sd197	Sdy	NC_007606
E2	0.044	*E. coli* 536	Eco$_2$	NC_008253
		E. coli ED1a	Eco$_3$	NC_011745
		E. coli APEC 01	Eco$_4$	NC_008563
		E. coli O127:H6 str. E2348/69	Eco$_5$	NC_011601
		E. coli IAI39	Eco$_6$	NC_011750
E3	0.387	*E. coli* IAI1	Eco$_7$	NC_011741
		E. coli HS	Eco$_8$	NC_009800
		E. coli S88	Eco$_9$	NC_011742
		E. fergusonii ATCC 35469	Efe	NC_011740
		Salmonella enterica subsp. arizonae serovar 62:z4,z23	Sen	NC_010067
P1	0.312	*Pseudomonas putida* F1	Ppu$_1$	NC_009512
		P. putida KT2440	Ppu$_2$	NC_002947
		P. putida GB-1	Ppu$_3$	NC_010322
		P. putida W619	Ppu$_4$	NC_010501
		P. entomophila L48	Pen	NC_008027

(*continued*)

TABLE 1

(Continued)

Clade	Tree length	Species/strain name	Abbreviation	Accession
P2	0.885	*P. fluorescens* Pf0-1	Pfl_1	NC_007492
		P. fluorescens SBW25	Pfl_2	NC_012660
		P. fluorescens Pf-5	Pfl_3	NC_004129
		P. syringae pv. phaseolicola 1448A	Psy_1	NC_005773
		P. mendocina ymp	Pme	NC_009439
P3	1.118	*P. putida* KT2440	Ppu_2	NC_002947
		P. syringae pv. tomato str. DC3000	Psy_2	NC_004578
		P. mendocina ymp	Pme	NC_009439
		P. stutzeri A1501	Pst	NC_009434
		P. aeruginosa PA7	Pae	NC_009656

numbers of gene insertions/deletions do not provide enough statistical power for the estimation of the rate variation parameter despite the high estimated rates. A lack of statistical power was previously documented in some phylogenetic groups with small genome sizes and/or closely related species (Hao and Golding 2008b). Indeed, the removal of the E1, E2, and C1 clades yields a strong positive association (with *P*-value = 0.0054) between tree length and α_Γ in the remaining nine clades (Figure S8).

In the instantaneous rate matrix, α, β, and γ are the rate ratios between the state pairs *af*, *ap*, and *fp*, respectively, and are also known as the exchangeability terms. They are plotted for each clade in Figure 3. Here, α was fixed to be 1, and β and γ were estimated under the $M_0 + \pi$ model. The trend seems to be that the β- and γ-values increase as the clades become more diverse. There are two exceptions (in two clades, E1 and C3) to this trend:

1. The β- and γ-values in the E1 clade are larger than those in the E2 clade. This could possibly be due to the low number of commonly present gene families in E1 (Figure S6), which is very likely associated with the process of genome reduction and gene pseudogenization in the *S. flexneri* (Sfl) genome (Wei *et al.* 2003; Lerat and Ochman 2004; Dagan *et al.* 2006). By contrast, the number of commonly present gene families generally decreases as the clade divergence increases. Furthermore, the similar level of divergence between E1 and E2 could potentially lead to the lack of statistical power to estimate parameters in very closely related clades as suggested in Hao and Golding (2008b).
2. The β-value in the C3 clade is smaller than the β-value in the C2 clade. Genome size was found to vary greatly in both clades, *e.g.*, from 2.9 to 6.0 Mb in the C2 clade and from 2.5 to 4.8 Mb in the C3 clade. In the C2 clade, Cbe is significantly larger than the remaining four genomes, while in the C3 clade, Cno and Cte are significantly smaller than the remaining three genomes. We sought to address whether the unexpected pattern of the β-parameter in the instantaneous rate matrix could be explained by the highly variable genome sizes. A separate instantaneous rate matrix was assumed on the branches associated with the strain(s) with substantially different genome sizes (Cbe in C2, Cno and Cte in C3). The parameters β and γ are higher on the branch leading to the large genome (Cbe) and lower on the branches associated with the two small genomes (Cno and Cte), compared with on the rest of the phylogeny (Figure S9). The β-values on the rest of the phylogeny are 1.224 for C2 and 1.486 for C3, and they yield an increasing trend from C1 (β = 0.315 in C1) to C2 and to C3. Such a trend has been observed in Pseudomonas and Bacillaceae (Figure 3).

Furthermore, we computed the product of the scaled instantaneous rate matrix *Q* and the rate parameter μ (Table S12), which presents the instantaneous rates for all possible transitions. There is a clear trend that the instantaneous rates for all parameters increase as the clade becomes more closely related. As a part of the picture, the increased rates associated with character *f* in more closely related clades suggest that many truncated genes are in the process of being rapidly deleted from the genome.

We then sought to address the question whether a false diagnosis of gene absence systematically overestimates the rates of gene insertion/deletion. First, we conducted an analysis as in Hao and Golding (2006), in which the truncated genes were classified as absent ($f \rightarrow a$), in a false diagnosis of gene absence in Zhaxybayeva *et al.* (2007). To make a comparison, we conducted another set of analyses by forcing all truncated genes to be classified entirely as present ($f \rightarrow p$). Maximum-likelihood estimation was then conducted for both scenarios and the MLEs are shown in Table 3. When all truncated genes were classified as present ($f \rightarrow p$), rather than absent ($f \rightarrow a$), all 12 clades showed a lower μ under the M_0 model. Under the $M_0 + \pi$ model, 5 clades (B1, C1, E1, E2, and E3) showed a lower μ, while the remaining 7 clades showed a

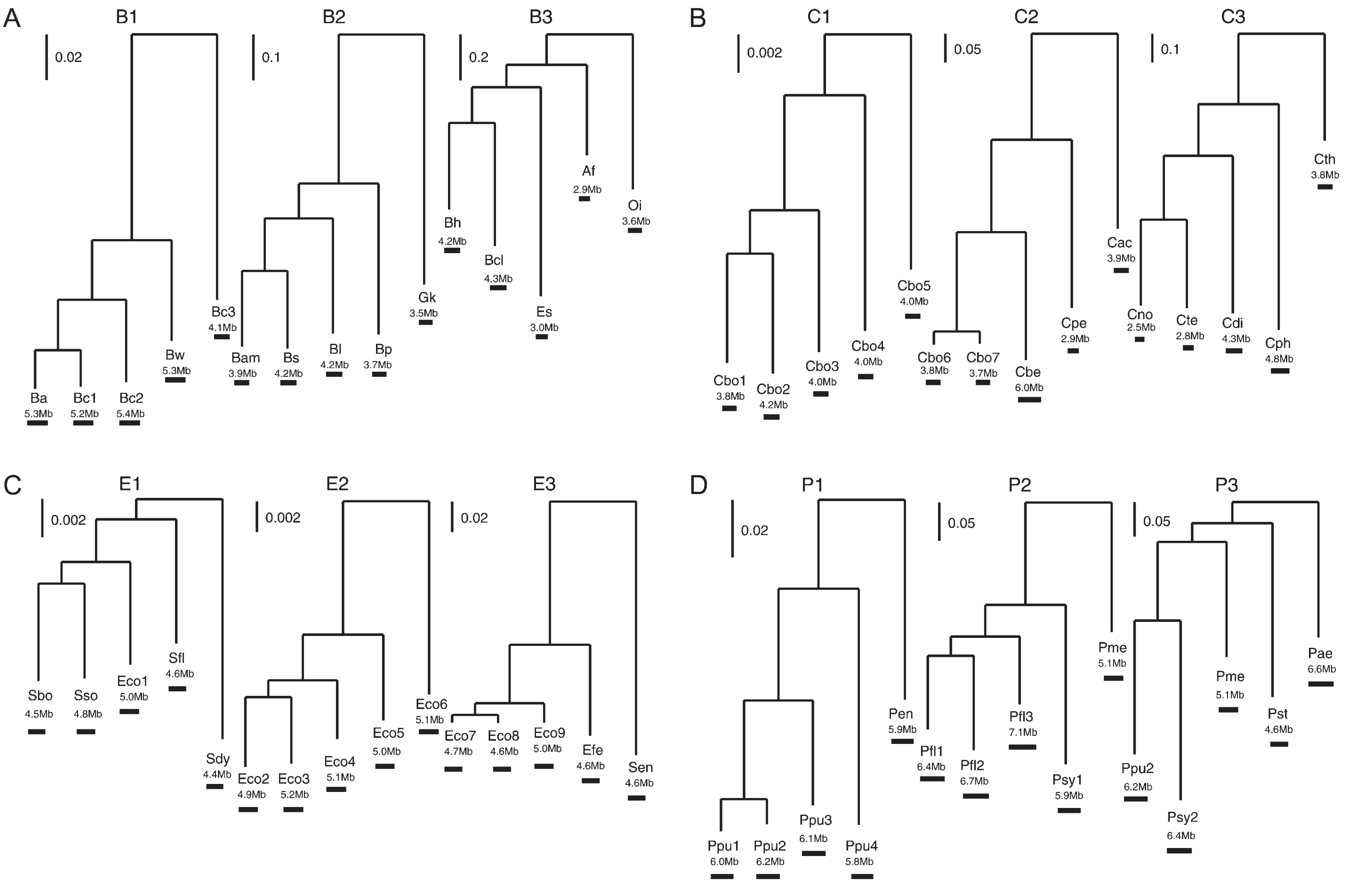


Figure 1.—(A–D) Phylogenies with varied levels of divergence. Clade names and strain abbreviations are as in Table 1. Horizontal scale bar indicates genome size.

higher μ. This is consistent with Hao and Golding (2008a) that the effect of false diagnosis of gene absence does not lead to a systematic bias but rather has a more complicated effect. As expected from the change of the frequency of state *a*, the stationary probability π_a became smaller in every clade after all truncated genes were classified as present, compared with when truncated genes were classified as absent.

Finally, gene insertion/deletion rates were distinguished on different parts of the phylogeny, namely between external branches and internal branches (as shown in Figure 4). Table 4 shows that nine clades have significant improvement when rates on external branches and internal branches were distinguished. All nine clades show higher rates of gene insertions/deletions on external branches than on internal branches. Similar results were observed previously in Bacillaceae strains, Streptococcus strains, and Corynebacterium strains (Hao and Golding 2006; Marri *et al.* 2006, 2007). The three clades that do not show significant improvement are B2, B3, and C3 (Table 4), and they are the three most diverse clades in the study (Table 1 and Figure 2). Since most of the dynamics of gene insertions/deletions occur at the tip of the phylogeny (Hao and Golding 2006, 2008b), it is perhaps not surprising that little or no difference in the rates of gene insertions/deletions was found between external branches and internal branches in such diverse clades. In fact, substantially different results were observed on gene gains/losses by applying similar parsimony methods on distantly related species (Mirkin *et al.* 2003) *vs.* on closely related species (Hao and Golding 2004).

DISCUSSION

Here, we are primarily interested in modeling gene insertions/deletions with consideration for truncated genes. We have not attempted to infer the functionality of any truncated genes. First, there has never been a standard criterion in the literature for pseudogene identification (Chain *et al.* 2004; Lerat and Ochman 2004). Second, detection of pseudogenes requires extensive knowledge of each gene's transcription and its protein's function but this is beyond the scope of this study. Finally, the boundary between gene and pseudogene might rather be ambiguous (Zheng and Gerstein 2007). Presence of an annotated gene within a genome does not necessarily suggest its functionality, but ironically, some shortened homologs might still carry out some function (Ogata *et al.* 2001).

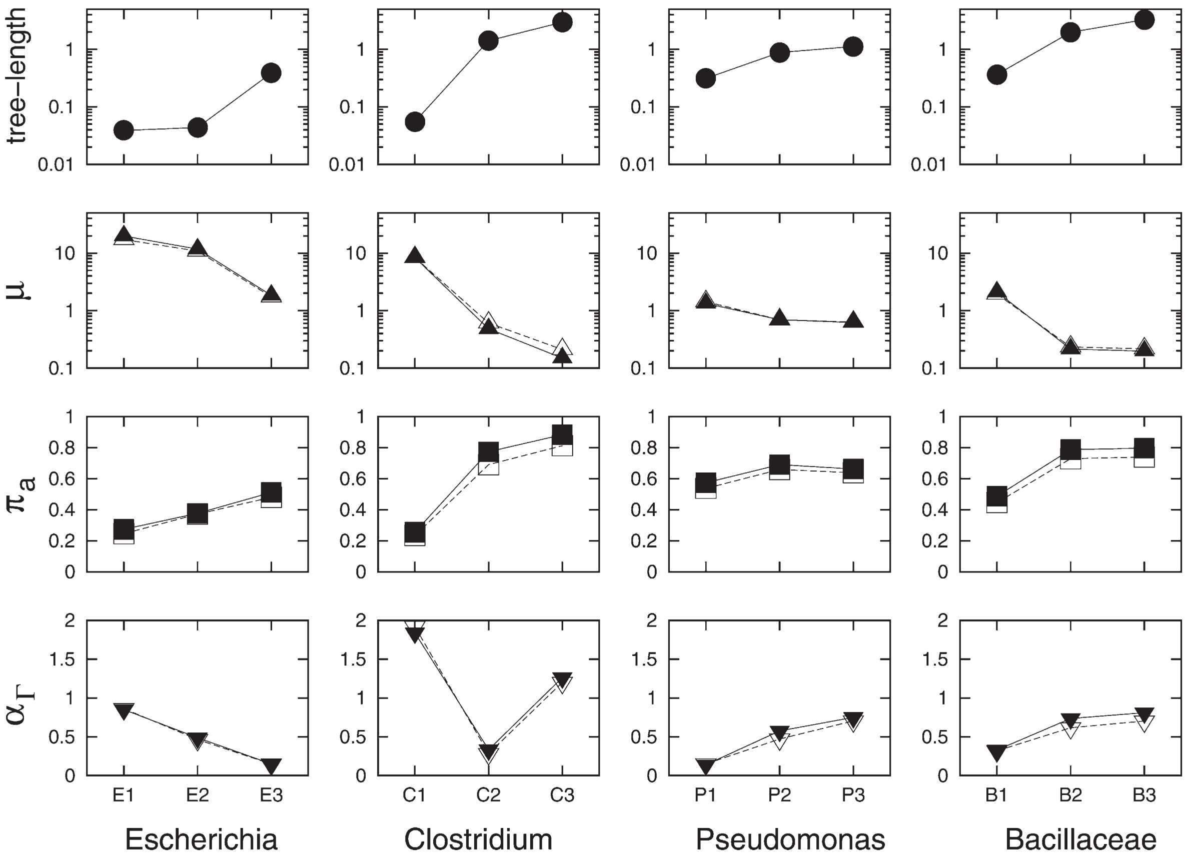


FIGURE 2.—Association between estimated parameters and phylogenetic divergence in each group. Three maximum-likelihood estimates (MLEs), μ, π_a, and α_Γ, are estimated under the $M_0 + \pi$ model using different cutoff thresholds in BLASTP/TBLASTN searches. Solid symbols represent MLEs using a criterion of *E*-value $\leq 10^{-20}$ and match length ≥85%, while open symbols represent MLEs using a criterion of *E*-value $\leq 10^{-10}$ and match length ≥70%. The four groups are sorted from the least diverse group (Escherichia) on the left to the most diverse group (Bacillaceae) on the right. Although shown along with three estimates, tree length is *not* an estimate from the gene insertion/deletion model. Indeed, it is the sum of branch lengths based on nucleotide substitution and used as an indicator for the degree of divergence in the clade.

Our current study classifies genes into three categories (presence/absence/fragment) and makes no attempt to examine any sequence divergence at the gene or subgenic levels. If a whole gene or a fraction of it was replaced via a lateral transfer with a similar gene, the methods employed here would fail to detect these transfers. In other words, homologous recombination (HR) would *not* directly contribute to any rate changes under our current model. However, if HR has occurred in the genes we used for phylogeny reconstruction, it might affect the maximum-likelihood estimates (MLEs) since our estimation was based on the gene phylogeny (*e.g.*, *t* in Equation 4 is branch length). The effect of such HRs on the MLEs is likely complex. If HR has occurred between two examined genomes in a clade and the observed sequence diversity is likely to have been diminished, the branch lengths on the gene phylogeny would have been underestimated and the estimated rates of gene insertions/deletions might have been overestimated. If HR has occurred between one examined genome and any unsampled distantly related genome, the recombined branch length would have been overestimated and, as a result, the estimated rates of gene insertions/deletions might have been underestimated. If HR has occurred and even altered the branching order(s), the estimated rates of gene insertions/deletions would likely have been overestimated, since one generally expects to infer more evolutionary events on a less parsimonious tree. If HR has occurred in the combination of two scenarios or more, the effect on the MLEs could be even more complex. In our study, the phylogeny of each clade was constructed using concatenated sequences of a large number of single-copy genes. Although not completely immune to lateral transfer (YAP *et al.* 1999; BROCHIER *et al.* 2000), commonly present single-copy genes have been shown to exhibit mostly vertical descent (HOOPER and BERG 2003; WELLNER *et al.* 2007). We therefore believe that the effect of any potential HRs in the genes used for phylogeny reconstruction should be small.

As in previous studies (HAO and GOLDING 2006, 2008b; COHEN *et al.* 2008), we initially assumed a constant rate of gene insertions/deletions on each phylogeny. This simplifying assumption is not realistic.

TABLE 2

Maximum log-likelihood comparison of different evolutionary models

		Bacillaceae			Clostridium			Escherichia			Pseudomonas		
Model	Parameter	B1 (5,453)[c]	B2 (5,614)	B3 (6,813)	C1 (3,546)	C2 (6,526)	C3 (7,645)	E1 (4,681)	E2 (5,063)	E3 (5,307)	P1 (6,505)	P2 (8,761)	P3 (9,140)
M_0	μ	2.705	0.515	0.404	8.315	1.041	0.478	20.143	12.774	2.683	1.847	1.162	0.947
	β	0.577	1.281	1.700	0.326	2.430	1.851	0.372	0.285	0.439	0.587	1.117	1.466
	γ	1.311	2.096	2.725	0.590	3.498	3.050	0.686	0.495	0.817	0.831	1.703	2.209
	ln L	−16,551	−18,432	−24,799	−7,424	−22,783	−27,618	−13,019	−12,016	−15,071	−17,175	−29,926	−32,327
$M_0 + \pi$	μ	2.102	0.214	0.197	8.385	0.482	0.149	19.805	11.699	1.837	1.332	0.695	0.619
	β	0.440	0.845	1.314	0.315	1.578	1.191	0.335	0.259	0.353	0.535	0.892	1.364
	γ	1.900	3.862	4.489	0.728	4.397	4.742	0.860	0.704	1.109	1.210	2.638	3.254
	π_a	0.487	0.788	0.797	0.255	0.774	0.883	0.274	0.377	0.511	0.574	0.691	0.662
	π_f	0.041	0.036	0.035	0.029	0.036	0.037	0.074	0.040	0.044	0.041	0.047	0.055
	ln L	−16,243	−17,799	−24,069	−7,385	−22,112	−26,491	−12,924	−11,876	−14,822	−16,787	−29,169	−31,715
	Δ AIC[a]	−612	−1,262	−1,456	−74	−1,338	−2,250	−186	−276	−494	−772	−1,510	−1,220
$M_0 + \Gamma + \pi$	μ	3.811	0.264	0.285	8.446	1.623	0.193	22.112	14.150	8.257	4.123	0.948	0.809
	α_Γ[b]	0.330	0.736	0.811	1.839	0.334	1.263	0.850	0.487	0.150	0.139	0.575	0.751
	ln L	−15,861	−17,553	−23,738	−7,370	−21,312	−26,253	−12,830	−11,728	−13,951	−16,334	−28,674	−31,325
	Δ AIC[a]	−762	−490	−660	−28	−1598	−474	−190	−294	−1740	−904	−988	−778

[a] Akaike's information criteria (Δ AICs) are shown as $(M_0 + \pi)$ *vs.* M_0 and $(M_0 + \Gamma + \pi)$ *vs.* $(M_0 + \pi)$. By definition, AIC $= 2(-\ln L + K)$, where K is the number of parameters that are estimated from the data; thereafter, Δ AIC $= 2(-\Delta \ln L + \Delta K)$. The model that best approximates the data is the one with smallest AIC.

[b] This is the shape parameter in a Γ-distribution, which is traditionally described as α. The use of α_Γ is to distinguish it from the symbol α used in the instantaneous rate matrix.

[c] The number of gene families for each clade is shown in parentheses underneath the clade name.

TABLE 3

Maximum log-likelihood comparison of different evolutionary models considering only gene presence/absence (*p*/*a*)

			Bacillaceae			Clostridium			Escherichia			Pseudomonas		
Scenario	Model	Parameter	B1	B2	B3	C1	C2	C3	E1	E2	E3	P1	P2	P3
$f \rightarrow a$	M_{00}[a]	μ	1.886	0.463	0.388	5.459	1.061	0.441	12.125	8.250	1.723	1.450	1.038	0.870
		ln L	−13,682	−15,268	−21,207	−6,491	−19,623	−23,554	−10,428	−10,080	−12,546	−14,607	−24,708	−26,670
	M_0[a]	μ	2.474	0.483	0.381	7.611	1.041	0.444	17.150	11.306	2.495	1.736	1.084	0.884
		ln L	−13,970	−15,487	−20,743	−6,444	−19,359	−22,168	−10,457	−10,195	−12,766	−14,961	−25,029	−26,803
	$M_0 + \pi$	μ	1.842	0.185	0.173	7.514	0.440	0.093	16.075	9.785	1.570	1.138	0.577	0.542
		π_a	0.510	0.816	0.831	0.286	0.807	0.933	0.345	0.407	0.534	0.616	0.745	0.725
		ln L	−13,681	−14,866	−20,017	−6,404	−18,705	−21,012	−10,362	−10,062	−12,543	−14,572	−24,253	−26,154
		Δ ln L[b]	289	621	726	40	654	1156	95	133	223	389	776	649
$f \rightarrow p$	M_{00}	μ	1.490	0.393	0.356	4.593	0.912	0.416	8.072	6.969	1.405	1.289	0.894	0.741
		ln L	−12,470	−14,472	−20,761	−5,902	−19,171	−23,250	−8,578	−9,293	−11,857	−13,785	−23,565	−25,408
	M_0	μ	2.045	0.419	0.337	6.747	0.873	0.377	12.813	10.198	2.116	1.605	0.956	0.762
		ln L[b]	−12,579	−14,613	−20,443	−5,794	−18,990	−22,318	−8,425	−9,296	−11,944	−13,989	−23,749	−25,478
	$M_0 + \pi$	μ	1.681	0.209	0.191	6.686	0.471	0.156	12.765	9.263	1.555	1.207	0.639	0.567
		π_a	0.442	0.745	0.774	0.242	0.753	0.857	0.238	0.341	0.460	0.534	0.657	0.639
		ln L	−12,460	−14,274	−20,070	−5,784	−18,635	−21,804	−8,425	−9,249	−11,853	−13,782	−23,402	−25,243
		Δ ln L[b]	119	339	373	10	355	514	0	47	91	207	347	235

In the two extreme scenarios, all truncated genes were entirely classified as absent ($f \rightarrow a$) or present ($f \rightarrow p$).

[a] As described in METHODS, model M_{00} is the special case of $M_0 + \pi$ when $\pi_a = \pi_p = 0.5$, while M_0 is the special case of $M_0 + \pi$ when π_a and π_p are the frequencies of each character in the data.

[b] Δ ln L's are shown as ($M_0 + \pi$) *vs.* M_0. Since $2\Delta \ln L \approx \chi^2$, Δ ln L's > 1.97 (d.f. = 1) are considered significant.

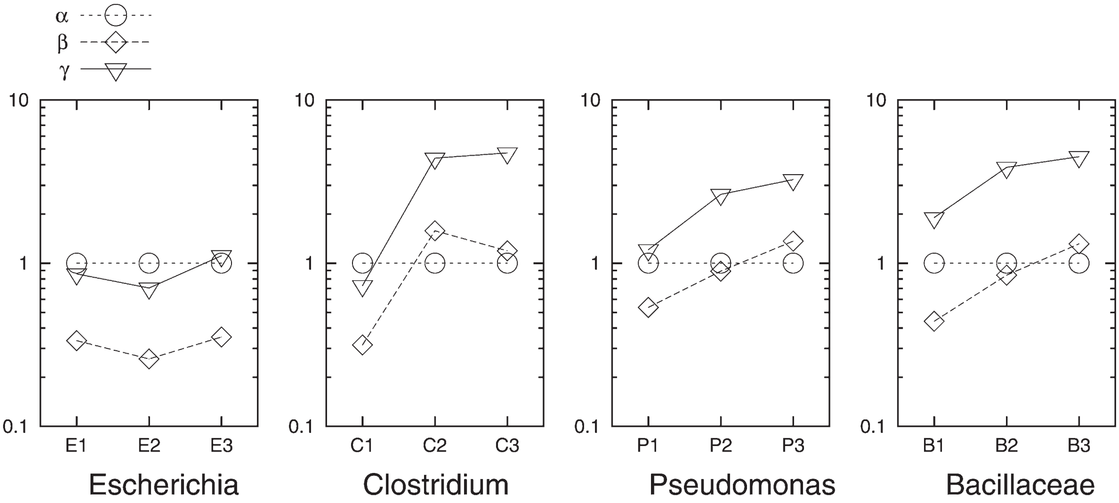


Figure 3.—Estimated parameters of the instantaneous rate matrix in each clade. As described in methods, α was fixed to be 1, and β and γ were estimated under the $M_0 + \pi$ model.

Since the number of gene insertions/deletions is proportional to μt, here t is the branch length; when the rate of gene insertions/deletions μ is constant, the number of gene insertions/deletions would be proportional to the corresponding branch length. The assumption of a constant rate μ would result in a bias that high numbers of gene insertions/deletions are inferred on the fast-evolving branches. This bias exists as long as the members of a clade do not evolve at the exact same rate. Clearly, the members in each studied clade do not all evolve at the exact same rate since none of the studied clades support a strict molecular clock tree (Figure 1). Furthermore, previous studies have shown that the inferred rates of gene insertions/deletions are not constant, and instead, recent branches tend to have higher numbers of gene insertions/deletions (Hao and Golding 2006). Given the use of a simplifying model in the study, it is essential to address the robustness of the results upon such an assumption. Our findings reveal that there is a strong negative association between the rate parameter μ and the degree of divergence in the clade (Figure 2). The same trend was found in a previous study on groups with different sets of genomes (Hao and Golding 2008b). Both studies showed that closely related clades tend to have high rates of gene insertions/deletions, suggesting many recently transferred genes are to be rapidly deleted from the genome. Importantly, the same conclusion can be drawn by comparing the rates of gene insertions/deletions between recent branches and ancient branches in the same phylogeny. Table 4 shows that the rates of gene insertions/deletions on external branches, when significant, are always higher than those on internal branches. This pattern has also been found in Hao and Golding (2006) and Marri *et al.* (2006). Furthermore, branch lengths, when estimated from the sequence data, have often been systematically overestimated on recent branches (Ho *et al.* 2005). In our study, we observed high rates of gene insertions/deletions on recent branches. One can easily imagine that the high rates of gene insertions/deletions on recent branches will be further inflated after correcting for the overestimation of the recent branch lengths.

Currently our method assumes reversibility in the transition processes among genes present (p), fragment (f), and absent (a). This is not likely realistic. For instance, the transition from p to f could easily be explained by gene truncation, while the reverse is not so clear. In our data, the state changes from f to p could result from homologous recombination or acquisition of a new full-length homolog. First, homologous recombination has been widely reported in bacterial genomes, especially between closely related strains (Guttman and Dykhuizen 1994; Gogarten *et al.* 2002; Fraser *et al.* 2007; Lefebure and Stanhope 2007), and recombination could involve long stretches of sequences (Didelot *et al.* 2007; Sheppard *et al.* 2008). When the recombinant sequences are long, truncated genes embedded in the recombinant region could potentially be converted to full-length genes. Second, the three different states are for gene families rather than for individual genes, and full-length genes with "truncated" paralogs are always classified as p. As a consequence, acquisition of full-length homolog(s) will result in the change of gene state from f to p. Since the rates of gene insertions are high (Hao and Golding 2006) and genes with high duplicability are more prone to gene transfer (Wellner *et al.* 2007), it should be appropriate to consider the transition from f to p. Although the transitions and their reverse forms are all possible, there is no good reason to believe that the actual transitions are mathematically reversible. Our current model assumes reversibility and assigns a single instantaneous rate parameter to both directions of each transition. For instance, the instantaneous rate parameter for the state pair fp is γ (Equation 1). Further improvement can be made in future studies by distinguishing the two directions of each transition and ultimately introducing an irreversible rate matrix. Future studies by incorporating an irreversible rate matrix

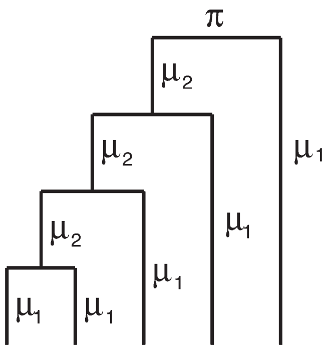


FIGURE 4.—Rate parameters estimated on a five-taxon phylogeny. Rates on external branches are μ_1, and rates on internal branches are μ_2.

would be able to further improve the MLEs and address how asymmetric each transition is in the instantaneous rate matrix. In the data *f*'s are very much outnumbered by *a*'s and *p*'s (Table S2, Table S3, Table S4, Table S5, Table S6, Table S7, Table S8, and Table S9). We suspect that the asymmetry between *f* and *p* and between *f* and *a* might not result in dramatic changes of MLEs when genome size remains roughly constant. On the flip side, when genome size varies significantly among taxa, models incorporating an irreversible rate matrix would be highly desirable. Furthermore, our current study assumes one instantaneous rate matrix on the entire phylogeny. This is also not likely realistic, especially when genome size varies among genomes (as shown in Figure S9). A more thorough study on genome size variation is in progress and will be reported later. Future studies by incorporating an irreversible rate matrix would shed new light on understanding the dynamics of genome size during bacterial genome evolution.

This study models insertions/deletions (or gains/losses) of gene families and requires the identification of the full-length gene in at least one genome in each examined clade. Recently DIDELOT *et al.* (2009) presented a method to reconstruct genomic flux on the basis of raw genomic sequences without relying on gene identification. In their study, each sequence rather than each gene was treated as a unit and sequence gains/losses were modeled on the basis of the presence or the absence of each sequence unit. One advantage of their method is its ability to model genomic flux beyond the gene boundary, since gene transfer could occur both at subgenic levels (RILEY and LABEDAN 1997; MILLER *et al.* 2005; CHAN *et al.* 2009) and in large gene clusters (LAWRENCE 1999). However, their model, as with previous models that consider a gene family as a unit, does not allow any intermediate states other than sequence presence or absence. In contrast, our study identifies truncated genes by comparing the full-length gene in a closely related species and should yield more accurate estimates of gene insertion/deletion events. Adding an intermediate state and considering insertions/deletions in the unit of genes, our method has a

TABLE 4

Rate comparison on different branches under the $M_0 + \pi$ model

	Bacillaceae			Clostridium			Escherichia			Pseudomonas		
Model	B1	B2	B3	C1	C2	C3	E1	E2	E3	P1	P2	P3
$\mu_1 = \mu_2$	2.102	0.214	0.197	8.385	0.482	0.149	19.805	11.699	1.837	1.332	0.695	0.619
ln L	−16,243.9	−17,799.0	−24,069.1	−7,384.6	−22,112.1	−26,490.9	−12,924.1	−11,875.9	−14,822.4	−16,787.4	−29,169.0	−31,715.2
μ_1	3.169	0.215	0.200	10.949	0.661	0.146	22.154	15.765	2.533	1.750	0.764	0.644
μ_2	0.180	0.210	0.180	0.000058	0.199	0.158	5.797	2.161	0.737	0.556	0.454	0.558
ln L	−15,879.1	−17,798.9	−24,067.8	−7,233.0	−21,879.1	−26,489.6	−12,888.7	−11,722.1	−14,722.1	−16,608.3	−29,128.4	−31,710.8
Δ ln L[a]	364.8	0.1	1.3	151.6	233.0	1.43	35.4	153.8	100.3	79.1	40.6	4.4

[a] Δ ln L's > 1.94 (d.f. = 1) are considered significant. As a result, B2, B3, and C3 are not significant.

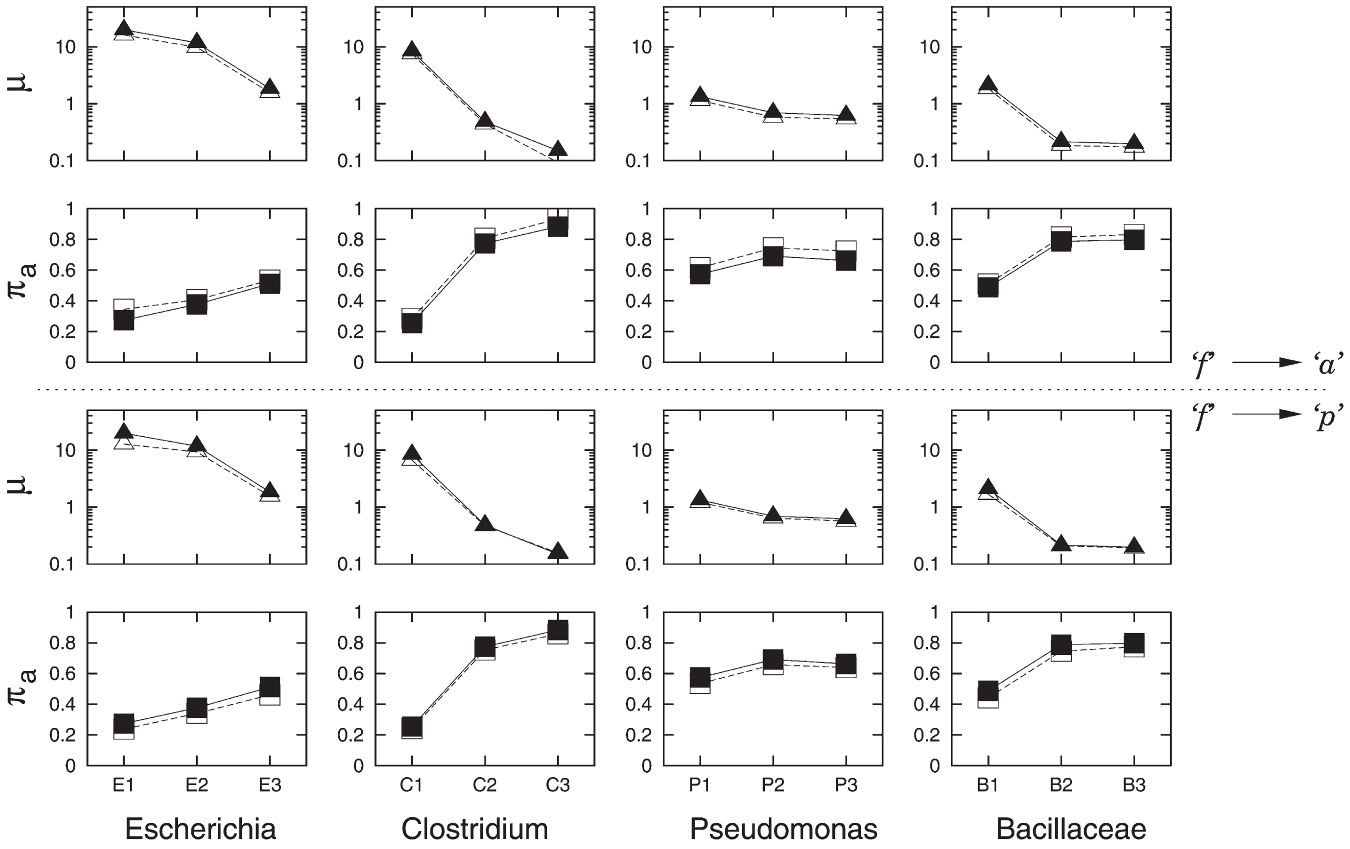


Figure 5.—Comparison of MLEs with (solid symbols) or without (open symbols) considering truncated genes in the model. MLEs were estimated under the $M_0 + \pi$ model. When not considering truncated genes, truncated genes were entirely classified as absent (top half) or present (bottom half).

potential to model gene decay, which could be frequent and rapid in some genomes (Cole *et al.* 2001; Dagan *et al.* 2006).

It is noteworthy that the rates of gene insertion/deletion were estimated from the data of currently present gene families. If gene deletion largely takes place in recently transferred genes, the number of anciently transferred genes will decrease during the course of evolution and will be reflected by a small number of gene gains or a slow rate of gene gains/losses in estimations (for a detailed illustration, see Hao and Golding 2010). This study reveals that closely related clades have high rates of gene insertions/deletions, while distantly related clades have low rates of gene insertions/deletions (Figure 2). This suggests that the fate of many recently transferred genes is to be deleted from the genome. When rates were distinguished between external branches and internal branches, the rate on external branches is, when significant, always higher than the one on internal branches (Table 4). These data are consistent with our previous observations (Hao and Golding 2004, 2006) that many of the recently transferred genes have a fast turnover. Several lines of evidence have previously documented that truncated genes are involved in the fast turnover of laterally transferred genes. An early study has shown that truncated genes arose from failed lateral gene transfer events (Liu *et al.* 2004). We reported that many of the truncated genes are recently acquired into the host genome (Hao and Golding 2008a) and are associated with gene translocation and gene deletion (Hao and Golding 2009). To address whether a false diagnosis of gene absence leads to systematic overestimation of any estimates, the maximum-likelihood estimates considering truncated genes were compared with the estimates after forcing truncated genes to be classified as either absent or present (Figure 5). Classifying truncated genes as absent does not always yield smaller rates (μ) than classifying truncated genes as present. As suggested in Hao and Golding (2008a), the effect of false diagnoses is not systematically biased, but rather complex. A more thorough understanding of the effect of false diagnoses requires further studies. Under the $M_0 + \pi$ model in the study, the rate parameter μ with truncated genes is almost always (with one exception in C3) slightly higher than the rate when classifying truncated genes as either absent or present (Figure 4). We believe that the higher rates observed when considering truncated genes are likely due to the richness of parameters. An analogous situation is often seen in observing greater nucleotide substitution distances when the substitution model is more parameter rich [*e.g.*, from

JC (Jukes and Cantor 1969) to K2P (Kimura 1980) and then to HKY (Hasegawa *et al.* 1985)].

The expected equilibrium frequencies of the three character states (π_a, π_f, π_p) were also assumed to be the frequencies at the ancestral node. Although the stationary probabilities are associated with the empirical frequencies of the character states in the data, they may deviate from the empirical frequencies since the stationary probabilities have taken into account the gene families that were once present in the ancestral genome but are no longer observable in the current data. One should expect a higher frequency of *a* in the stationary probability than in the empirical data. In fact, the stationary probability π_a is always higher than the observed frequency of *a* (Figure S7). One should not misinterpret it as any systematic bias that might favor more gene losses. In the results, the stationary probability π_a is positively associated with the tree length of each clade (Figure 2). These data suggest that a large number of ancient gene families have been lost in highly diverse clades, while only a small number of ancient gene families have been lost in low diversity clades. However, the large value of π_a in a more diverse clade might not necessarily suggest a smaller ancestral genome size of the clade, since an accurate estimation of ancestral genome size relies on the total number of gene families including the absolute number of unobservable patterns. When truncated genes were forced to be classified as present, the estimate of π_a became smaller in every clade compared to when truncated genes were considered. While if truncated genes were forced to be classified as absent, the estimate of π_a became larger in every clade compared to when truncated genes were considered (Figure 2). These estimates might be a reflection of the change of frequency of state *a* among clades.

It is widely acknowledged that gene family data contain phylogenetic signals (Fitz-Gibbon and House 1999; Snel *et al.* 1999; Tekaia *et al.* 1999) and many studies have used such data for the reconstruction of phylogenetic trees (Dutilh *et al.* 2004; Gu and Zhang 2004; Huson and Steel 2004; Zhang and Gu 2004; Spencer *et al.* 2007a,b) and phylogenetic topologies in more complex forms (Rivera and Lake 2004; Lake 2008). Even though the primary purpose of our study is to infer the dynamics of gene content during bacterial genome evolution, the model incorporating truncated genes could be potentially applied for phylogeny reconstruction using gene family data. We note that using three characters (*p*, *f*, *a*), compared with using only two characters (*p*, *a*) increased the –ln *L* values (Tables 2 and 3). In Figure S10, we plotted the –ln *L* values of using two gene characters against the ln *L* differences after adding the third gene character *f*. It is clear that there is a significantly positive association between the –ln *L* values and the ln *L* differences. This suggests a significant improvement in the probability of observing the data given three rather than two characters.

Given the nature of high rates of gene insertions/deletions in closely related species and low rates of gene insertions/deletions in distantly related species, the model presented in this study is expected to be useful among closely related taxa but less so for deep phylogeny questions.

The results from the improved model reveal fast rates of gene insertions/deletions/truncations on recent branches. This holds true when comparing different rates both between internal branches and external branches and among clades with different levels of divergence. The estimates of the rate ratio parameters suggest that many recently truncated genes are in the process of being rapidly deleted from the genome. We also demonstrated that using simplifying models, in which truncated genes are classified as absent, does not result in a systematic bias, but has a complex effect on rate estimates. Furthermore, the improved model is sensitive to the variation of genome size, and it opens the door to more thorough and comprehensive studies on the variation and dynamics of genome size during bacterial genome evolution.

We thank the anonymous reviewers for their helpful comments on previous versions of this manuscript. This work was supported by a Natural Sciences and Engineering Council of Canada (NSERC) postdoctoral fellowship to W.H. and an NSERC research grant to G.B.G.

LITERATURE CITED

Altschul, S. F., T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Archibald, J. M., and A. J. Roger, 2002 Gene duplication and gene conversion shape the evolution of archaeal chaperonins. J. Mol. Biol. **316:** 1041–1050.

Berg, O. G., and C. G. Kurland, 2002 Evolution of microbial genomes: sequence acquisition and loss. Mol. Biol. Evol. **19:** 2265–2276.

Brochier, C., H. Philippe and D. Moreira, 2000 The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Genet. **16:** 529–533.

Chain, P. S., E. Carniel, F. W. Larimer, J. Lamerdin, P. O. Stoutland *et al.*, 2004 Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. Proc. Natl. Acad. Sci. USA **101:** 13826–13831.

Chan, C. X., R. G. Beiko, A. E. Darling and M. A. Ragan, 2009 Lateral transfer of genes and gene fragments in prokaryotes. Genome Biol. Evol. **2009:** 429–438.

Choi, I. G., and S. H. Kim, 2007 Global extent of horizontal gene transfer. Proc. Natl. Acad. Sci. USA **104:** 4489–4494.

Cohen, O., and T. Pupko, 2010 Inference and characterization of horizontally transferred gene families using stochastic mapping. Mol. Biol. Evol. **27:** 703–713.

Cohen, O., N. D. Rubinstein, A. Stern, U. Gophna and T. Pupko, 2008 A likelihood framework to analyse phyletic patterns. Philos. Trans. R. Soc. Lond. B Biol. Sci. **363:** 3903–3911.

Cole, S. T., K. Eiglmeier, J. Parkhill, K. D. James, N. R. Thomson *et al.*, 2001 Massive gene decay in leprosy bacillus. Nature **409:** 1007–1011.

Dagan, T., R. Blekhman and D. Graur, 2006 The "domino theory" of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. Mol. Biol. Evol. **23:** 310–316.

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

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

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

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

Didelot, X., A. Darling and D. Falush, 2009 Inferring genomic flux in bacteria. Genome Res. **19:** 306–317.

Dutilh, B. E., M. A. Huynen, W. J. Bruno and B. Snel, 2004 The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J. Mol. Evol. **58:** 527–539.

Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:** 1792–1797.

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

Felsenstein, J., 1992 Phylogenies from restriction sites: a maximum-likelihood approach. Evolution **46:** 159–173.

Felsenstein, J., 2001 Taking variation of evolutionary rates between sites into account in inferring phylogenies. J. Mol. Evol. **53:** 447–455.

Fitz-Gibbon, S. T., and C. H. House, 1999 Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. **27:** 4218–4222.

Fraser, C., W. P. Hanage and B. G. Spratt, 2007 Recombination and the nature of bacterial speciation. Science **315:** 476–480.

Fraser-Liggett, C. M., 2005 Insights on biology and evolution from microbial genome sequencing. Genome Res. **15:** 1603–1610.

Friedman, R., and A. L. Hughes, 2003 The temporal distribution of gene duplication events in a set of highly conserved human gene families. Mol. Biol. Evol. **20:** 154–161.

Garcia-Vallvé, S., A. Romeu and J. Palau, 2000 Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. **10:** 1719–1725.

Gogarten, J. P., W. F. Doolittle and J. G. Lawrence, 2002 Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. **19:** 2226–2238.

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

Guttman, D. S., and D. E. Dykhuizen, 1994 Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science. **266:** 1380–1383.

Hao, W., 2010 OrgConv: detection of gene conversion using consensus sequences and its application in plant mitochondrial and chloroplast homologs. BMC Bioinformatics **11:** 114.

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

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

Hao, W., and G. B. Golding, 2008a High rates of lateral gene transfer are not due to false diagnosis of gene absence. Gene **421:** 27–31.

Hao, W., and G. B. Golding, 2008b Uncovering rate variation of lateral gene transfer during bacterial genome evolution. BMC Genomics **9:** 235.

Hao, W., and G. B. Golding, 2009 Does gene translocation accelerate the evolution of laterally transferred genes? Genetics **182:** 1365–1375.

Hao, W., and G. B. Golding, 2010 Patterns of horizontal gene transfer in bacteria, pp. 49–60 in *Microbial Population Genetics*, edited by J. Xu. Caister Academic Press, Norfolk, UK.

Hao, W., and J. D. Palmer, 2009 Fine-scale mergers of chloroplast and mitochondrial genes create functional, transcompartmentally chimeric mitochondrial genes. Proc. Natl. Acad. Sci. USA **106:** 16728–16733.

Hasegawa, M., H. Kishino and T. Yano, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

Ho, S. Y., M. J. Phillips, A. Cooper and A. J. Drummond, 2005 Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol. Biol. Evol. **22:** 1561–1568.

Hooper, S. D., and O. G. Berg, 2003 Duplication is more common among laterally transferred genes than among indigenous genes. Genome Biol. **4:** R48.

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

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Keeling, P. J., and J. D. Palmer, 2001 Lateral transfer at the gene and subgenic levels in the evolution of eukaryotic enolase. Proc. Natl. Acad. Sci. USA **98:** 10745–10750.

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

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

Lake, J. A., 2008 Reconstructing evolutionary graphs: 3D parsimony. Mol. Biol. Evol. **25:** 1677–1682.

Lawrence, J., 1999 Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Curr. Opin. Genet. Dev. **9:** 642–648.

Lefebure, T., and M. J. Stanhope, 2007 Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. **8:** R71.

Lerat, E., and H. Ochman, 2004 Ψ-Φ: exploring the outer limits of bacterial pseudogenes. Genome Res. **14:** 2273–2278.

Lerat, E., and H. Ochman, 2005 Recognizing the pseudogenes in bacterial genomes. Nucleic Acids Res. **33:** 3125–3132.

Liu, Y., P. M. Harrison, V. Kunin and M. Gerstein, 2004 Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. Genome Biol. **5:** R64.

Marri, P. R., W. Hao and G. B. Golding, 2006 Gene gain and gene loss in *Streptococcus*: Is it driven by habitat? Mol. Biol. Evol. **23:** 2379–2391.

Marri, P. R., W. Hao and G. B. Golding, 2007 The role of laterally transferred genes in adaptive evolution. BMC Evol. Biol. **7** (Suppl. 1): S8.

Miller, S. R., S. Augustine, T. L. Olson, R. E. Blankenship, J. Selker *et al.*, 2005 Discovery of a free-living chlorophyll *d*-producing cyanobacterium with a hybrid proteobacterial/cyanobacterial small-subunit rRNA gene. Proc. Natl. Acad. Sci. USA **102:** 850–855.

Mira, A., H. Ochman and N. A. Moran, 2001 Deletional bias and the evolution of bacterial genomes. Trends Genet. **17:** 589–596.

Mirkin, B. G., T. I. Fenner, M. Y. Galperin and E. V. Koonin, 2003 Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol. Biol. **3:** 2.

Nelder, J. A., and R. Mead, 1965 A simplex method for function minimization. Comput. J. **7:** 308–313.

Nielsen, K. M., and J. P. Townsend, 2004 Monitoring and modeling horizontal gene transfer. Nat. Biotechnol. **22:** 1110–1114.

Novozhilov, A. S., G. P. Karev and E. V. Koonin, 2005 Mathematical modeling of evolution of horizontally transferred genes. Mol. Biol. Evol. **22:** 1721–1732.

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

Ogata, H., S. Audic, P. Renesto-Audiffren, P. E. Fournier, V. Barbe *et al.*, 2001 Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science **293:** 2093–2098.

R Development Core Team, 2008 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Richardson, A. O., and J. D. Palmer, 2007 Horizontal gene transfer in plants. J. Exp. Bot. **58:** 1–9.

Riley, M., and B. Labedan, 1997 Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. J. Mol. Biol. **268:** 857–868.

Rivera, M. C., and J. A. Lake, 2004 The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature **431:** 152–155.

Sheppard, S. K., N. D. McCarthy, D. Falush and M. C. J. Maiden, 2008 Convergence of *Campylobacter* species: implications for bacterial evolution. Science **320:** 237–239.

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

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

Spencer, M., and A. Sangaralingam, 2009 A phylogenetic mixture model for gene family loss in parasitic bacteria. Mol. Biol. Evol. **26:** 1901–1908.

Spencer, M., D. Bryant and E. Susko, 2007a Conditioned genome reconstruction: how to avoid choosing the conditioning genome. Syst. Biol. **56:** 25–43.

Spencer, M., E. Susko and A. J. Roger, 2007b Modelling prokaryote gene content. Evol. Bioinform. Online **2:** 157–178.

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

Tekaia, F., A. Lazcano and B. Dujon, 1999 The genomic tree as revealed from whole proteome comparisons. Genome Res. **9:** 550–557.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini *et al.*, 2005 Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. USA **102:** 13950–13955.

Wei, J., M. B. Goldberg, V. Burland, M. M. Venkatesan, W. Deng *et al.*, 2003 Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. Infect. Immun. **71:** 2775–2786.

Welch, R. A., V. Burland, G. Plunkett, 3rd, P. Redford, P. Roesch *et al.*, 2002 Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc. Natl. Acad. Sci. USA **99:** 17020–17024.

Wellner, A., M. N. Lurie and U. Gophna, 2007 Complexity, connectivity, and duplicability as barriers to lateral gene transfer. Genome Biol. **8:** R156.

Yang, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39:** 306–314.

Yap, V. B., and T. Speed, 2005 Rooting a phylogenetic tree with nonreversible substitution models. BMC Evol. Biol. **5:** 2.

Yap, W. H., Z. Zhang and Y. Wang, 1999 Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. J. Bacteriol. **181:** 5201–5209.

Zhang, H., and X. Gu, 2004 Maximum likelihood for genome phylogeny on gene content. Stat. Appl. Genet. Mol. Biol. **3:** 31.

Zhaxybayeva, O., C. L. Nesbo and W. F. Doolittle, 2007 Systematic overestimation of gene gain through false diagnosis of gene absence. Genome Biol. **8:** 402.

Zheng, D., and M. B. Gerstein, 2007 The ambiguous boundary between genes and pseudogenes: The dead rise up, or do they? Trends Genet. **23:** 219–224.

Communicating editor: D. Voytas

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.118448/DC1

Inferring Bacterial Genome Flux While Considering Truncated Genes

Weilong Hao and G. Brian Golding

DOI: 10.1534/genetics.110.118448

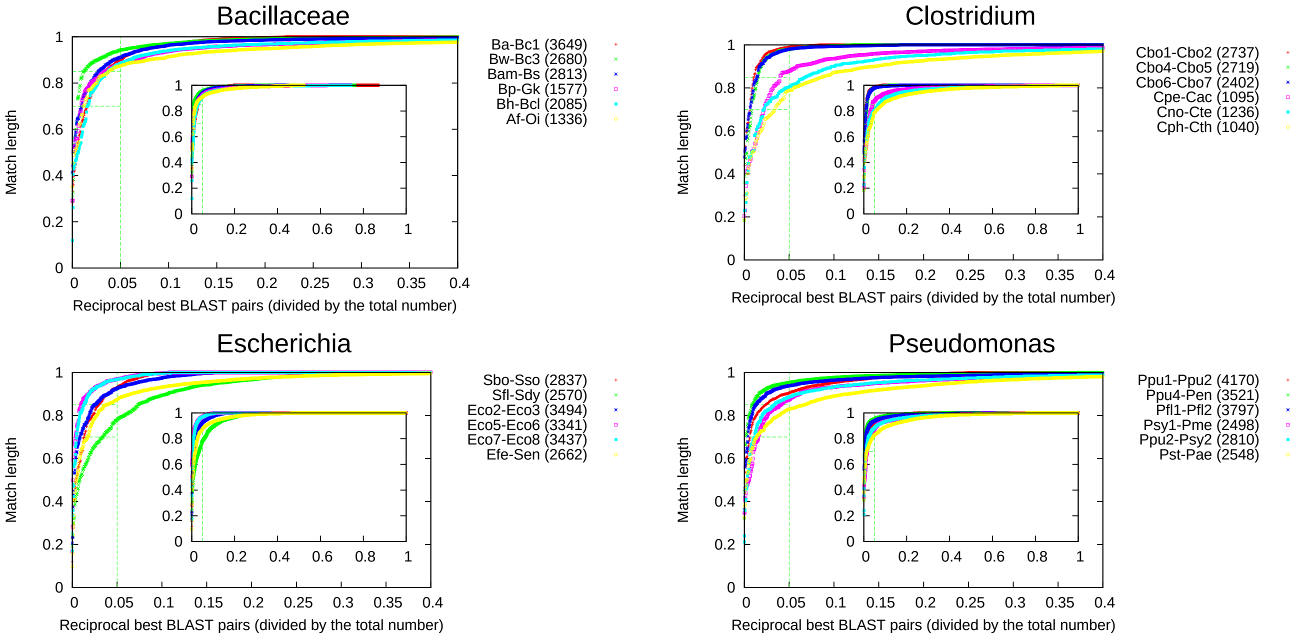


FIGURE S1.—Distribution of length variation in reciprocal best BLASTP pairs. From each clade, one closely related genome pair (the two innermost taxa) and one distantly related genome pair (the two outermost taxa on the phylogeny) were analyzed. All homologs are required to have an E-value $< 10^{-10}$ and only annotated genes were examined. Gene pairs were sorted by match length and divided by the total number (shown in parentheses). A vertical line was drawn at the lower 5% of the number of gene pairs and two horizontal lines were drawn at the match length of 85% and 70% respectively.

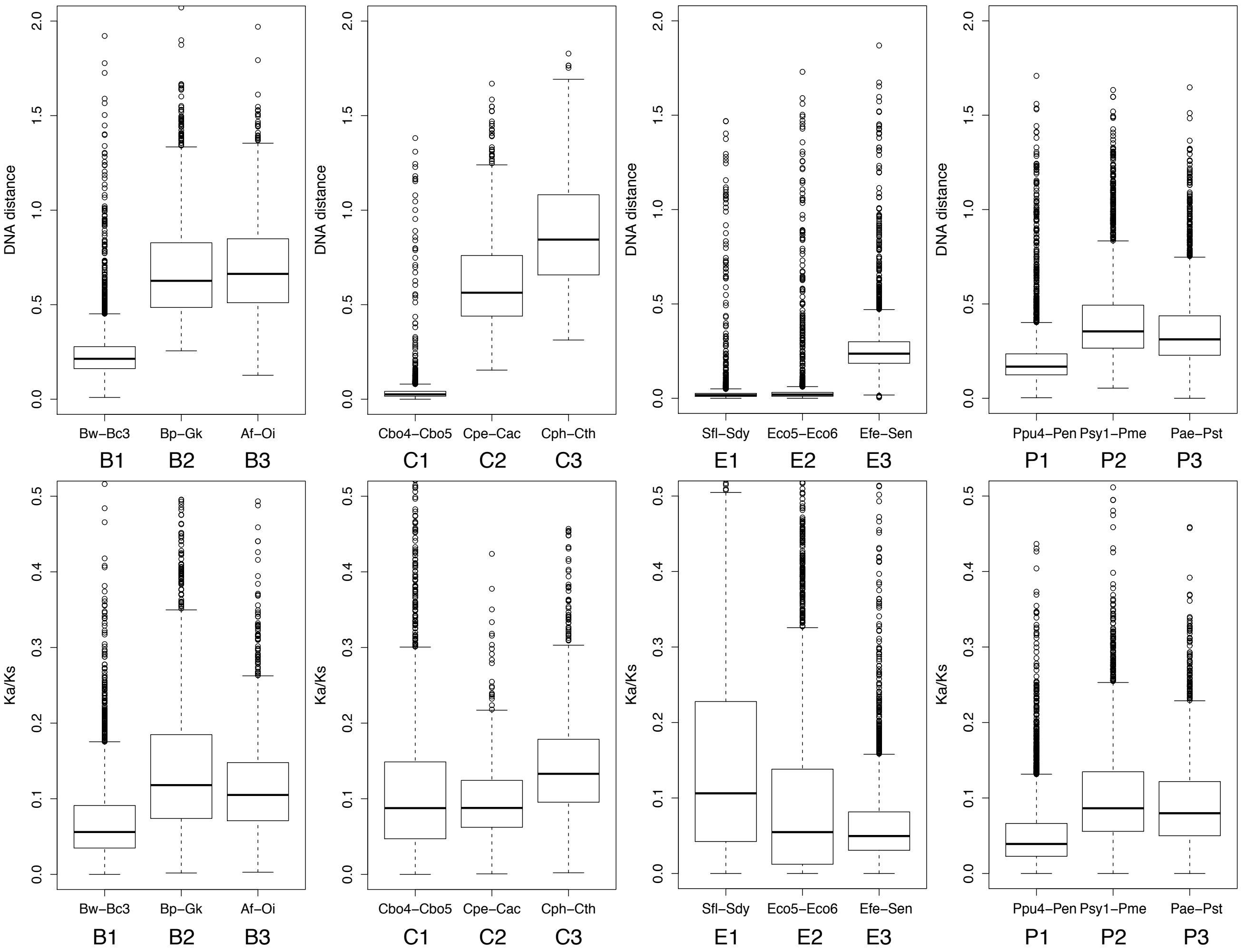


FIGURE S2.—Distribution of DNA distance and Ka/Ks (ω) ratio in reciprocal best BLASTP pairs. The gene pairs are from the distantly related genome pair (the two outermost taxa) in each clade as in Figure S.5. DNA distance was measured by DNADIST in the PHYLIP package, while the Ka/Ks ratio was measured by the Yang and Nielsen (2000) method, using yn00 in the PAML package.

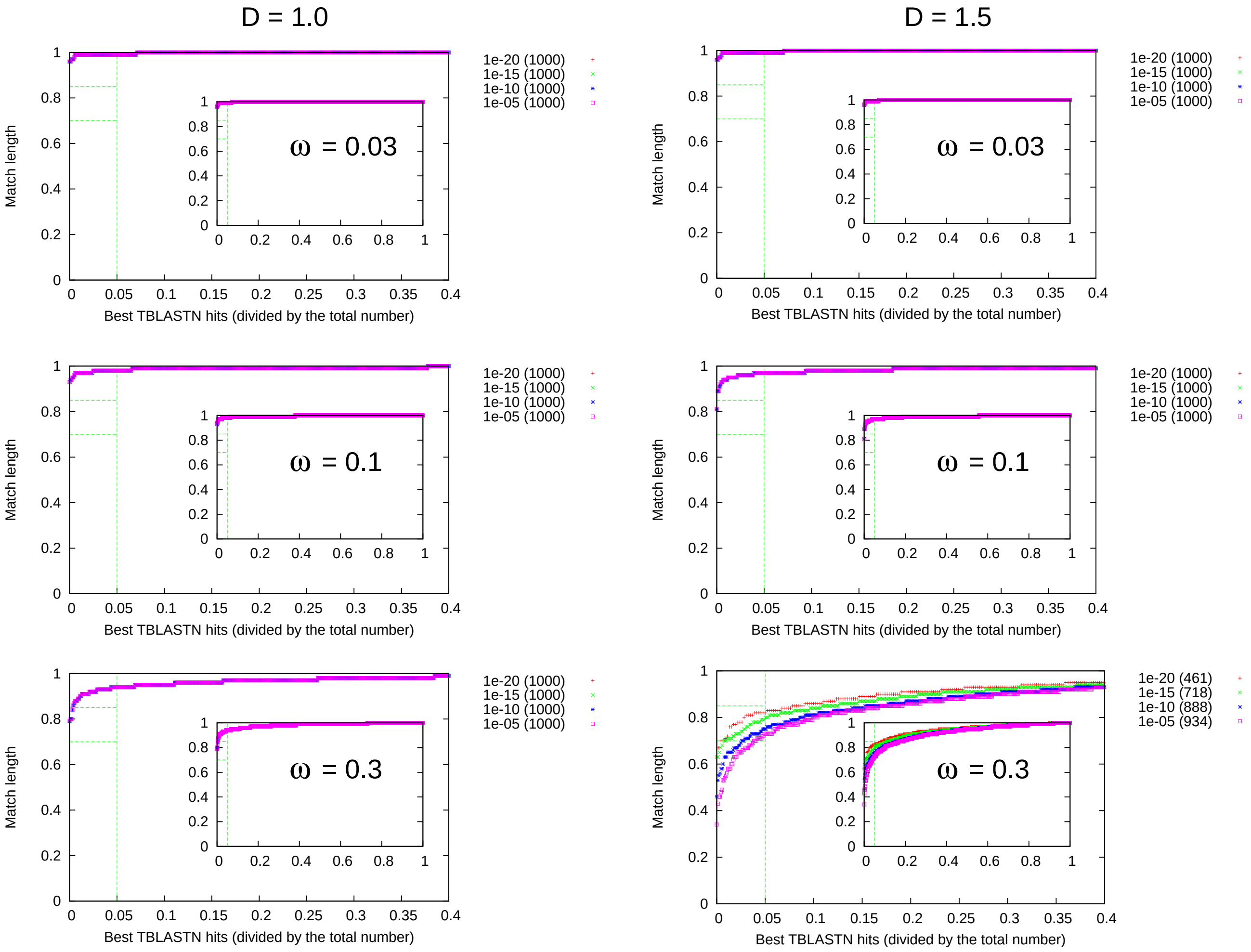


FIGURE S3.—Distribution of match length of `TBLASTN` hits in simulated sequences. Nucleotide sequences (350 aa in length) were simulated using `evolver` in the `PAML` package. Substitutions were introduced at given DNA distance and different ω (Ka/Ks) ratios to generate "evolved" sequences and no indels were allowed. Then the "ancestral" protein sequences were used as query sequences to `TBLASTN` the "evolved" DNA sequences and the best `TBLASTN` hits were plotted. Homologs were determined by a series of E-values (10^{-20}, 10^{-15}, 10^{-10}, and 10^{-05}) and the number of hits was shown in parentheses. 1000 iterations were conducted for each set of parameters.

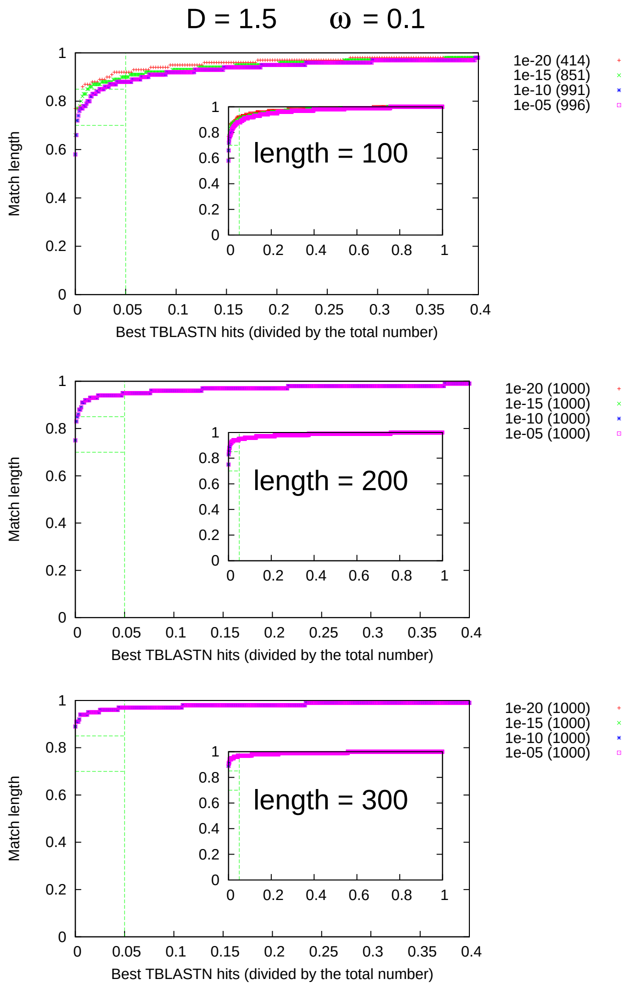


FIGURE S4.—Distribution of match length of `TBLASTN` hits associated with the length of the query sequences in simulated sequences. DNA distance was fixed to be 1.5 ($D = 1.5$) with $\omega = 0.1$. Three settings (100 aa, 200 aa, and 300 aa) for sequence length were simulated. Homologs were determined by a series of E-values (10^{-20}, 10^{-15}, 10^{-10}, and 10^{-05}) and the number of hits was shown in parentheses. 1000 iterations were conducted for each set of parameters.

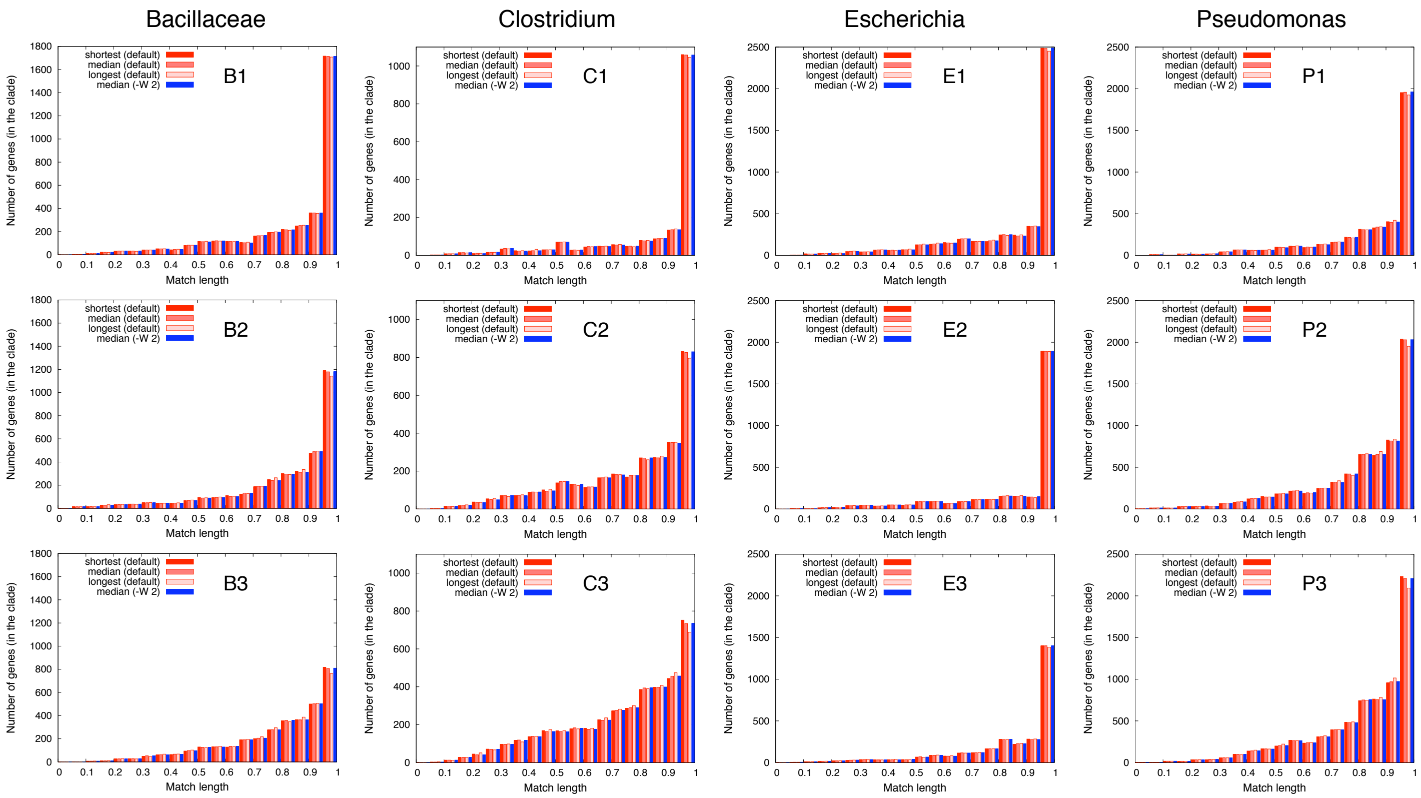


FIGURE S5.—Distribution of match length of TBLASTN hits. Query sequences (shortest, median, and longest in length) were chosen from each gene family clustered using parameters of E-value $< 10^{-20}$ and match length $> 85\%$, and a TBLASTN search in a genome was conducted when the genome does not have any annotated gene clustered in the family. Searches were performed using either the default (-W 3) or reduced (-W 2) parameter for word size. The numbers presented here are TBLASTN hits with an E-value $< 10^{-20}$ in each clade (containing 5 genomes).

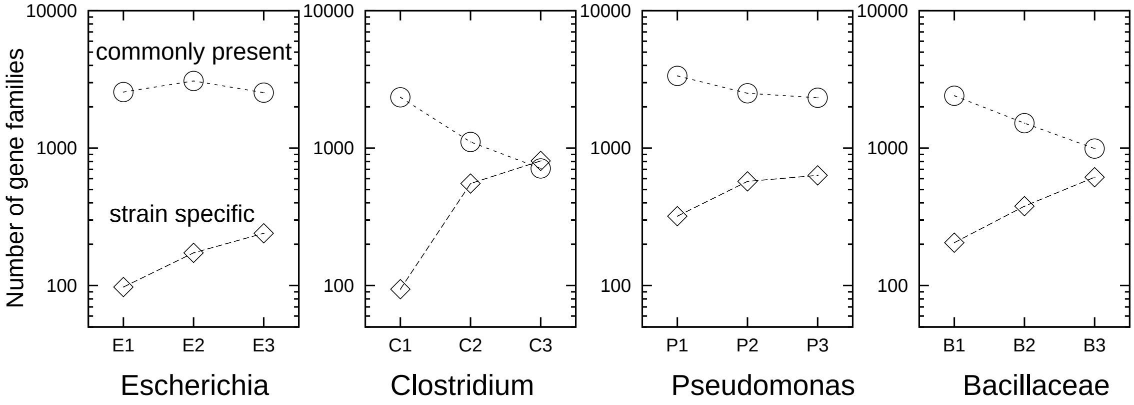


FIGURE S6.—Dynamics of gene families in different clades. Gene families shared by all five taxa in a clade (labeled as 'core') are shown in circles, while strain specific gene families are shown in diamonds.

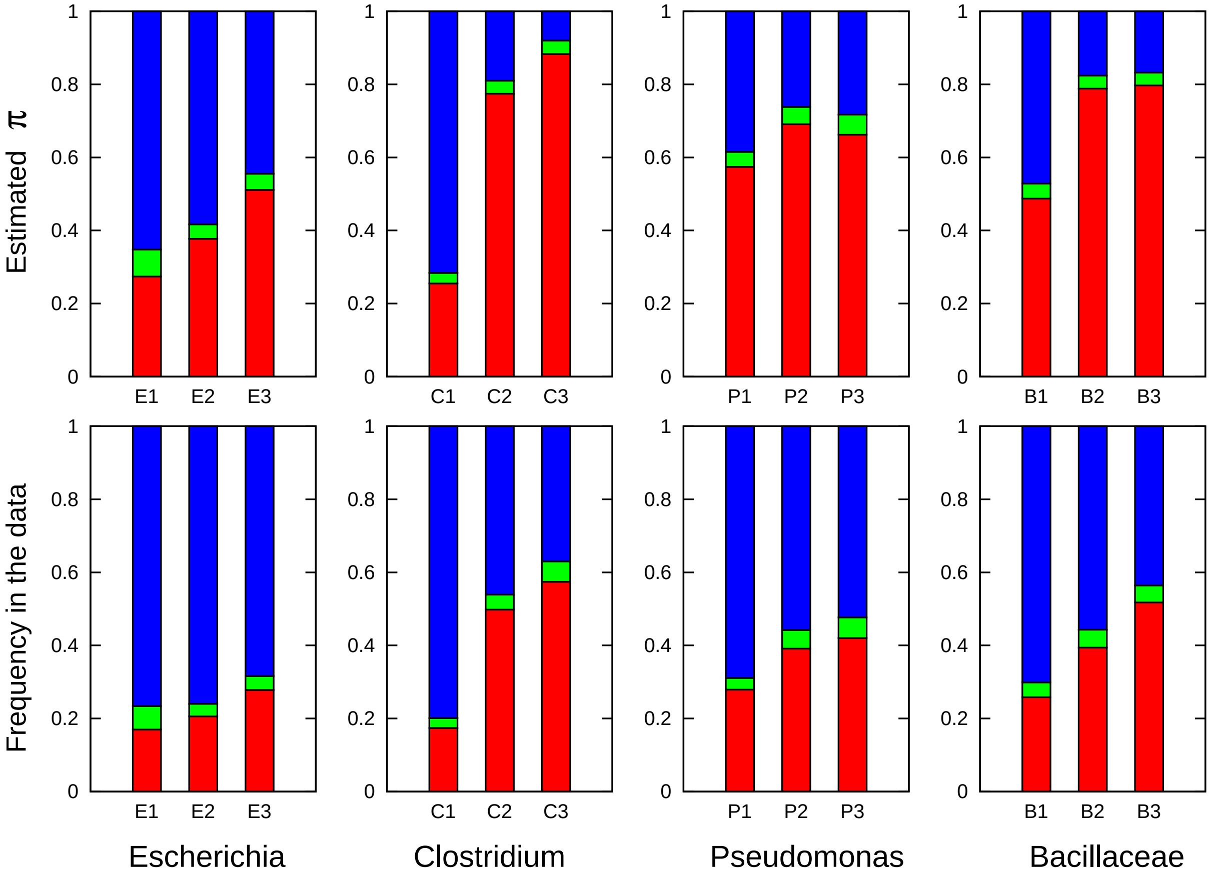


FIGURE S7.—Estimated stationary probabilities π and the observed character frequencies in each clade. 'a' is colored in red, 'f' is colored in green, and 'p' is colored in blue. The stationary probabilities (π) were estimated under the M_0+π model.

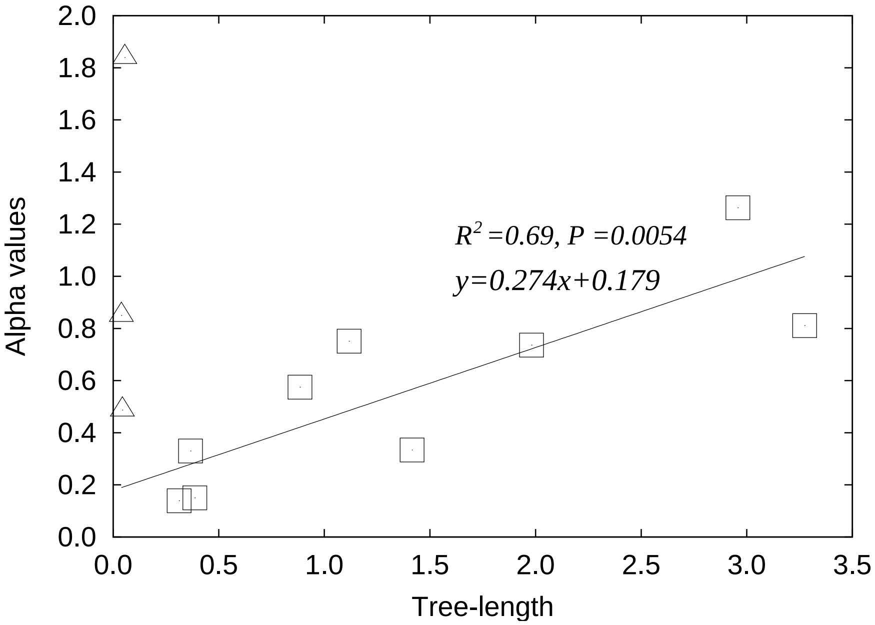
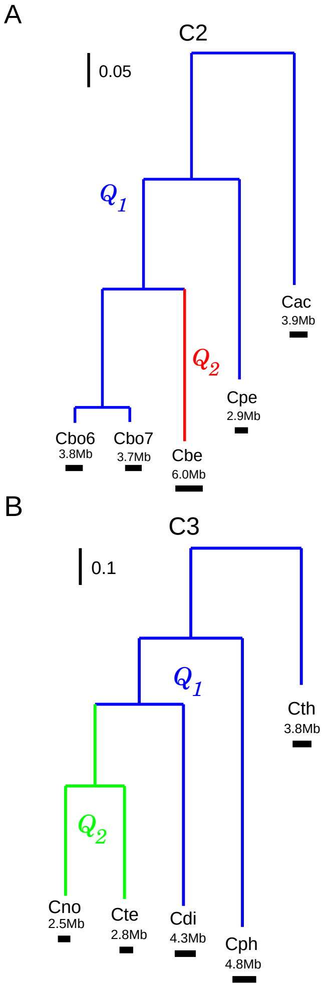


FIGURE S8.—Association between tree length and the optimized α in a Γ distribution in each clade. The three clades (E1, E2, and C1) with extremely small tree length (< 0.1) are shown in open triangles and excluded from the linear regression analysis.



MLEs	$Q_1 = Q_2$	(Q_1, Q_2)
μ	0.482	0.484
$\beta\ (Q_1)$	1.578	1.224
$\gamma\ (Q_1)$	4.397	2.964
$\beta\ (Q_2)$	-	2.442
$\gamma\ (Q_2)$	-	7.925
π_a	0.774	0.774
π_f	0.036	0.038
$\ln L$	-22112	-22101
ΔAIC	-18	

MLEs	$Q_1 = Q_2$	(Q_1, Q_2)
μ	0.149	0.158
$\beta\ (Q_1)$	1.191	1.486
$\gamma\ (Q_1)$	4.742	5.818
$\beta\ (Q_2)$	-	0.877
$\gamma\ (Q_2)$	-	3.555
π_a	0.883	0.876
π_f	0.037	0.037
$\ln L$	-26491	-26476
ΔAIC	-26	

FIGURE S9.—MLEs in the C2 (A) and C3 (B) clades when different transition matrices were considered on the branches associated with the strains with substantially different genome sizes. In the C2 clade, the branch leading to Cbe is colored in red. In the C3 clade, the Cno and Cte branches and the branch leading to their common ancestor are colored in green. The calculation of ΔAIC was shown in Table 2.

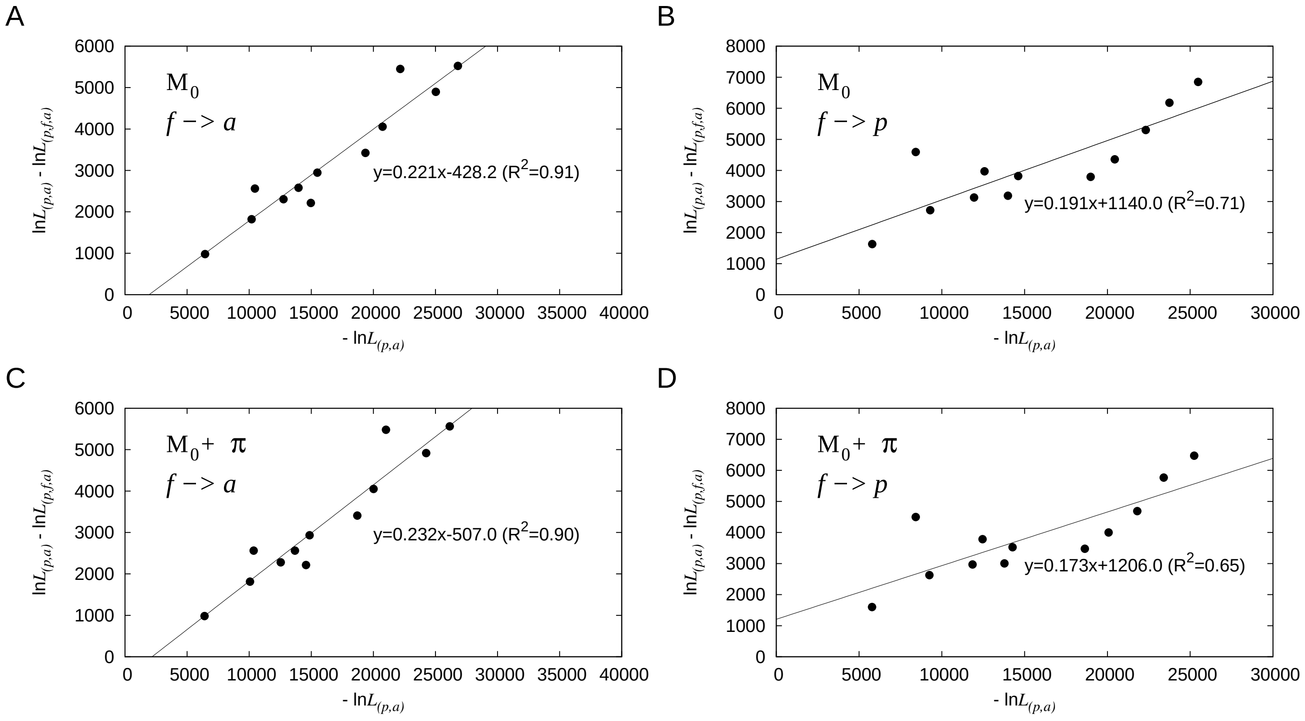


FIGURE S10.—Association between -ln$L$s of using two gene characters (p,a) and the ln$L$ differences caused by adding the third character 'f'. The ln$L$ values under model M_0 (A,B) and model M_0+π (C,D) were extracted from Tables 2 and 3. Both scenarios $f \rightarrow p$ (A,C) and $f \rightarrow a$ (B,D) were plotted for the two-character models. When 'f' was treated as 'p' (B,D), the E1 clade stands out as an outlier. This is likely, at least in part, due to the large number of pseudogenes in the Sfl genome in E1.

TABLE S1

List of outgroup species for each clade

Clade	Outgroup species	Accession
B1	*Bacillus halodurans*	NC_002570
B2	*Bacillus halodurans*	NC_002570
B3	*Lysinibacillus sphaericus*	NC_010382
C1	*Clostridium tetani* E88	NC_004557
	Clostridium kluyveri DSM 555	NC_009706
C2	*Clostridium cellulolyticum* H10	NC_011898
C3	*Moorella thermoacetica* ATCC 39073	NC_007644
E1	*Escherichia fergusonii* ATCC 35469	NC_011740
E2	*Escherichia fergusonii* ATCC 35469	NC_011740
E3	*Yersinia pestis* Antiqua	NC_008150
P1	*Pseudomonas mendocina* ymp	NC_009439
P2	*Pseudomonas aeruginosa* PA7	NC_009656
P3	*Azotobacter vinelandii* DJ	NC_012560

TABLE S2

Phylogenetic patterns in the Bacillaceae group (using a cutoff threshold of E-value $\leq 10^{-20}$ and match length $\geq 85\%$)

Number			B1			Number			B2			Number			B3		
of genes	Ba	Bc_1	Bc_2	Bw	Bc_3	of genes	Bam	Bs	Bl	Bp	Gk	of genes	Bh	Bcl	Oi	Es	Af
2408	*p*	*p*	*p*	*p*	*p*	1521	*p*	*p*	*p*	*p*	*p*	992	*p*	*p*	*p*	*p*	*p*
567	*p*	*p*	*p*	*p*	*a*	725	*a*	*a*	*a*	*a*	*p*	720	*a*	*a*	*p*	*a*	*a*
238	*a*	*a*	*a*	*a*	*p*	509	*p*	*p*	*p*	*p*	*a*	707	*a*	*p*	*a*	*a*	*a*
225	*a*	*a*	*p*	*a*	*a*	412	*a*	*a*	*p*	*a*	*a*	602	*p*	*a*	*a*	*a*	*a*
216	*a*	*p*	*a*	*a*	*a*	313	*a*	*a*	*a*	*p*	*a*	570	*a*	*a*	*a*	*a*	*p*
194	*a*	*a*	*a*	*p*	*a*	282	*a*	*p*	*a*	*a*	*a*	471	*a*	*a*	*a*	*p*	*a*
153	*p*	*a*	*a*	*a*	*a*	160	*p*	*a*	*a*	*a*	*a*	327	*p*	*p*	*a*	*a*	*a*
87	*p*	*p*	*p*	*p*	*f*	159	*p*	*p*	*a*	*a*	*a*	238	*p*	*p*	*a*	*p*	*p*
71	*p*	*p*	*p*	*a*	*a*	152	*p*	*p*	*p*	*a*	*a*	199	*p*	*p*	*a*	*a*	*p*
67	*p*	*a*	*p*	*p*	*a*	101	*p*	*p*	*p*	*p*	*f*	132	*a*	*p*	*a*	*a*	*p*
66	*p*	*a*	*a*	*p*	*a*	90	*a*	*p*	*p*	*a*	*a*	125	*p*	*p*	*p*	*a*	*p*
52	*p*	*p*	*a*	*a*	*a*	84	*a*	*a*	*p*	*p*	*a*	91	*p*	*a*	*a*	*p*	*a*
51	*a*	*p*	*p*	*p*	*a*	66	*p*	*p*	*a*	*p*	*a*	85	*p*	*a*	*a*	*a*	*p*
50	*p*	*a*	*p*	*a*	*a*	59	*a*	*p*	*p*	*p*	*a*	72	*a*	*a*	*p*	*p*	*a*
50	*a*	*a*	*p*	*p*	*a*	52	*f*	*f*	*f*	*f*	*p*	72	*a*	*a*	*p*	*a*	*p*
42	*f*	*p*	*p*	*p*	*p*	48	*p*	*p*	*p*	*a*	*p*	70	*p*	*p*	*a*	*p*	*a*
42	*p*	*p*	*a*	*p*	*a*	40	*a*	*p*	*a*	*p*	*a*	49	*a*	*a*	*a*	*p*	*p*
40	*a*	*a*	*p*	*a*	*p*	39	*a*	*a*	*p*	*a*	*p*	45	*a*	*p*	*p*	*a*	*p*
37	*p*	*p*	*f*	*p*	*p*	33	*p*	*p*	*a*	*a*	*p*	42	*p*	*p*	*p*	*p*	*a*
37	*a*	*p*	*a*	*p*	*a*	26	*p*	*f*	*p*	*p*	*p*	42	*p*	*p*	*p*	*a*	*a*
760	Other patterns					743	Other patterns					1162	Other patterns				

TABLE S3

Phylogenetic patterns in the *Clostridium* group (using a cutoff threshold of E-value $\leq 10^{-20}$ and match length $\geq 85\%$)

Number			C1			Number			C2			Number			C3		
of genes	Cbo$_1$	Cbo$_2$	Cbo$_3$	Cbo$_4$	Cbo$_5$	of genes	Cbo$_6$	Cbo$_7$	Cbe	Cpe	Cac	of genes	Cno	Cte	Cdi	Cph	Cth
2345	*p*	*p*	*p*	*p*	*p*	1110	*p*	*p*	*p*	*p*	*p*	1290*	*a*	*a*	*a*	*p*	*a*
114	*a*	*a*	*p*	*a*	*a*	977	*a*	*a*	*p*	*a*	*a*	994	*a*	*a*	*a*	*a*	*p*
111	*a*	*a*	*a*	*p*	*a*	892	*a*	*a*	*a*	*a*	*p*	949	*a*	*a*	*p*	*a*	*a*
99	*a*	*p*	*a*	*a*	*a*	436	*a*	*a*	*a*	*p*	*a*	712*	*p*	*p*	*p*	*p*	*p*
99	*a*	*a*	*a*	*a*	*p*	327	*a*	*a*	*p*	*a*	*p*	479	*a*	*p*	*a*	*a*	*a*
84	*a*	*p*	*a*	*p*	*a*	309	*p*	*p*	*p*	*a*	*a*	326	*p*	*a*	*a*	*a*	*a*
63	*a*	*p*	*p*	*p*	*p*	266	*p*	*p*	*a*	*a*	*a*	220	*p*	*p*	*a*	*a*	*a*
47	*p*	*a*	*a*	*a*	*a*	259	*p*	*a*	*a*	*a*	*a*	175	*a*	*a*	*a*	*p*	*p*
38	*a*	*a*	*p*	*a*	*p*	236	*p*	*p*	*p*	*a*	*p*	122	*a*	*a*	*p*	*p*	*a*
36	*p*	*p*	*p*	*a*	*p*	196	*a*	*p*	*a*	*a*	*a*	114	*p*	*p*	*p*	*a*	*a*
34	*p*	*p*	*p*	*p*	*a*	178	*p*	*p*	*p*	*p*	*a*	103	*a*	*p*	*p*	*a*	*a*
33	*f*	*p*	*p*	*p*	*p*	104	*p*	*a*	*p*	*a*	*a*	100	*p*	*p*	*p*	*p*	*a*
28	*a*	*p*	*a*	*a*	*p*	101	*p*	*p*	*a*	*p*	*a*	94	*p*	*p*	*p*	*a*	*p*
19	*p*	*p*	*a*	*p*	*p*	64	*a*	*a*	*p*	*p*	*a*	77	*a*	*a*	*p*	*a*	*p*
17	*p*	*p*	*p*	*f*	*p*	54	*a*	*a*	*p*	*p*	*p*	67	*a*	*a*	*a*	*p*	*f*
16	*p*	*p*	*a*	*a*	*p*	49	*a*	*a*	*a*	*p*	*p*	66	*a*	*p*	*p*	*p*	*a*
16	*p*	*a*	*a*	*a*	*p*	46	*f*	*f*	*p*	*f*	*f*	58	*p*	*p*	*a*	*a*	*p*
15	*p*	*a*	*p*	*a*	*a*	41	*p*	*p*	*p*	*f*	*p*	57	*p*	*a*	*p*	*p*	*p*
12	*p*	*p*	*f*	*p*	*p*	36	*a*	*p*	*p*	*a*	*a*	55	*a*	*p*	*a*	*p*	*a*
12	*a*	*p*	*p*	*a*	*a*	32	*p*	*p*	*p*	*p*	*f*	54	*a*	*a*	*p*	*p*	*p*
308			Other patterns			813			Other patterns			1533			Other patterns		

* Unlike in other clades, the pattern of the genes present in all taxa is *not* the most gene-family-rich pattern in C3.

TABLE S4

Phylogenetic patterns in the *Escherichia* group (using a cutoff threshold of E-value $\leq 10^{-20}$ and match length $\geq 85\%$)

Number			E1			Number			E2			Number			E3		
of genes	Sbo	Sso	Eco$_1$	Sfl	Sdy	of genes	Eco$_2$	Eco$_3$	Eco$_4$	Eco$_5$	Eco$_6$	of genes	Eco$_7$	Eco$_8$	Eco$_9$	Efe	Sen
2561	*p*	*p*	*p*	*p*	*p*	3082	*p*	*p*	*p*	*p*	*p*	2532	*p*	*p*	*p*	*p*	*p*
200	*a*	*a*	*p*	*a*	*a*	242	*a*	*a*	*a*	*p*	*a*	552	*a*	*a*	*a*	*a*	*p*
139	*p*	*p*	*p*	*p*	*a*	205	*a*	*p*	*a*	*a*	*a*	269	*a*	*a*	*a*	*p*	*a*
124	*p*	*p*	*p*	*p*	*f*	189	*a*	*a*	*a*	*a*	*p*	238	*p*	*p*	*p*	*p*	*a*
92	*a*	*p*	*p*	*a*	*a*	129	*a*	*a*	*p*	*a*	*a*	216	*a*	*a*	*p*	*a*	*a*
89	*a*	*a*	*a*	*p*	*a*	101	*p*	*a*	*a*	*a*	*a*	130	*p*	*p*	*p*	*a*	*a*
83	*p*	*a*	*a*	*a*	*a*	71	*p*	*p*	*p*	*a*	*a*	95	*p*	*a*	*a*	*a*	*a*
80	*a*	*a*	*a*	*a*	*p*	60	*p*	*p*	*p*	*p*	*a*	91	*p*	*p*	*p*	*p*	*f*
68	*f*	*p*	*p*	*p*	*p*	54	*p*	*p*	*p*	*a*	*p*	89	*p*	*p*	*p*	*a*	*p*
67	*p*	*p*	*p*	*f*	*p*	44	*a*	*a*	*p*	*p*	*p*	81	*p*	*p*	*a*	*a*	*a*
60	*a*	*p*	*p*	*p*	*a*	39	*p*	*a*	*p*	*p*	*p*	81	*a*	*a*	*p*	*p*	*a*
52	*p*	*p*	*p*	*a*	*p*	39	*a*	*p*	*p*	*p*	*p*	69	*a*	*p*	*a*	*a*	*a*
44	*a*	*p*	*p*	*p*	*p*	36	*a*	*p*	*a*	*a*	*p*	62	*a*	*a*	*a*	*p*	*p*
37	*p*	*f*	*p*	*p*	*p*	32	*f*	*p*	*p*	*p*	*p*	60	*a*	*p*	*p*	*a*	*a*
36	*a*	*p*	*a*	*a*	*a*	32	*p*	*p*	*p*	*f*	*p*	39	*f*	*f*	*f*	*f*	*p*
34	*p*	*p*	*p*	*a*	*a*	31	*a*	*p*	*p*	*a*	*a*	35	*p*	*p*	*a*	*p*	*a*
34	*a*	*a*	*p*	*p*	*a*	30	*a*	*p*	*a*	*p*	*a*	34	*a*	*a*	*p*	*a*	*p*
27	*f*	*p*	*p*	*p*	*f*	29	*a*	*a*	*p*	*a*	*p*	28	*p*	*a*	*p*	*p*	*a*
27	*p*	*p*	*a*	*p*	*p*	28	*p*	*p*	*p*	*p*	*f*	25	*a*	*p*	*p*	*p*	*a*
26	*p*	*p*	*a*	*p*	*a*	28	*p*	*a*	*p*	*a*	*p*	21	*f*	*f*	*f*	*p*	*f*
801	Other patterns					562	Other patterns					560	Other patterns				

TABLE S5

Phylogenetic patterns in the *Pseudomonas* group (using a cutoff threshold of E-value $\leq 10^{-20}$ and match length $\geq 85\%$)

Number			P1			Number			P2			Number			P3		
of genes	Ppu$_1$	Ppu$_2$	Ppu$_3$	Ppu$_4$	Pen	of genes	Pfl$_1$	Pfl$_2$	Pfl$_3$	Psy$_1$	Pme	of genes	Ppu$_2$	Psy$_2$	Pme	Pst	Pae
3354	*p*	*p*	*p*	*p*	*p*	2506	*p*	*p*	*p*	*p*	*p*	2325	*p*	*p*	*p*	*p*	*p*
526	*a*	*a*	*a*	*a*	*p*	670	*a*	*a*	*a*	*a*	*p*	872	*a*	*a*	*a*	*a*	*p*
333	*a*	*a*	*p*	*a*	*a*	637	*a*	*a*	*p*	*a*	*a*	791	*a*	*p*	*a*	*a*	*a*
317	*a*	*a*	*a*	*p*	*a*	559	*a*	*a*	*a*	*p*	*a*	658	*p*	*a*	*a*	*a*	*a*
243	*a*	*p*	*a*	*a*	*a*	511	*a*	*p*	*a*	*a*	*a*	436	*a*	*a*	*a*	*p*	*a*
208	*p*	*p*	*p*	*p*	*a*	487	*p*	*a*	*a*	*a*	*a*	413	*a*	*a*	*p*	*a*	*a*
178	*p*	*a*	*a*	*a*	*a*	485	*p*	*p*	*p*	*p*	*a*	227	*p*	*p*	*p*	*a*	*p*
105	*p*	*p*	*p*	*a*	*a*	284	*p*	*p*	*p*	*a*	*a*	188	*p*	*p*	*a*	*a*	*a*
92	*p*	*p*	*a*	*a*	*a*	217	*p*	*p*	*p*	*a*	*p*	178	*p*	*a*	*p*	*p*	*p*
78	*p*	*p*	*p*	*a*	*p*	175	*a*	*p*	*p*	*a*	*a*	172	*a*	*a*	*p*	*a*	*p*
71	*a*	*a*	*p*	*a*	*p*	163	*p*	*a*	*p*	*a*	*a*	151	*p*	*p*	*a*	*a*	*p*
65	*p*	*p*	*a*	*p*	*a*	88	*p*	*p*	*a*	*a*	*a*	140	*a*	*p*	*a*	*a*	*p*
58	*p*	*a*	*a*	*p*	*a*	87	*a*	*p*	*a*	*p*	*a*	136	*a*	*a*	*p*	*p*	*p*
56	*a*	*a*	*a*	*p*	*p*	81	*p*	*p*	*p*	*p*	*f*	129	*p*	*a*	*a*	*a*	*p*
46	*a*	*p*	*p*	*a*	*a*	78	*a*	*a*	*p*	*a*	*p*	129	*a*	*a*	*p*	*p*	*a*
44	*a*	*p*	*a*	*p*	*a*	66	*a*	*p*	*p*	*p*	*a*	109	*a*	*a*	*a*	*p*	*p*
39	*f*	*f*	*f*	*f*	*p*	63	*p*	*p*	*p*	*f*	*p*	88	*p*	*a*	*p*	*a*	*p*
33	*p*	*p*	*p*	*p*	*f*	59	*p*	*a*	*p*	*a*	*p*	80	*a*	*p*	*p*	*p*	*p*
30	*p*	*p*	*a*	*p*	*p*	57	*f*	*f*	*f*	*f*	*p*	66	*p*	*p*	*a*	*p*	*p*
30	*a*	*a*	*p*	*p*	*a*	55	*a*	*p*	*p*	*a*	*p*	66	*a*	*p*	*p*	*a*	*p*
599			Other patterns			1433			Other patterns			1786			Other patterns		

TABLE S6

Phylogenetic patterns in the Bacillaceae group (using a cutoff threshold of E-value $\leq 10^{-10}$ and match length $\geq 70\%$)

Number			B1			Number			B2			Number			B3		
of genes	Ba	Bc$_1$	Bc$_2$	Bw	Bc$_3$	of genes	Bam	Bs	Bl	Bp	Gk	of genes	Bh	Bcl	Oi	Es	Af
2168	p	p	p	p	p	1453	p	p	p	p	p	1017	p	p	p	p	p
428	p	p	p	p	a	555	a	a	a	a	p	521	a	a	p	a	a
187	a	a	p	a	a	372	p	p	p	p	a	467	a	p	a	a	a
184	a	a	a	a	p	301	a	a	p	a	a	426	p	a	a	a	a
172	a	p	a	a	a	232	a	a	a	p	a	403	a	a	a	a	p
148	a	a	a	p	a	210	a	p	a	a	a	327	a	a	a	p	a
116	p	a	a	a	a	116	p	a	a	a	a	242	p	p	a	a	a
77	p	p	p	p	f	106	p	p	p	a	a	200	p	p	a	p	p
59	p	p	p	a	a	99	p	p	a	a	a	161	p	p	a	a	p
53	p	a	p	p	a	67	a	p	p	a	a	108	a	p	a	a	p
50	p	a	a	p	a	60	a	a	p	p	a	103	p	p	p	a	p
41	p	a	p	a	a	59	p	p	p	p	f	72	p	a	a	p	a
37	p	p	a	a	a	57	p	p	a	p	a	69	a	a	p	p	a
36	p	p	a	p	a	54	a	p	p	p	a	62	p	p	a	p	a
36	a	a	p	p	a	51	p	p	p	a	p	61	a	a	p	a	p
34	a	p	p	p	a	35	a	a	p	a	p	59	p	a	a	a	p
34	a	a	p	a	p	33	a	p	a	p	a	46	p	p	p	p	a
27	a	p	p	a	a	27	p	p	a	a	p	39	p	a	p	p	p
26	a	p	a	p	a	26	f	f	f	f	p	36	p	a	p	a	a
24	f	p	p	p	p	19	p	a	p	p	a	33	p	p	p	a	a
540			Other patterns			516			Other patterns			855			Other patterns		

TABLE S7

Phylogenetic patterns in the *Clostridium* group (using a cutoff threshold of E-value $\leq 10^{-10}$ and match length $\geq 70\%$)

Number			C1			Number			C2			Number			C3		
of genes	Cbo$_1$	Cbo$_2$	Cbo$_3$	Cbo$_4$	Cbo$_5$	of genes	Cbo$_6$	Cbo$_7$	Cbe	Cpe	Cac	of genes	Cno	Cte	Cdi	Cph	Cth
2032	*p*	*p*	*p*	*p*	*p*	1071	*p*	*p*	*p*	*p*	*p*	900*	*a*	*a*	*a*	*p*	*a*
95	*a*	*a*	*p*	*a*	*a*	645	*a*	*a*	*p*	*a*	*a*	741	*a*	*a*	*a*	*a*	*p*
81	*a*	*a*	*a*	*p*	*a*	629	*a*	*a*	*a*	*a*	*p*	719*	*p*	*p*	*p*	*p*	*p*
77	*a*	*a*	*a*	*a*	*p*	316	*a*	*a*	*a*	*p*	*a*	646	*a*	*a*	*p*	*a*	*a*
73	*a*	*p*	*a*	*p*	*a*	251	*a*	*a*	*p*	*a*	*p*	347	*a*	*p*	*a*	*a*	*a*
71	*a*	*p*	*a*	*a*	*a*	225	*p*	*p*	*p*	*a*	*a*	214	*p*	*a*	*a*	*a*	*a*
41	*a*	*p*	*p*	*p*	*p*	217	*p*	*p*	*p*	*a*	*p*	164	*a*	*a*	*a*	*p*	*p*
38	*p*	*a*	*a*	*a*	*a*	199	*p*	*a*	*a*	*a*	*a*	134	*p*	*p*	*a*	*a*	*a*
37	*a*	*a*	*p*	*a*	*p*	194	*p*	*p*	*a*	*a*	*a*	128	*a*	*a*	*p*	*p*	*a*
29	*f*	*p*	*p*	*p*	*p*	154	*p*	*p*	*p*	*p*	*a*	99	*p*	*p*	*p*	*p*	*a*
28	*p*	*p*	*p*	*a*	*p*	153	*a*	*p*	*a*	*a*	*a*	89	*p*	*p*	*p*	*a*	*a*
27	*p*	*p*	*p*	*p*	*a*	95	*p*	*a*	*p*	*a*	*a*	78	*a*	*a*	*p*	*a*	*p*
25	*a*	*p*	*a*	*a*	*p*	79	*p*	*p*	*a*	*p*	*a*	77	*a*	*p*	*p*	*a*	*a*
14	*a*	*a*	*a*	*p*	*p*	53	*a*	*a*	*p*	*p*	*a*	73	*p*	*p*	*p*	*a*	*p*
12	*p*	*p*	*p*	*f*	*p*	52	*a*	*a*	*p*	*p*	*p*	61	*a*	*a*	*p*	*p*	*p*
12	*p*	*p*	*a*	*p*	*p*	47	*p*	*p*	*p*	*f*	*p*	57	*p*	*a*	*p*	*a*	*a*
12	*p*	*p*	*a*	*a*	*p*	42	*a*	*p*	*p*	*a*	*a*	55	*p*	*p*	*a*	*a*	*p*
10	*p*	*a*	*p*	*a*	*a*	38	*a*	*a*	*a*	*p*	*p*	54	*p*	*a*	*p*	*p*	*p*
10	*p*	*a*	*a*	*a*	*p*	33	*p*	*p*	*a*	*p*	*p*	51	*p*	*p*	*a*	*p*	*a*
9	*p*	*p*	*p*	*p*	*f*	26	*f*	*f*	*p*	*f*	*f*	51	*a*	*p*	*a*	*p*	*a*
269			Other patterns			694			Other patterns			1308			Other patterns		

[*] Unlike in other clades, the pattern of the genes present in all taxa is *not* the most gene-family-rich pattern in C3.

TABLE S8

Phylogenetic patterns in the *Escherichia* group (using a cutoff threshold of E-value $\leq 10^{-10}$ and match length $\geq 70\%$)

Number			E1			Number			E2			Number			E3		
of genes	Sbo	Sso	Eco_1	Sfl	Sdy	of genes	Eco_2	Eco_3	Eco_4	Eco_5	Eco_6	of genes	Eco_7	Eco_8	Eco_9	Efe	Sen
2311	p	p	p	p	p	2687	p	p	p	p	p	2250	p	p	p	p	p
160	a	a	p	a	a	207	a	a	a	p	a	453	a	a	a	a	p
110	p	p	p	p	a	178	a	p	a	a	a	215	a	a	a	p	a
73	a	p	p	a	a	144	a	a	a	a	p	197	p	p	p	p	a
71	p	p	p	p	f	108	a	a	p	a	a	175	a	a	p	a	a
70	a	a	a	p	a	84	p	a	a	a	a	86	p	p	p	a	a
64	p	a	a	a	a	63	p	p	p	a	a	79	p	a	a	a	a
60	a	a	a	a	p	45	p	p	p	a	p	76	p	p	p	a	p
50	a	p	p	p	a	41	a	p	p	p	p	66	p	p	p	p	f
47	f	p	p	p	p	40	p	p	p	p	a	62	p	p	a	a	a
45	p	p	p	a	p	40	a	a	p	p	p	58	a	p	a	a	a
38	a	p	p	p	p	32	a	p	a	p	a	57	a	a	p	p	a
36	p	p	p	f	p	31	a	p	p	p	a	51	a	a	a	p	p
31	p	p	p	a	a	28	p	p	p	f	p	45	a	p	p	a	a
28	a	a	p	p	a	27	a	a	p	a	p	35	p	p	a	p	a
27	a	p	a	a	a	26	p	a	p	a	p	31	p	a	p	p	a
24	p	p	a	p	a	25	a	p	p	a	a	28	a	p	p	p	a
23	p	a	p	a	a	24	a	p	a	a	p	26	f	f	f	f	p
23	a	p	a	p	a	24	a	a	a	p	p	24	a	a	p	a	p
21	f	f	p	f	f	23	p	a	p	p	p	19	p	p	a	a	p
537			Other patterns			425			Other patterns			435			Other patterns		

TABLE S9

Phylogenetic patterns in the *Pseudomonas* group (using a cutoff threshold of E-value $\leq 10^{-10}$ and match length $\geq 70\%$)

Number			P1			Number			P2			Number			P3		
of genes	Ppu$_1$	Ppu$_2$	Ppu$_3$	Ppu$_4$	Pen	of genes	Pfl$_1$	Pfl$_2$	Pfl$_3$	Psy$_1$	Pme	of genes	Ppu$_2$	Psy$_2$	Pme	Pst	Pae
2803	*p*	*p*	*p*	*p*	*p*	2214	*p*	*p*	*p*	*p*	*p*	2025	*p*	*p*	*p*	*p*	*p*
419	*a*	*a*	*a*	*a*	*p*	497	*a*	*a*	*p*	*a*	*a*	661	*a*	*a*	*a*	*a*	*p*
258	*a*	*a*	*p*	*a*	*a*	497	*a*	*a*	*a*	*a*	*p*	594	*a*	*p*	*a*	*a*	*a*
240	*a*	*a*	*a*	*p*	*a*	414	*a*	*a*	*a*	*p*	*a*	476	*p*	*a*	*a*	*a*	*a*
188	*a*	*p*	*a*	*a*	*a*	391	*a*	*p*	*a*	*a*	*a*	335	*a*	*a*	*a*	*p*	*a*
171	*p*	*p*	*p*	*p*	*a*	355	*p*	*a*	*a*	*a*	*a*	294	*a*	*a*	*p*	*a*	*a*
141	*p*	*a*	*a*	*a*	*a*	330	*p*	*p*	*p*	*p*	*a*	196	*p*	*p*	*p*	*a*	*p*
78	*p*	*p*	*p*	*a*	*a*	177	*p*	*p*	*p*	*a*	*a*	177	*p*	*p*	*a*	*a*	*a*
77	*p*	*p*	*a*	*a*	*a*	159	*p*	*p*	*p*	*a*	*p*	141	*p*	*a*	*p*	*p*	*p*
64	*p*	*p*	*p*	*a*	*p*	132	*a*	*p*	*p*	*a*	*a*	135	*a*	*a*	*p*	*a*	*p*
59	*p*	*p*	*a*	*p*	*a*	125	*p*	*a*	*p*	*a*	*a*	126	*p*	*p*	*a*	*a*	*p*
48	*a*	*a*	*p*	*a*	*p*	72	*a*	*p*	*a*	*p*	*a*	112	*a*	*p*	*a*	*a*	*p*
46	*a*	*a*	*a*	*p*	*p*	68	*p*	*p*	*a*	*a*	*a*	103	*a*	*a*	*p*	*p*	*p*
44	*p*	*a*	*a*	*p*	*a*	68	*a*	*a*	*p*	*a*	*p*	96	*p*	*a*	*a*	*a*	*p*
40	*a*	*p*	*a*	*p*	*a*	54	*p*	*a*	*p*	*a*	*p*	88	*a*	*a*	*p*	*p*	*a*
38	*f*	*f*	*f*	*f*	*p*	51	*p*	*p*	*p*	*p*	*f*	75	*a*	*a*	*a*	*p*	*p*
34	*a*	*p*	*p*	*a*	*a*	50	*a*	*p*	*p*	*p*	*a*	67	*p*	*a*	*p*	*a*	*p*
25	*p*	*a*	*p*	*a*	*a*	49	*a*	*a*	*a*	*p*	*p*	61	*a*	*p*	*p*	*p*	*p*
24	*a*	*a*	*p*	*p*	*a*	46	*f*	*f*	*f*	*f*	*p*	60	*a*	*p*	*p*	*a*	*p*
22	*a*	*p*	*p*	*p*	*p*	46	*p*	*a*	*p*	*p*	*a*	53	*a*	*p*	*p*	*a*	*a*
475			Other patterns			1027			Other patterns			1224			Other patterns		

TABLE S10

List of unobservable patterns*

taxon1	taxon2	taxon3	taxon4	taxon 5
a	a	a	a	a
a	a	a	a	f
a	a	a	f	a
a	a	a	f	f
a	a	f	a	a
a	a	f	a	f
a	a	f	f	a
a	a	f	f	f
a	f	a	a	a
a	f	a	a	f
a	f	a	f	a
a	f	a	f	f
a	f	f	a	a
a	f	f	a	f
a	f	f	f	a
a	f	f	f	f
f	a	a	a	a
f	a	a	a	f
f	a	a	f	a
f	a	a	f	f
f	a	f	a	a
f	a	f	a	f
f	a	f	f	a
f	a	f	f	f
f	f	a	a	a
f	f	a	a	f
f	f	a	f	a
f	f	a	f	f
f	f	f	a	a
f	f	f	a	f
f	f	f	f	a
f	f	f	f	f

*Identification of truncated genes requires a full length gene present in at least one taxon in the clade, as a consequence, genuinely truncated genes that do not have any full length homologues present in the clade would be unobservable in the study.

TABLE S11

Maximum log-likelihood comparison of different evolutionary models (using a cutoff threshold of E-value $\leq 10^{-10}$ and match length $\geq 70\%$)

Models and		Bacillaceae			Clostridium			Escherichia			Pseudomonas		
Parameters		B1	B2	B3	C1	C2	C3	E1	E2	E3	P1	P2	P3
		(4477)[§]	(4448)	(5307)	(3002)	(5213)	(6046)	(3849)	(4302)	(4468)	(5294)	(6822)	(7099)
M_0	μ	2.329	0.458	0.370	8.167	0.989	0.464	17.109	11.901	2.442	1.834	1.067	0.893
	β	0.523	1.014	1.342	0.269	1.424	1.787	0.348	0.319	0.412	0.465	0.923	1.110
	γ	1.003	1.466	2.075	0.473	2.093	2.713	0.607	0.515	0.707	0.639	1.248	1.670
	ln$L$	-12742	-13776	-18923	-6180	-18440	-22231	-9627	-9887	-12168	-13750	-22419	-24172
$M_0+\pi$	μ	1.945	0.233	0.216	8.372	0.609	0.211	17.196	10.870	1.737	1.429	0.692	0.633
	β	0.455	0.816	1.113	0.267	1.346	1.303	0.328	0.298	0.342	0.441	0.813	0.944
	γ	1.381	2.578	3.274	0.555	3.495	3.756	0.738	0.713	1.036	0.936	1.819	2.124
	π_a	0.446	0.729	0.740	0.234	0.691	0.812	0.246	0.372	0.480	0.538	0.659	0.639
	π_f	0.041	0.037	0.035	0.032	0.039	0.048	0.057	0.039	0.035	0.038	0.047	0.047
	ln$L$	-12522	-13351	-18466	-6153	-18016	-21557	-9567	-9769	-11979	-13474	-21898	-23736
	ΔAIC	-436	-846	-910	-50	-844	-1344	-116	-232	-374	-548	-1038	-868
$M_0+\Gamma+\pi$	μ	3.450	0.294	0.316	8.365	2.543	0.270	18.737	13.152	8.205	4.513	1.001	0.817
	α_Γ	0.318	0.616	0.702	1.919	0.283	1.200	0.862	0.464	0.147	0.149	0.470	0.705
	ln$L$	-12241	-13149	-18203	-6142	-17317	-21380	-9500	-9649	-11311	-13126	-21476	-23454
	ΔAIC	-560	-402	-524	-20	-1396	-352	-132	-238	-1334	-694	-842	-562

[†] ΔAICs are shown as ($M_0+\pi$) *vs.* M_0 and ($M_0+\Gamma+\pi$) *vs.* ($M_0+\pi$). By definition, AIC = 2(-ln$L$ + K), where K is the number of parameters that are estimated from the data, thereafter, ΔAIC = 2(-Δln$L$ + ΔK). The model that best approximates the data is the one with smallest AIC.

[§] The number of gene families for each clade is shown in parenthesis underneath the clade name.

TABLE S12

Product of the scaled instantaneous rate matrix Q and the rate parameter μ (or $Q\mu$) under model M_0+π

	Bacillaceae				Clostridium				Escherichia				Pseudomonas		
	a	f	p		a	f	p		a	f	p		a	f	p
B1	−1.655	0.273	1.382	C1	−11.171	2.578	8.593	E1	−23.810	6.025	17.785	P1	−1.023	0.170	0.853
	3.242	−9.212	5.970		22.670	−69.009	46.339		22.310	−67.966	45.656		2.376	−4.305	1.929
	1.426	0.519	−1.945		3.060	1.878	−4.937		7.474	5.182	−12.655		1.271	0.205	−1.476
B2	−0.116	0.023	0.093	C2	− 0.279	0.030	0.249	E2	−12.635	2.646	9.989	P2	−0.431	0.072	0.359
	0.496	−0.924	0.428		0.643	− 1.337	0.694		24.940	−52.091	27.151		1.060	−2.121	1.061
	0.419	0.087	−0.506		1.015	0.132	−1.147		6.459	1.863	− 8.322		0.946	0.190	−1.136
B3	−0.109	0.015	0.094	C3	− 0.075	0.021	0.054	E3	− 1.484	0.325	1.159	P3	−0.398	0.050	0.349
	0.341	−0.664	0.323		0.503	− 0.719	0.216		3.771	− 7.413	3.642		0.598	−1.430	0.832
	0.448	0.067	−0.515		0.599	0.100	−0.699		1.331	0.360	− 1.691		0.816	0.161	−0.977