Inferring bacterial genome flux while considering truncated genes

Weilong Hao¹, and G Brian Golding²

¹Department of Biology, Indiana University, Bloomington, Indiana, 47405 USA
²Department of Biology, McMaster University, Hamilton, Ontario, L8S4K1 Canada

Running head: Genome flux with gene truncation

Key words: gene family, gene truncation, lateral gene transfer, maximum likelihood, pseudogene

Corresponding author: Weilong Hao
Department of Biology
Indiana University
Bloomington, Indiana
47405 USA
E-mail: haoweilong@gmail.com
Tel: 812-855-2549
Abstract

Bacterial gene content variation during the course of evolution has been widely acknowledged and its pattern has been hotly 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 turn-over 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.

Introduction

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
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 alpha-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 homologues might erroneously lead to false inferences of “gene gain” rather than multiple instances of “gene loss”. This raises the question of how a false diagnosis of gene absence affects the estimation of insertion/deletion rates. Recently, we have shown 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 have 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 \(\Gamma\) distribution. We also allow rates to vary on different parts of the phylogeny (external branches versus 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 turn-over 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 Table S.1). 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 (Fig. 1). A large number of universally present non-duplicated genes from each clade were extracted and examined to obtain a robust phylogenetic tree. The number 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; 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 alpha in a Γ distribution was estimated using the PUZZLE program (Strimmer and V.H. 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 paralogues 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 \( E \text{-value} \leq 10^{-20} \) and match length \( \geq 85\% \) in a BLASTp search (Altschul et al., 1997). Gene families retained for further analysis were required to have more than 100 amino acids in at least one gene member, since similarity searches using BLAST have low power to detect homologues in short sequences (Altschul et al., 1997). Genes clustered in gene families were considered as gene presence or 'p'. Second, we further analyze 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 members, 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 \( E \text{-value} \leq 10^{-20} \) and match length \( \geq 85\% \). (2) gene truncation ('f') was inferred, if the BLAST hit has an \( E \text{-value} \leq 10^{-20} \) but match length <85%. (3) hits that have an \( E \text{-value} > 10^{-20} \) were considered as gene absence ('a'). The observed patterns of gene presence/absence/truncation are shown in Tables S.2-S.5. To access the robustness of the analysis, a different
The criterion of $E$-value $\leq 10^{-10}$ and match length $\geq 70\%$ in both BLASTP and TBLASTN searches was used in gene family identification (Tables S.6-S.9 and Fig. 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 5, which also makes the computation less demanding.

The mathematical model

The transitions among $p$’s, $f$’s, and $a$’s are defined by a $3 \times 3$ instantaneous rate matrix $Q$ with stationary probabilities $(\pi_a, \pi_f, \pi_p)$. Here $\pi_a + \pi_f + \pi_p = 1$, and the matrix $Q$ is reversible:

$$Q = \begin{pmatrix}
  a & f & p \\
  a & -\pi_f \alpha - \pi_p \beta & \pi_f \alpha & \pi_p \beta \\
  f & \pi_a \alpha & -\pi_a \alpha - \pi_p \gamma & \pi_p \gamma \\
  p & \pi_a \beta & \pi_f \gamma & -\pi_a \beta - \pi_f \gamma
\end{pmatrix}$$

where $\alpha$, $\beta$, and $\gamma$ 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

$$Q = \begin{pmatrix}
  a & p \\
  a & -\pi_p & \pi_p \\
  p & \pi_a & -\pi_a
\end{pmatrix}$$

which has been used in previous studies (COHEN et al., 2008; COHEN and PUPKO, 2010; SPENCER and SANGAR-ALINGAM, 2009). When $\pi_a = \pi_p = 0.5$, the matrix entries $Q(a,p)$ and $Q(p,a)$ are equal to 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; COHEN et al., 2008). This model is labelled $M_{00}$ in Table 3. To reduce the number of parameters to be optimized, the $\alpha$ parameter in the $Q_{3 \times 3}$ matrix was
It is a standard practice to allow only calibrated rate matrices, i.e., $Q$ satisfies
\[ -\sum_a \pi_a Q(a, a) = 1, \] (3)
so that a rate parameter (shown as $\mu$ below) is the average number of transition events per gene family per evolutionary time. The transition probability matrix is
\[ P = e^{Q\mu t}, \] (4)
where $t$ is the branch length based on nucleotide sequences, $\mu$ 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).
\[ L'_i = \frac{L_j}{1 - L'_j}, \] (5)
where $L'_j$ 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, summation is taken over all unobservable patterns to get $L'_i$.
\[ L'_i = \sum_{j \in U} L_j, \] (6)
where $U$ is the set of unobservable patterns (Table S.10), and $L_j$ is the likelihood of the $j$th unobservable pattern. Here $L'_i$ has the same value for all $i$.

Rate variation was also considered in a similar manner as nucleotide rate heterogeneity in phylogeny reconstruction (Yang, 1994; Felsenstein, 2001). A discrete $\Gamma$ 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 $\nu$ for that gene family weighted by the category probabilities $p_{\nu}$.
\[ L^i = \sum_{\nu=1}^M p_{\nu} L^i_{\nu}, \] (7)
where $\sum_{\nu=1}^M p_{\nu} = 1$. After incorporating a discrete $\Gamma$ model as done in (Spencer and Sangaralingam, 2009), the likelihood of observing the pattern of gene family $i$ will be
\[ L'_i = \frac{L^i}{1 - L'_i} = \frac{\sum_{\nu=1}^M p_{\nu} L^i_{\nu}}{1 - \sum_{\nu=1}^M p_{\nu} L^i_{\nu}}. \] (8)

First, parameters were estimated by assuming $\pi_a$, $\pi_f$, $\pi_p$ 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 $\pi_a$ and $\pi_p$ are the frequencies of each character-state in the data, and another model, in which $\pi_a$ and $\pi_p$ are 0.5, was introduced (let’s call it model
M_{00}) since it has been used in previous studies (Hao and Golding, 2006, 2008b; Cohen et al., 2008). Then $\pi_a$, $\pi_f$, $\pi_p$ were treated as parameters to be optimized, hence called model $M_{0+\pi}$. Finally, a discrete $\Gamma$ 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; Development Core Team, 2008). To further reduce the chance of being trapped in a 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 Fig. 1). The results reveal that closely related clades have higher rates of gene insertions/deletions ($\mu$), than distantly related clades (Fig. 2 and Table 2). This trend holds throughout all the four groups, and the use of different cutoff thresholds on identifying gene families yields remarkably similar results (Fig. 2 and Table S.11). 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 (Fig. 2), the stationary probability $\pi_a$ 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.

When the clade is more diverse, more gene families that were once present in the ancestral genome are lost from some descendants. Fig. S.1 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 $\pi_a$ appears to be greater than the observed frequency of ‘$a$’ in each clade (Fig. S.2). 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 S.10), but have been taken into account in the maximum likelihood estimation (equation 5). The rate variation parameter alpha (shown as $\alpha_{\Gamma}$ in Table 2) has a positive association with tree-length in the *Pseudomonas* and Bacillaceae groups (Fig. 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 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 has been 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 \( \alpha_\Gamma \) in the remaining 9 clades (Fig. S.3).

In the instantaneous rate matrix, \( \alpha, \beta, \) and \( \gamma \) 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 Fig. 3. Here, \( \alpha \) was fixed to be 1, \( \beta \) and \( \gamma \) were estimated under the MR+π model. The trend seems to be that the \( \beta \) and \( \gamma \) values increase as the clades become more diverse. There are three exceptions (in two clades, E1 and C3) to this trend. 1), the \( \beta \) and \( \gamma \) 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 (Fig.S.1), which is very likely associated with the process of genome reduction and gene pseudogenization in the *Shigella 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 \( \beta \) value in the C3 clade is smaller than the \( \beta \) in the C2 clade. Genome size was found to vary greatly in both clades, e.g., from 2.9 Mb to 6.0Mb in the C2 clade and from 2.5 Mb 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 \( \beta \) 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 \( \beta \) and \( \gamma \) 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 (Fig.S.4). The \( \beta \) 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 (\( \beta = 0.315 \) in C1) to C2, and to C3. Such a trend has been observed in *Pseudomonas* and *Bacillaceae* (Fig. 3). Furthermore, we computed the product of the scaled instantaneous rate matrix \( Q \) and the rate parameter \( \mu \) (Table S.12), which presents the instantaneous rates for all possible transitions. There is a clear trend that the instantaneous rates for all parameters to 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’→’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’→’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’→’p’), rather than absent (‘f’→’a’), all 12 clades showed a lower $\mu$ under the $M_0$ model. Under the $M_0+\pi$ model, 5 clades (B1, B1, E1, E2, E3) showed a lower $\mu$, while the remaining 7 clades showed a higher $\mu$. 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 $\pi_a$ 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 Fig. 4). Table 4 shows that nine clades have significant improvement when rates on external branches and internal branches were distinguished. All the nine clades show higher rates of gene insertions/deletions on external branches over on internal branches. Similar results have been 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 Fig. 2). Since most of the dynamics of gene insertions/deletions occurs 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 insertion/deletion was found between external branches and internal branches in such diverse clades. In fact, substantially different results have been observed on gene gains/losses by applying similar parsimony methods on distantly related species (MIRKIN et al., 2003) versus on closely related species (HAO and GOLDING, 2004).

**Discussion**

Here, we are primarily interested in modeling gene insertions/deletion 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 homologues might still carry out some function (OGATA et al., 2001).
To further assess 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 Fig. S.5 for details). Genome pairs Cph-Cth in C3 and Sfl-Sdy in E1 show the highest level of length variation (Fig. S.5). Then we plotted the DNA distance and Ka/Ks ratio of these homologs (Fig. S.6). All examined genome pairs have a medium DNA distance < 1.0 and a medium Ks/Ks ratio < 0.2 (Fig. S.6). 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 (Fig. S.7) and the query sequences are shorter (Fig. S.8). 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 (Fig. S.6) but a significantly high level of length variation (Fig. S.5). The Ks/Ks 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 (Fig. S.9). 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, (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 have used the longest, medium, and shortest sequences from each gene family as query sequences (see Fig. S.9 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, medium, or shortest sequences as query sequences and using a smaller word size in the search led to remarkably similar results in our examined genomes (Fig. S.9).

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 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 un-sampled 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; Cohen et al., 2008; Hao and Golding, 2008b), we initially assumed a constant rate of gene insertions/deletions on each phylogeny. This simplifying assumption is not realistic. Since the number of gene insertions/deletions is proportional to \( \mu t \), here \( t \) is the branch length, when the rate of gene insertions/deletions \( \mu \) is constant, the number of gene insertions/deletions would be proportional to the corresponding branch length. The assumption of a constant rate \( \mu \) 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 (Fig. 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 \( \mu \) and the degree of divergence in the clade (Fig. 2). The same trend has been found in a previous study on groups with different sets of genomes (Hao and Golding, 2008b). Both studies have shown 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 Marr 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 gene present ‘p’, fragment ‘f’ and absent ‘a’. This is not likely realistic. For instance, the transition from ‘p’ (present) to ‘f’ (fragment) 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 homologue. 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; Lefebvre and Stanhope, 2007), and recombination could involve long stretches of sequences (Didelot et al., 2007; Shepard 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 present ‘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. Thus, the instantaneous rate parameters for the transitions from ‘f’ to ‘p’ and from ‘p’ to ‘f’ could be distinguished as γ(f→p) and γ(p→f) respectively. It is not clear how different the estimated rates would be under the assumptions of rate reversibility versus rate irreversibility. Future studies by incorporating an irreversible rate matrix 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 (Tables S.2-S.9). We suspect that the asymmetry between ‘f’ and ‘p’ 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 Fig. S.4). 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) have presented a method to reconstruct genomic flux based on 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 based on the presence or 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 gene family as a unit, do 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 insertions/deletions events. Adding an intermediate state and considering insertions/deletions in the unit of genes, our method has a 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 become decreasingly small 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 (Fig. 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 turn-over. 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 have 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 (Fig. 5). Classifying truncated genes as absent does not always yield smaller rates ($\mu$) 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 $M_{\theta+\pi}$ model in the study, the rate parameter $\mu$ 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 (Fig. 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), then to HKY (Hasegawa et al., 1985)).

The expected equilibrium frequencies of the three character states ($\pi_a$, $\pi_f$, $\pi_p$) 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 of 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 genome than in the empirical data. In fact, the stationary probability $\pi_a$ is always higher than the observed frequency of ‘a’ (Fig. S.2). One should not misinterpret it as any systematic bias that might favour more gene losses. In the results, the stationary probability $\pi_a$ is positively associated with the tree-length of each clade (Fig. 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 $\pi_a$ 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 $\pi_a$ became smaller in every clade compared than when truncated genes were considered. While if truncated genes were forced to be classified as absent, the estimate of $\pi_a$ became larger in every clade compared than when truncated genes were considered (Fig. 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; Snell et al., 1999; Tekia 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 ($p, a$) has increased the $-\ln L$ values (Tables 2 and 3). In Fig. S.10, 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.
Conclusions

The results from the improved model reveal fast rates of gene insertions/deletions/truncations on recent branches. This holds true when both comparing different rates 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.

Authors contributions

W.H. and G.B.G. designed the study. W.H. carried out all analyses. W.H. and G.B.G. wrote the manuscript.

Acknowledgements

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. We would also like to thank the anonymous reviewers for their helpful comments on previous versions of this manuscript.

References


<table>
<thead>
<tr>
<th>Clade</th>
<th>Tree-length</th>
<th>Species/strain name</th>
<th>Abbreviation</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>0.366</td>
<td>Bacillus anthracis</td>
<td>Ba</td>
<td>NC_003997</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bacillus cereus ATCC10987</td>
<td>Bc1</td>
<td>NC_003909</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bacillus cereus ATCC14579</td>
<td>Bc2</td>
<td>NC_004722</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bacillus weihenstephaniensis</td>
<td>Bw</td>
<td>NC_001084</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bacillus cereus subsp. cyotoxis</td>
<td>Bc3</td>
<td>NC_009674</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bacillus amyloliquifaciens</td>
<td>Bani</td>
<td>NC_009725</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bacillus subtilis</td>
<td>Bs</td>
<td>NC_009694</td>
</tr>
<tr>
<td></td>
<td>1.981</td>
<td>Bacillus licheniformis ATCC 14580</td>
<td>Bl1</td>
<td>NC_008322</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bacillus pseudus</td>
<td>Bp</td>
<td>NC_009848</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Geobacillus kaustophilus</td>
<td>Gk</td>
<td>NC_008510</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bacillus halodurans</td>
<td>Bh</td>
<td>NC_002570</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bacillus clausii</td>
<td>Bcl</td>
<td>NC_006582</td>
</tr>
<tr>
<td></td>
<td>3.274</td>
<td>Oceanobacillus iheyensis</td>
<td>Oi</td>
<td>NC_004193</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exiguobacterium sibiricum</td>
<td>Es</td>
<td>NC_010556</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Anoxybacillus flavithermus</td>
<td>Af</td>
<td>NC_011567</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium botulinum A str. Hall</td>
<td>Cbo1</td>
<td>NC_009698</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium botulinum A2 str. Kyoto</td>
<td>Cbo2</td>
<td>NC_012563</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium botulinum A3 str. Loch Mare</td>
<td>Cbo3</td>
<td>NC_010520</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium botulinum B4 str. 657</td>
<td>Cbo4</td>
<td>NC_012658</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium botulinum A3 str. Loch Mare</td>
<td>Cbo5</td>
<td>NC_010520</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium botulinum B str. Ekholm 178</td>
<td>Cbo6</td>
<td>NC_010674</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium butyricum NCTM 8052</td>
<td>Cbe</td>
<td>NC_009617</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium perfringens SM101</td>
<td>Cpe</td>
<td>NC_008262</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium acetobutylicum ATCC 824</td>
<td>Cac</td>
<td>NC_009300</td>
</tr>
<tr>
<td>C1</td>
<td>0.055</td>
<td>Clostridium novyi NT</td>
<td>Cno1</td>
<td>NC_008593</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium tetani E88</td>
<td>Cte</td>
<td>NC_004557</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium difficile 630</td>
<td>Cdf</td>
<td>NC_009089</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium phytotoxins JSdg</td>
<td>Cph</td>
<td>NC_008258</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clostridium thermocellum ATCC 27405</td>
<td>Cth</td>
<td>NC_009012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shigella boydii S8227</td>
<td>Sbo</td>
<td>NC_007613</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shigella sonnei Ss046</td>
<td>Sso</td>
<td>NC_007384</td>
</tr>
<tr>
<td></td>
<td>0.039</td>
<td>Escherichia coli E24377A</td>
<td>Eco1</td>
<td>NC_009901</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shigella flexneri 5 str. 8401</td>
<td>Sfl</td>
<td>NC_008258</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shigella dysenteriae Sd197</td>
<td>Sdy</td>
<td>NC_007906</td>
</tr>
<tr>
<td>E1</td>
<td></td>
<td>Escherichia coli 536</td>
<td>Eco2</td>
<td>NC_008253</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Escherichia coli ED1a</td>
<td>Eco3</td>
<td>NC_011745</td>
</tr>
<tr>
<td></td>
<td>0.044</td>
<td>Escherichia coli OPEC 01</td>
<td>Eco4</td>
<td>NC_008563</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Escherichia coli O127:H6 str. E2348/69</td>
<td>Eco5</td>
<td>NC_011601</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Escherichia coli IA19</td>
<td>Eco6</td>
<td>NC_011750</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Escherichia coli IA1</td>
<td>Eco7</td>
<td>NC_011741</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Escherichia coli IBS</td>
<td>Eco8</td>
<td>NC_009900</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Salmonella enterica subsp. arizonae serovar 62:a:-,23</td>
<td>Sen</td>
<td>NC_010667</td>
</tr>
<tr>
<td>E2</td>
<td>0.312</td>
<td>Pseudomonas putida F1</td>
<td>Pp1</td>
<td>NC_009512</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonas putida KT2440</td>
<td>Pp2</td>
<td>NC_002947</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonas putida GB-1</td>
<td>Pp3</td>
<td>NC_010322</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonas putida W619</td>
<td>Pp4</td>
<td>NC_010501</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonas entomophila L48</td>
<td>Pp5</td>
<td>NC_008027</td>
</tr>
<tr>
<td></td>
<td>0.312</td>
<td>Pseudomonas fluorescens PRL-1</td>
<td>Pfl1</td>
<td>NC_007492</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonas fluorescens SBW25</td>
<td>Pfl2</td>
<td>NC_012660</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonas fluorescens PRL-5</td>
<td>Pfl3</td>
<td>NC_004129</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonas syringae pv. phaseolicola 1448A</td>
<td>Pps1</td>
<td>NC_005733</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonas mendocina ymp</td>
<td>Pmm</td>
<td>NC_009439</td>
</tr>
<tr>
<td></td>
<td>0.885</td>
<td>Pseudomonas putida KT2440</td>
<td>Ppp1</td>
<td>NC_002947</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonas syringae pv. tomato subsp. DC3000</td>
<td>Pps2</td>
<td>NC_004578</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonas mendocina ymp</td>
<td>Pmp</td>
<td>NC_009439</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonas stutzeri A1501</td>
<td>Pst</td>
<td>NC_009414</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonas aeruginosa PA7</td>
<td>Pae</td>
<td>NC_009656</td>
</tr>
</tbody>
</table>

Table 1: Strain information from a variety of phylogenetic groups
Table 2: Maximum log-likelihood comparison of different evolutionary models

<table>
<thead>
<tr>
<th>Models and Parameters</th>
<th>Bacillaceae</th>
<th>Clostridium</th>
<th>Escherichia</th>
<th>Pseudomonas</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B1</td>
<td>B2</td>
<td>B3</td>
<td>C1</td>
</tr>
<tr>
<td></td>
<td>(5453)</td>
<td>(5614)</td>
<td>(6813)</td>
<td>(3546)</td>
</tr>
<tr>
<td>$M_0$</td>
<td>$\mu$ 2.705</td>
<td>0.515</td>
<td>0.404</td>
<td>8.315</td>
</tr>
<tr>
<td></td>
<td>$\beta$ 0.577</td>
<td>1.281</td>
<td>1.700</td>
<td>0.326</td>
</tr>
<tr>
<td></td>
<td>$\gamma$ 1.311</td>
<td>2.096</td>
<td>2.725</td>
<td>0.590</td>
</tr>
<tr>
<td></td>
<td>$\ln L$ -16551</td>
<td>-18432</td>
<td>-24799</td>
<td>-7424</td>
</tr>
<tr>
<td>$M_0+\pi$</td>
<td>$\mu$ 2.102</td>
<td>0.214</td>
<td>0.197</td>
<td>8.385</td>
</tr>
<tr>
<td></td>
<td>$\beta$ 0.440</td>
<td>0.845</td>
<td>1.314</td>
<td>0.315</td>
</tr>
<tr>
<td></td>
<td>$\gamma$ 1.900</td>
<td>3.862</td>
<td>4.489</td>
<td>0.728</td>
</tr>
<tr>
<td></td>
<td>$\pi_a$ 0.487</td>
<td>0.788</td>
<td>0.797</td>
<td>0.255</td>
</tr>
<tr>
<td></td>
<td>$\pi_f$ 0.041</td>
<td>0.036</td>
<td>0.035</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>$\ln L$ -16243</td>
<td>-17799</td>
<td>-24069</td>
<td>-7385</td>
</tr>
<tr>
<td>$M_0+\Gamma+\pi$</td>
<td>$\mu$ 3.811</td>
<td>0.264</td>
<td>0.285</td>
<td>8.446</td>
</tr>
<tr>
<td></td>
<td>$\alpha_f$ 0.330</td>
<td>0.736</td>
<td>0.811</td>
<td>1.839</td>
</tr>
<tr>
<td></td>
<td>$\ln L$ -15861</td>
<td>-17553</td>
<td>-23738</td>
<td>-7370</td>
</tr>
<tr>
<td>$\Delta AIC$†</td>
<td>-612</td>
<td>-1262</td>
<td>-1456</td>
<td>-74</td>
</tr>
</tbody>
</table>

$\Delta AIC$s 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, $\Delta AIC = 2(-\ln L + \Delta K)$. The model that best approximates the data is the one with smallest $AIC$.

†This is the shape parameter in a $\Gamma$ distribution, which is traditionally described as $\alpha$. The use of $\alpha_f$ is to distinguish it from the symbol $\alpha$ used in the instantaneous rate matrix.

‡This is the shape parameter in a $\Gamma$ distribution, which is traditionally described as $\alpha$. The use of $\alpha_f$ is to distinguish it from the symbol $\alpha$ used in the instantaneous rate matrix.

§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’’), in the two extreme scenarios, all truncated genes were entirely classified as absent (‘\( f’→’a’’ or present (‘\( f’→’p’’).

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Models and Parameters</th>
<th>Bacillaceae</th>
<th>Clostridium</th>
<th>Escherichia</th>
<th>Pseudomonas</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_{00} )</td>
<td>( \mu )</td>
<td>1.886</td>
<td>2.474</td>
<td>1.842</td>
<td>1.581</td>
</tr>
<tr>
<td></td>
<td>( \ln L )</td>
<td>-13682</td>
<td>-13970</td>
<td>-13681</td>
<td>-13680</td>
</tr>
<tr>
<td>( M_{0} ^{+\pi} )</td>
<td>( \mu )</td>
<td>1.490</td>
<td>2.045</td>
<td>1.681</td>
<td>1.426</td>
</tr>
<tr>
<td></td>
<td>( \ln L )</td>
<td>-12470</td>
<td>-12579</td>
<td>-12460</td>
<td>-11983</td>
</tr>
</tbody>
</table>

\( \pi_a \) and \( \pi_p \) are the frequencies of each character in the data. \( \Delta \ln L \) are shown as \( (M_{0} ^{+\pi}) \) vs. \( M_{0} \). Since \( 2 \Delta \ln L \approx \chi^2 \), \( \Delta \ln L > 1.97 \) (\( df = 1 \)) are considered significant.

As described in the Method section, 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 \( \pi_a \) and \( \pi_p \) are the frequencies of each character in the data. \( \Delta \ln L \) are shown as \( (M_{0} ^{+\pi}) \) vs. \( M_{0} \). Since \( 2 \Delta \ln L \approx \chi^2 \), \( \Delta \ln L > 1.97 \) (\( df = 1 \)) are considered significant.
Table 4: Rate comparison on different branches under the $M_{0}+\pi$ model

<table>
<thead>
<tr>
<th>Models</th>
<th>Bacillaceae</th>
<th>Clostridium</th>
<th>Escherichia</th>
<th>Pseudomonas</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1=$ &amp; B1</td>
<td>B2</td>
<td>B3</td>
<td>C1</td>
<td>C2</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>-16243.9</td>
<td>-17799.0</td>
<td>-24069.1</td>
<td>-7384.6</td>
</tr>
<tr>
<td>$\ln L$</td>
<td>-15879.1</td>
<td>-17798.9</td>
<td>-24067.8</td>
<td>-7233.0</td>
</tr>
<tr>
<td>$\Delta \ln L$</td>
<td>364.8</td>
<td>0.1</td>
<td>1.3</td>
<td>151.6</td>
</tr>
</tbody>
</table>

$\Delta \ln L \geq 1.94$ ($df = 1$) are considered significant. As a result, B2, B3, and C3 are not significant.
Figure 1: Phylogenies with varied levels of divergence. Clade names and strain abbreviations are as in Table 1. Horizontal scale bar indicates genome size.
Figure 2: Association between estimated parameters and phylogenetic divergence in each group. Three maximum likelihood estimates (MLEs), $\mu$, $\pi$, $\alpha$, are estimated under the $M_0+\pi$ model using different cutoff thresholds in BLASTP/TBLASTN searches. Filled symbols represent MLEs using a criterion of $E$-value $\leq 10^{-20}$ and match length $\geq 85\%$, while open symbols represent MLEs using a criterion of $E$-value $\leq 10^{-10}$ and match length $\geq 70\%$. The four groups are sorted as 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.
Figure 3: Estimated parameters of the instantaneous rate matrix in each clade. As described in the methods section, $\alpha$ was fixed to be 1, $\beta$ and $\gamma$ were estimated under the $M_{0+\pi}$ model.
Figure 4: Rate parameters estimated on a 5-taxon phylogeny. Rates on external branches are $\mu_1$, and rates on internal branches are $\mu_2$. 
Figure 5: Comparison of MLEs with (filled 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).
### Supplementary Information

Table S.1: List of outgroup species for each clade

<table>
<thead>
<tr>
<th>Clade</th>
<th>Outgroup species</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td><em>Bacillus halodurans</em></td>
<td>NC_002570</td>
</tr>
<tr>
<td>B2</td>
<td><em>Bacillus halodurans</em></td>
<td>NC_002570</td>
</tr>
<tr>
<td>B3</td>
<td><em>Lysinibacillus sphaericus</em></td>
<td>NC_010382</td>
</tr>
<tr>
<td>C1</td>
<td><em>Clostridium tetani</em> E88</td>
<td>NC_004557</td>
</tr>
<tr>
<td></td>
<td><em>Clostridium kluyveri</em> DSM 555</td>
<td>NC_009706</td>
</tr>
<tr>
<td>C2</td>
<td><em>Clostridium cellulolyticum</em> H10</td>
<td>NC_011898</td>
</tr>
<tr>
<td>C3</td>
<td><em>Moorella thermoacetica</em> ATCC 39073</td>
<td>NC_007644</td>
</tr>
<tr>
<td>E1</td>
<td><em>Escherichia fergusonii</em> ATCC 35469</td>
<td>NC_011740</td>
</tr>
<tr>
<td>E2</td>
<td><em>Escherichia fergusonii</em> ATCC 35469</td>
<td>NC_011740</td>
</tr>
<tr>
<td>E3</td>
<td><em>Yersinia pestis</em> Antiqua</td>
<td>NC_008150</td>
</tr>
<tr>
<td>P1</td>
<td><em>Pseudomonas mendocina</em> ymp</td>
<td>NC_009439</td>
</tr>
<tr>
<td>P2</td>
<td><em>Pseudomonas aeruginosa</em> PA7</td>
<td>NC_009656</td>
</tr>
<tr>
<td>P3</td>
<td><em>Azotobacter vinelandii</em> DJ</td>
<td>NC_012560</td>
</tr>
</tbody>
</table>
Table S.2: Phylogenetic patterns in the Bacillaceae group (using a cutoff threshold of $E$-value $\leq 10^{-20}$ and match length $\geq 85\%$)

<table>
<thead>
<tr>
<th>Number of genes</th>
<th>B1 Ba Bc1 Bc2 Bw Bc3</th>
<th>Number of genes</th>
<th>B2 Bam Bs B2 Bp Gk</th>
<th>Number of genes</th>
<th>B3 Bh Bcl Oi Es Af</th>
</tr>
</thead>
<tbody>
<tr>
<td>567</td>
<td>p p p p a</td>
<td>725</td>
<td>a a a a a p</td>
<td>720</td>
<td>a a p a a a</td>
</tr>
<tr>
<td>238</td>
<td>a a a a p</td>
<td>509</td>
<td>p p p p a a</td>
<td>707</td>
<td>a p a a a a</td>
</tr>
<tr>
<td>225</td>
<td>a a p a a</td>
<td>412</td>
<td>a a p a a a</td>
<td>602</td>
<td>p a a a a a</td>
</tr>
<tr>
<td>216</td>
<td>a p a a a</td>
<td>313</td>
<td>a a a a p a</td>
<td>570</td>
<td>a a a a a p</td>
</tr>
<tr>
<td>194</td>
<td>a a a p a</td>
<td>282</td>
<td>a p a a a a</td>
<td>471</td>
<td>a a a a p a</td>
</tr>
<tr>
<td>153</td>
<td>p a a a a</td>
<td>160</td>
<td>p a a a a a</td>
<td>327</td>
<td>p p a a a a</td>
</tr>
<tr>
<td>87</td>
<td>p p p p f</td>
<td>159</td>
<td>p a a a a a</td>
<td>238</td>
<td>p p a p p p</td>
</tr>
<tr>
<td>71</td>
<td>p p p a a</td>
<td>152</td>
<td>p p p a a a</td>
<td>199</td>
<td>p p a a a p</td>
</tr>
<tr>
<td>67</td>
<td>p a p p a</td>
<td>101</td>
<td>p p p p f</td>
<td>132</td>
<td>a p a a a p</td>
</tr>
<tr>
<td>66</td>
<td>p a a p a</td>
<td>90</td>
<td>a p p a a a</td>
<td>125</td>
<td>p p p a p p</td>
</tr>
<tr>
<td>52</td>
<td>p p a a a</td>
<td>84</td>
<td>a a p p a a</td>
<td>91</td>
<td>p a a p a a p</td>
</tr>
<tr>
<td>51</td>
<td>a p p p p</td>
<td>66</td>
<td>p p a p a a</td>
<td>85</td>
<td>p a a a p a</td>
</tr>
<tr>
<td>50</td>
<td>p a p a a</td>
<td>59</td>
<td>a p p p a a</td>
<td>72</td>
<td>a a p p a a</td>
</tr>
<tr>
<td>50</td>
<td>a a p p a</td>
<td>52</td>
<td>f p f a p p</td>
<td>72</td>
<td>a a a p a a</td>
</tr>
<tr>
<td>42</td>
<td>f p p p p</td>
<td>48</td>
<td>p p p p a a</td>
<td>70</td>
<td>p p a a a a</td>
</tr>
<tr>
<td>42</td>
<td>p p a p a</td>
<td>40</td>
<td>a p a a p a</td>
<td>49</td>
<td>a a a a p p</td>
</tr>
<tr>
<td>40</td>
<td>a a p a p</td>
<td>39</td>
<td>a a p a p a</td>
<td>45</td>
<td>a p p a a p</td>
</tr>
<tr>
<td>37</td>
<td>p p f p p</td>
<td>33</td>
<td>p p a a p a</td>
<td>42</td>
<td>p p p p a a</td>
</tr>
<tr>
<td>37</td>
<td>p p a p p</td>
<td>26</td>
<td>f p p p p p</td>
<td>42</td>
<td>p p p a a a</td>
</tr>
<tr>
<td>760</td>
<td>Other patterns</td>
<td>743</td>
<td>Other patterns</td>
<td>1162</td>
<td>Other patterns</td>
</tr>
</tbody>
</table>
Table S.3: Phylogenetic patterns in the *Clostridium* group (using a cutoff threshold of \( E\)-value \(\leq 10^{-20}\) and match length \(\geq 85\%\))

<table>
<thead>
<tr>
<th>Number of genes</th>
<th>Cbo₁</th>
<th>Cbo₂</th>
<th>Cbo₃</th>
<th>Cbo₄</th>
<th>Number of genes</th>
<th>Cbo₅</th>
<th>Cbo₆</th>
<th>Cbo₇</th>
<th>Cbe</th>
<th>Cpe</th>
<th>Cac</th>
<th>Number of genes</th>
<th>Cno</th>
<th>Cte</th>
<th>Cdi</th>
<th>Cph</th>
<th>Cth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2345</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>1110</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>1290⁺</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>114</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>977</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>994</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>892</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>949</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>436</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>712⁺</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>327</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>479</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>84</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>309</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>326</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>266</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>220</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>259</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>175</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>236</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>122</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>196</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>114</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>178</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>103</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>f</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>104</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>100</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>101</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>94</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>64</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>77</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>p</td>
<td>54</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>67</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>f</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>49</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>66</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>46</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td>58</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>41</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>57</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>p</td>
<td>36</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>55</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>a</td>
<td>p</td>
<td>f</td>
<td>a</td>
<td>32</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>54</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>308</td>
<td>Other patterns</td>
<td>813</td>
<td>Other patterns</td>
<td>1533</td>
<td>Other patterns</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>114</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Unlike in other clades, the pattern of the genes present in all taxa is *not* the most gene-family-rich pattern in C3.
Table S.4: Phylogenetic patterns in the \textit{Escherichia} group (using a cutoff threshold of $E$-value $\leq 10^{-20}$ and match length $\geq 85\%$)

<table>
<thead>
<tr>
<th>Number of genes</th>
<th>Sbo</th>
<th>Sso</th>
<th>Eco$_1$</th>
<th>Sfl</th>
<th>Sdy</th>
<th>Number of genes</th>
<th>Eco$_2$</th>
<th>Eco$_3$</th>
<th>Eco$_4$</th>
<th>Eco$_5$</th>
<th>Eco$_6$</th>
<th>Number of genes</th>
<th>Eco$_7$</th>
<th>Eco$_8$</th>
<th>Efe</th>
<th>Sen</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>a a p a a a</td>
<td>242</td>
<td>a a a a p a a</td>
<td>269</td>
<td>a a a p a a</td>
<td>238</td>
<td>p p p p p a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>199</td>
<td>p p p p a a a</td>
<td>205</td>
<td>p a a a a a</td>
<td>216</td>
<td>a a p a a a</td>
<td>130</td>
<td>p p p p a a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>124</td>
<td>p p p p f</td>
<td>189</td>
<td>a a a a p p a</td>
<td>95</td>
<td>p a a a a a</td>
<td>91</td>
<td>p p p p f</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>a p p a a a</td>
<td>129</td>
<td>a a p a a a</td>
<td>54</td>
<td>p p p a p a</td>
<td>89</td>
<td>p p p a a a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>a a a p a a</td>
<td>101</td>
<td>p a a a a a</td>
<td>52</td>
<td>a p p a a a</td>
<td>81</td>
<td>p p p a a a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>p a a a a a</td>
<td>71</td>
<td>p p p a a a</td>
<td>60</td>
<td>p p p a a a</td>
<td>44</td>
<td>a a a p p p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>a a a a a p</td>
<td>60</td>
<td>p p p a a a</td>
<td>66</td>
<td>a a a a a a</td>
<td>69</td>
<td>a a a a a a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>f p p p p p</td>
<td>64</td>
<td>p p p a a a</td>
<td>39</td>
<td>p a a p p p</td>
<td>62</td>
<td>a a a a p a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>p p p p p f</td>
<td>44</td>
<td>a a a p p p</td>
<td>39</td>
<td>p a a p p p</td>
<td>60</td>
<td>a a a a p a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>a p p p a a</td>
<td>39</td>
<td>p a a p p p</td>
<td>32</td>
<td>f f p p p p</td>
<td>60</td>
<td>a a a p a a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>p p p a a a</td>
<td>39</td>
<td>a p p p p p</td>
<td>32</td>
<td>f f p p p p</td>
<td>39</td>
<td>a a a a p a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>a p p p p p</td>
<td>36</td>
<td>a a p a a a</td>
<td>32</td>
<td>f f p p p p</td>
<td>39</td>
<td>a a a a p a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>p f p p p p</td>
<td>32</td>
<td>f f p p p p</td>
<td>32</td>
<td>f f p p p p</td>
<td>39</td>
<td>a a a a p a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>a p a a a a</td>
<td>32</td>
<td>f f p p p p</td>
<td>32</td>
<td>f f p p p p</td>
<td>39</td>
<td>a a a a p a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>p p p a a a</td>
<td>31</td>
<td>a a p a a a</td>
<td>32</td>
<td>a a a p p a</td>
<td>35</td>
<td>p p a a a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>a a p p a a</td>
<td>30</td>
<td>a a a p p a</td>
<td>30</td>
<td>a a a p p a</td>
<td>34</td>
<td>a a a a p a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>f p p p p f</td>
<td>29</td>
<td>a a a p a a</td>
<td>29</td>
<td>a a a p a a</td>
<td>28</td>
<td>p a a p p a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>p p a a p a</td>
<td>28</td>
<td>p a a a p p</td>
<td>28</td>
<td>p a a a p p</td>
<td>21</td>
<td>f f p p f f</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>801</td>
<td>Other patterns</td>
<td>562</td>
<td>Other patterns</td>
<td>560</td>
<td>Other patterns</td>
<td>560</td>
<td>Other patterns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The table above shows the phylogenetic patterns in the \textit{Escherichia} group, with columns indicating the number of genes for each species and the number of patterns matching those criteria.
Table S.5: Phylogenetic patterns in the *Pseudomonas* group (using a cutoff threshold of \( E\)-value \( \leq 10^{-20} \) and match length \( \geq 85\% \))

<table>
<thead>
<tr>
<th>Number of genes</th>
<th>Ppu1</th>
<th>Ppu2</th>
<th>Ppu3</th>
<th>Ppu4</th>
<th>Pen</th>
<th>Number of genes</th>
<th>Pfl1</th>
<th>Pfl2</th>
<th>Pfl3</th>
<th>Psy1</th>
<th>Pme</th>
<th>Number of genes</th>
<th>Pme</th>
<th>Pst</th>
<th>Pae</th>
</tr>
</thead>
<tbody>
<tr>
<td>526</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>670</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>872</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>333</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>637</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>791</td>
<td>a</td>
<td>p</td>
<td>a</td>
</tr>
<tr>
<td>317</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>559</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>658</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>243</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>511</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>436</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>208</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>487</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>413</td>
<td>a</td>
<td>a</td>
<td>p</td>
</tr>
<tr>
<td>178</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>485</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>227</td>
<td>p</td>
<td>p</td>
<td>a</td>
</tr>
<tr>
<td>105</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>284</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>188</td>
<td>p</td>
<td>p</td>
<td>a</td>
</tr>
<tr>
<td>92</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>217</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>178</td>
<td>p</td>
<td>a</td>
<td>p</td>
</tr>
<tr>
<td>78</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>175</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>172</td>
<td>a</td>
<td>a</td>
<td>p</td>
</tr>
<tr>
<td>71</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>163</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>151</td>
<td>p</td>
<td>p</td>
<td>a</td>
</tr>
<tr>
<td>65</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>88</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>140</td>
<td>a</td>
<td>p</td>
<td>a</td>
</tr>
<tr>
<td>58</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>87</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>136</td>
<td>a</td>
<td>a</td>
<td>p</td>
</tr>
<tr>
<td>56</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>81</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>129</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>46</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>78</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>129</td>
<td>a</td>
<td>a</td>
<td>p</td>
</tr>
<tr>
<td>44</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>66</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>109</td>
<td>a</td>
<td>a</td>
<td>p</td>
</tr>
<tr>
<td>39</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td>p</td>
<td>63</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>p</td>
<td>88</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>33</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>p</td>
<td>59</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>80</td>
<td>a</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>30</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>57</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td>p</td>
<td>66</td>
<td>p</td>
<td>p</td>
<td>a</td>
</tr>
<tr>
<td>30</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>55</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>66</td>
<td>a</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>599</td>
<td>Other patterns</td>
<td>1433</td>
<td>Other patterns</td>
<td>1786</td>
<td>Other patterns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table S.6: Phylogenetic patterns in the Bacillaceae group (using a cutoff threshold of $E$-value $\leq 10^{-10}$ and match length $\geq 70\%$)

<table>
<thead>
<tr>
<th>Number of genes</th>
<th>Ba</th>
<th>Bc\textsubscript{1}</th>
<th>Bc\textsubscript{2}</th>
<th>Bw</th>
<th>Bc\textsubscript{3}</th>
<th>Number of genes</th>
<th>Bam</th>
<th>Bs</th>
<th>Bl</th>
<th>Bp</th>
<th>Gk</th>
<th>Number of genes</th>
<th>Bh</th>
<th>Bcl</th>
<th>Oi</th>
<th>Es</th>
<th>Af</th>
</tr>
</thead>
<tbody>
<tr>
<td>428</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>555</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>521</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
</tr>
<tr>
<td>187</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>372</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>467</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
</tr>
<tr>
<td>184</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>301</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>426</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
</tr>
<tr>
<td>172</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>232</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>403</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
</tr>
<tr>
<td>148</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>210</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>327</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
</tr>
<tr>
<td>116</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>116</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>242</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
</tr>
<tr>
<td>77</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$f$</td>
<td>106</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>200</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>$p$</td>
</tr>
<tr>
<td>59</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>99</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>161</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
</tr>
<tr>
<td>53</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>67</td>
<td>$a$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>108</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
</tr>
<tr>
<td>50</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>60</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>103</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
</tr>
<tr>
<td>41</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>59</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$f$</td>
<td>72</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
</tr>
<tr>
<td>37</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>57</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>69</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
</tr>
<tr>
<td>36</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>54</td>
<td>$a$</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>62</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
</tr>
<tr>
<td>36</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>51</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>61</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
</tr>
<tr>
<td>34</td>
<td>$a$</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>35</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>59</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
</tr>
<tr>
<td>34</td>
<td>$a$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>33</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>46</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
</tr>
<tr>
<td>27</td>
<td>$a$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>27</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>39</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
</tr>
<tr>
<td>26</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>26</td>
<td>$f$</td>
<td>$f$</td>
<td>$f$</td>
<td>$f$</td>
<td>$p$</td>
<td>36</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
</tr>
<tr>
<td>24</td>
<td>$f$</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>$p$</td>
<td>19</td>
<td>$p$</td>
<td>$a$</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>33</td>
<td>$p$</td>
<td>$p$</td>
<td>$a$</td>
<td>$a$</td>
<td>$a$</td>
</tr>
<tr>
<td>540</td>
<td>Other patterns</td>
<td>516</td>
<td>Other patterns</td>
<td>855</td>
<td>Other patterns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table S.7: Phylogenetic patterns in the *Clostridium* group (using a cutoff threshold of $E$-value $\leq 10^{-10}$ and match length $\geq 70\%$)

<table>
<thead>
<tr>
<th>Number of genes</th>
<th>Cbo₁</th>
<th>Cbo₂</th>
<th>Cbo₃</th>
<th>Cbo₄</th>
<th>Cbo₅</th>
<th>Number of genes</th>
<th>Cbo₆</th>
<th>Cbo₇</th>
<th>Cbe</th>
<th>Cph</th>
<th>Cac</th>
<th>Number of genes</th>
<th>Cno</th>
<th>Cte</th>
<th>Cdi</th>
<th>Cph</th>
<th>Cth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2032</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>1071</td>
<td>645</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>900</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
</tr>
<tr>
<td>95</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>629</td>
<td>629</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>741</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
</tr>
<tr>
<td>81</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>316</td>
<td>251</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>719</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>77</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>225</td>
<td>217</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>646</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>73</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>199</td>
<td>194</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>347</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>71</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>194</td>
<td>194</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>214</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>41</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>194</td>
<td>194</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>164</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
</tr>
<tr>
<td>38</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>194</td>
<td>194</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>134</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>37</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>194</td>
<td>194</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>128</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
</tr>
<tr>
<td>29</td>
<td>f</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>154</td>
<td>99</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>99</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>28</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>153</td>
<td>78</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>89</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>27</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>95</td>
<td>78</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>78</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>25</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>79</td>
<td>77</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>77</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>14</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>53</td>
<td>73</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>73</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
</tr>
<tr>
<td>12</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>p</td>
<td>52</td>
<td>61</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>61</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>12</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>47</td>
<td>57</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>57</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
</tr>
<tr>
<td>12</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>42</td>
<td>55</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>55</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>10</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>38</td>
<td>54</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>54</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>10</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>33</td>
<td>51</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>51</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>9</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>26</td>
<td>51</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>51</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>269</td>
<td>Other patterns</td>
<td>694</td>
<td>Other patterns</td>
<td>1308</td>
<td>Other patterns</td>
<td>Other patterns</td>
<td>1308</td>
<td>Other patterns</td>
<td>1308</td>
<td>Other patterns</td>
<td>1308</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Unlike in other clades, the pattern of the genes present in all taxa is *not* the most gene-family-rich pattern in C3.
Table S.8: Phylogenetic patterns in the *Escherichia* group (using a cutoff threshold of $E$-value $\leq 10^{-10}$ and match length $\geq 70\%$)

<table>
<thead>
<tr>
<th>Number of genes</th>
<th>Sbo</th>
<th>Sso</th>
<th>Eco1</th>
<th>Sfl</th>
<th>Sdy</th>
<th>Number of genes</th>
<th>Eco2</th>
<th>Eco3</th>
<th>Efe</th>
<th>Other patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>2311</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>2687</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>425 Other patterns</td>
</tr>
<tr>
<td>160</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>207</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>453 Other patterns</td>
</tr>
<tr>
<td>110</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>178</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>73</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>144</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>71</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>108</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>70</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>84</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>64</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>63</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>60</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>45</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>50</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>41</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>47</td>
<td>f</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>40</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>45</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>40</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>38</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>32</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>36</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>p</td>
<td>31</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>31</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>28</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>28</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>27</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>27</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>26</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>24</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>25</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>23</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>24</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>23</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>24</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>21</td>
<td>f</td>
<td>f</td>
<td>p</td>
<td>f</td>
<td>f</td>
<td>23</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>435 Other patterns</td>
</tr>
<tr>
<td>537</td>
<td>Other patterns</td>
<td>425 Other patterns</td>
<td>435 Other patterns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table S.9: Phylogenetic patterns in the *Pseudomonas* group (using a cutoff threshold of $E$-value $\leq 10^{-10}$ and match length $\geq 70\%$)

<table>
<thead>
<tr>
<th>Number of genes</th>
<th>Ppu1</th>
<th>Ppu2</th>
<th>Ppu3</th>
<th>Ppu4</th>
<th>Pen</th>
<th>Number of genes</th>
<th>Pfl1</th>
<th>Pfl2</th>
<th>Pfl3</th>
<th>Psy1</th>
<th>Pme</th>
<th>Number of genes</th>
<th>Pme</th>
<th>Psy2</th>
<th>Pme</th>
<th>Pst</th>
<th>Pae</th>
</tr>
</thead>
<tbody>
<tr>
<td>419</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>497</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>661</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
</tr>
<tr>
<td>258</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>497</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>594</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>240</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>414</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>476</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>188</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>391</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>335</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
</tr>
<tr>
<td>171</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>355</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>294</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>141</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>336</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>196</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>78</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>177</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>177</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>77</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>159</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>141</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>64</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>132</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>135</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>59</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>125</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>126</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>48</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>72</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>112</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>46</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>68</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>103</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>44</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>68</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>96</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
</tr>
<tr>
<td>40</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>54</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>88</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>38</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td>p</td>
<td>51</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>f</td>
<td>75</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>34</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>50</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>67</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>25</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>49</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>61</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>24</td>
<td>a</td>
<td>a</td>
<td>p</td>
<td>a</td>
<td>a</td>
<td>46</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td>p</td>
<td>60</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>p</td>
</tr>
<tr>
<td>22</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>46</td>
<td>p</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>p</td>
<td>53</td>
<td>a</td>
<td>p</td>
<td>p</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>475</td>
<td>Other patterns</td>
<td>1027</td>
<td>Other patterns</td>
<td>1224</td>
<td>Other patterns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table S.10: List of unobservable patterns

<table>
<thead>
<tr>
<th>taxon1</th>
<th>taxon2</th>
<th>taxon3</th>
<th>taxon4</th>
<th>taxon 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>f</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td>a</td>
<td>f</td>
<td>a</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td>a</td>
<td>f</td>
<td>f</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td>f</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td>f</td>
<td>a</td>
<td>f</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td>f</td>
<td>f</td>
<td>a</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td>f</td>
<td>f</td>
<td>f</td>
</tr>
<tr>
<td>a</td>
<td>f</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>a</td>
<td>f</td>
<td>a</td>
<td>a</td>
<td>f</td>
</tr>
<tr>
<td>a</td>
<td>f</td>
<td>a</td>
<td>f</td>
<td>f</td>
</tr>
<tr>
<td>a</td>
<td>f</td>
<td>a</td>
<td>f</td>
<td>a</td>
</tr>
<tr>
<td>a</td>
<td>f</td>
<td>f</td>
<td>f</td>
<td>f</td>
</tr>
<tr>
<td>f</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>f</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>f</td>
</tr>
<tr>
<td>f</td>
<td>a</td>
<td>a</td>
<td>f</td>
<td>f</td>
</tr>
<tr>
<td>f</td>
<td>a</td>
<td>f</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>f</td>
<td>a</td>
<td>f</td>
<td>a</td>
<td>f</td>
</tr>
<tr>
<td>f</td>
<td>a</td>
<td>f</td>
<td>f</td>
<td>a</td>
</tr>
<tr>
<td>f</td>
<td>a</td>
<td>f</td>
<td>f</td>
<td>f</td>
</tr>
<tr>
<td>f</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>a</td>
<td>a</td>
<td>f</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>a</td>
<td>f</td>
<td>a</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>a</td>
<td>f</td>
<td>f</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>f</td>
<td>a</td>
<td>f</td>
</tr>
</tbody>
</table>

*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 S.11: Maximum log-likelihood comparison of different evolutionary models (using a cutoff threshold of $E$-value $\leq 10^{-10}$ and match length $\geq 70\%$)

<table>
<thead>
<tr>
<th>Models and Parameters</th>
<th>Bacillaceae</th>
<th>Clostridium</th>
<th>Escherichia</th>
<th>Pseudomonas</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_0$</td>
<td>$\mu$ 2.329 0.458 0.370 8.167 0.989 0.464 17.109 11.901 2.442 1.834 1.067 0.893</td>
<td>$\beta$ 0.523 1.014 1.342 0.269 1.424 1.787 0.348 0.319 0.412 0.465 0.923 1.110</td>
<td>$\gamma$ 1.003 1.466 2.075 0.473 2.093 2.713 0.607 0.515 0.707 0.639 1.248 1.670</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ln$L$ -12742 -13776 -18923 -6180 -18440 -22231 -9627 -9887 -12168 -13750 -22419 -24172</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M_{0} + \pi$</td>
<td>$\mu$ 1.945 0.233 0.216 8.372 0.609 0.211 17.196 10.870 1.737 1.429 0.692 0.633</td>
<td>$\beta$ 0.455 0.816 1.113 0.267 1.346 1.303 0.328 0.298 0.342 0.441 0.813 0.944</td>
<td>$\gamma$ 1.381 2.578 3.274 0.555 3.495 3.756 0.738 0.713 1.036 0.936 1.819 2.124</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ln$L$ -12522 -13351 -18466 -6153 -18016 -21380 -9567 -9769 -11979 -13474 -21898 -23736</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M_{0} + \pi + \sigma_\mu$</td>
<td>$\mu$ 3.450 0.294 0.316 8.365 2.543 0.270 18.737 13.152 8.205 4.513 1.001 0.817</td>
<td>$\beta$ 0.318 0.616 0.702 1.919 0.283 1.200 0.862 0.464 0.147 0.149 0.470 0.705</td>
<td>$\sigma_\mu$ 0.446 0.729 0.740 0.234 0.691 0.812 0.246 0.372 0.480 0.538 0.659 0.639</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ln$L$ -12241 -13149 -18203 -6142 -17317 -21380 -9500 -9649 -11311 -13126 -21476 -23454</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta AIC$</td>
<td>-560 -402 -524 -20 -524 -352 -132 -234 -1334 -694 -842 -562</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\Delta AIC$'s are shown as $(M_{0} + \pi)$ vs. $M_0$ and $(M_{0} + \pi + \sigma_\mu)$ 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, $\Delta AIC = 2(-\Delta \ln L + \Delta K)$. The model that best approximates the data is the one with smallest $AIC$.

$\dagger$The number of gene families for each clade is shown in parenthesis underneath the clade name.
Table S.12: Product of the scaled instantaneous rate matrix $Q$ and the rate parameter $\mu$ (or $Q\mu$) under model $M_{0+\pi}$

<table>
<thead>
<tr>
<th></th>
<th>Bacillaceae</th>
<th>Clostridium</th>
<th>Escherichia</th>
<th>Pseudomonas</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_1$</td>
<td>$\begin{pmatrix} -1.655 &amp; 0.273 &amp; 1.382 \ 3.242 &amp; -0.212 &amp; 5.970 \ 1.426 &amp; 0.519 &amp; -1.945 \ -0.116 &amp; 0.023 &amp; 0.093 \end{pmatrix}$</td>
<td>$\begin{pmatrix} -11.171 &amp; 2.578 &amp; 8.593 \ 22.670 &amp; -69.009 &amp; 46.339 \ 3.060 &amp; 1.878 &amp; -4.937 \ -0.279 &amp; 0.030 &amp; 0.249 \end{pmatrix}$</td>
<td>$\begin{pmatrix} -23.810 &amp; 6.925 &amp; 17.785 \ 22.310 &amp; -67.966 &amp; 45.656 \ 7.474 &amp; 5.182 &amp; -12.656 \ -12.635 &amp; 2.646 &amp; 9.989 \end{pmatrix}$</td>
<td>$\begin{pmatrix} -1.023 &amp; 0.170 &amp; 0.853 \ 2.376 &amp; -4.305 &amp; 1.929 \ 1.271 &amp; 0.205 &amp; -1.476 \ -0.431 &amp; 0.072 &amp; 0.359 \end{pmatrix}$</td>
</tr>
<tr>
<td>$B_2$</td>
<td>$\begin{pmatrix} 0.496 &amp; -0.924 &amp; 0.428 \ 0.419 &amp; 0.087 &amp; -0.506 \ -0.109 &amp; 0.015 &amp; 0.094 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 0.643 &amp; -1.337 &amp; 0.694 \ 1.015 &amp; 0.132 &amp; -1.147 \ -0.075 &amp; 0.021 &amp; 0.054 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 24.940 &amp; -52.091 &amp; 27.151 \ 6.459 &amp; 1.863 &amp; -8.322 \ -1.484 &amp; 0.325 &amp; 1.159 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 1.060 &amp; 2.121 &amp; 1.061 \ 0.946 &amp; 0.190 &amp; -1.136 \ -0.398 &amp; 0.050 &amp; 0.349 \end{pmatrix}$</td>
</tr>
<tr>
<td>$B_3$</td>
<td>$\begin{pmatrix} 0.341 &amp; -0.664 &amp; 0.323 \ 0.448 &amp; 0.067 &amp; -0.515 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 0.503 &amp; -0.719 &amp; 0.216 \ 0.599 &amp; 0.100 &amp; -0.699 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 3.771 &amp; -7.413 &amp; 3.642 \ 1.331 &amp; 0.360 &amp; -1.691 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 0.598 &amp; 1.430 &amp; 0.832 \ 0.816 &amp; 1.611 &amp; -0.977 \end{pmatrix}$</td>
</tr>
</tbody>
</table>
Figure S.1: 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 S.2: Estimated stationary probabilities $\pi$ 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 ($\pi$) were estimated under the $M_0+\pi$ model.
Figure S.3: Association between tree length and the optimized $\alpha$ in a $\Gamma$ 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.

$R^2 = 0.69$, $P = 0.0054$

$y = 0.274x + 0.179$
Figure S.4: 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 $\Delta AIC$ was shown in Table 2.
Figure S.5: 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 S.6: 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 S.7: 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 S.8: 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 S.9: Distribution of match length of TBLASTN hits. Query sequences (shortest, medium, 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 S.10: 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 $M_0+\pi$ (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.