

# Inferring genomic flux in bacteria

Xavier Didelot,<sup>1,5</sup> Aaron Darling,<sup>2,4</sup> and Daniel Falush<sup>3</sup>

<sup>1</sup>Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom; <sup>2</sup>Institute for Molecular Bioscience, University of Queensland, St. Lucia QLD 4072, Australia; <sup>3</sup>Environmental Research Institute, Department of Microbiology, University College Cork, Cork, Ireland

Acquisition and loss of genetic material are essential forces in bacterial microevolution. They have been repeatedly linked with adaptation of lineages to new lifestyles, and in particular, pathogenicity. Comparative genomics has the potential to elucidate this genetic flux, but there are many methodological challenges involved in inferring evolutionary events from collections of genome sequences. Here we describe a model-based method for using whole-genome sequences to infer the patterns of genome content evolution. A fundamental property of our model is that it allows the rates at which genetic elements are gained or lost to vary in time and from one lineage to another. Our approach is purely sequence based, and does not rely on gene identification. We show how inference can be performed under our model and illustrate its use on three datasets from *Francisella tularensis*, *Streptococcus pyogenes*, and *Escherichia coli*. In all three examples, we found interesting variations in the rates of genetic material gain and loss, which strongly correlate with their lifestyle. The algorithms we describe are implemented in a computer software named GenoPlast.

[Supplemental material is available online at [www.genome.org](http://www.genome.org). The GenoPlast software is freely available from <http://go.warwick.ac.uk/genoplast/>.]

Bacteria adapt to new environmental niches by remodeling their genomes. Genome sequencing has revealed a prominent role for gene gain and loss in the processes of niche adaptation, specialization, host-switching, and other lifestyle changes. Diverse bacterial species exhibit such genetic flux, which plays a crucial role in bacterial evolution (Ochman et al. 2000; Wren 2000; Dobrindt and Hacker 2001).

Previous studies of genomic flux have used annotated genes as the units of gain and loss. In the standard inference protocol, genes annotated in sequenced genomes are first assigned to orthologous groups on the basis of sequence homology. Paralogs in multicopy gene families are either disambiguated or discarded, and genes exhibiting only partial homology are usually subjected to a conservation threshold to be considered orthologous (e.g., 70% of the amino acid length must be conserved). The resulting one-to-one mapping of orthologous genes is then subjected to a gene flux analysis. Gene gains and losses are typically presumed to be equally likely to occur in all lineages, which enables parsimonious mapping of gene gains/losses to branches of a phylogenetic tree relating the organisms under study. Finally such studies usually investigate the relationship between gain, loss, and ecological niche.

However, some molecular processes underlying genomic flux operate without regard to gene boundaries. Short segments within genes, such as protein domains, are often gained or lost (Spratt 1988; Riley and Labeledan 1997), and intergenic regulatory regions may also be subject to such pressures. Clusters of neighboring genes and operons may be gained or lost in a single event (Lawrence and Roth 1996). A complete evolutionary account would annotate individual events while also detecting variations in rate over time in particular lineages. The discovery of reductive genome evolution (Silva et al. 2001; Hershberg et al. 2007) has

clearly demonstrated that in many cases, the process of genomic gain and loss is asymmetric in some lineages. Parsimony criteria are known to be unreliable when branch lengths are unequal (Felsenstein 1978; Pol and Siddall 2001; Swofford et al. 2001), meaning that statistical modeling of unequal rates is necessary for accurate evolutionary inference.

In the present work, we introduce a new method to reconstruct genomic flux based on raw genomic sequence (without annotated coding sequences) that can also infer lineage-specific changes in the rates of gain and loss. Our method takes whole-genome multiple alignments as input, and outputs a mapping of changes in genomic content to branches of a phylogenetic tree, along with confidence estimates. The method utilizes a stochastic model of genomic evolution by gain and loss, incorporating a compound Poisson process model (Huelsenbeck et al. 2000) to allow the rates of gain and loss to vary in time and between lineages. Therefore, our model does not assume that evolution proceeds according to a constant molecular clock (Linz et al. 2007). The importance of modeling the changes in the rate of gene flux has been recognized before (Hao and Golding 2004, 2008; Marri et al. 2006), but our method is the first to be able to infer from the data where such changes may have happened instead of relying on the user's prior knowledge.

Our method processes a whole-genome multiple alignment to identify the parts that are present in all genomes (the core genome) and the parts present in some, but not all of them (the dispensable genome). The core genome is used to robustly infer a phylogenetic tree. Since the parts in the dispensable genome are not found in all genomes, they must have been gained or lost at least once along the branches of the phylogeny. In order to model the overall rate of genetic material being gained and lost, the dispensable genome is broken up into small "features" of constant size. We encode the presence or absence of these features in a particular genome as a binary character, and model the evolution of these binary characters along the phylogenetic tree. Thus, the rates of gain  $r^+$  and loss  $r^-$  incorporated in our model reflect the total number of nucleotides gained and lost during the evolution

<sup>4</sup>Present address: Genome Center, University of California, Davis, CA 95616, USA.

<sup>5</sup>Corresponding author.

E-mail [X.Didelot@warwick.ac.uk](mailto:X.Didelot@warwick.ac.uk); fax 44-02476-524532.

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.082263.108>.

of a population among sequences found in at least one of the genomes. Figure 1 gives an illustration of our model.

Inference is performed under this model using a reversible-jump Markov chain Monte Carlo (MCMC) (Green 1995, 2003). Our prior model favors simple explanations for the observed patterns of feature presence and absence (i.e., low rates for gain and loss, and few changes in the rates). Thus, a change in the rate of gain and loss in a particular lineage must be supported by the data to be inferred by our method.

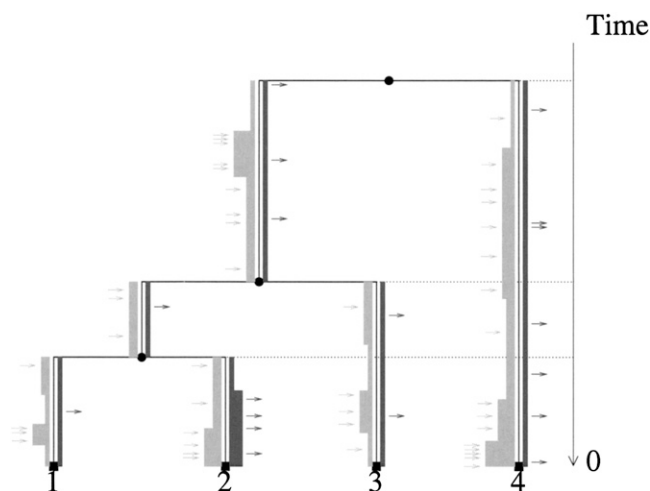
We assess the power of our method using a simulation study and illustrate its use for two groups of  $\gamma$ -Proteobacteria and one group of Firmicutes. In doing so, we demonstrate the ability of our approach to infer genomic flux that involves regulatory regions and fragments of genes. We further demonstrate that, using genome sequence alone, we are able to identify changes in the rate of genomic flux. The rate changes identified by our method are associated with microbial lifestyle changes such as transitions from generalist to host-restricted pathogen lifestyles. We have made a software implementation of the algorithm with a graphical interface freely available from <http://go.warwick.ac.uk/genoplast/>.

## Results

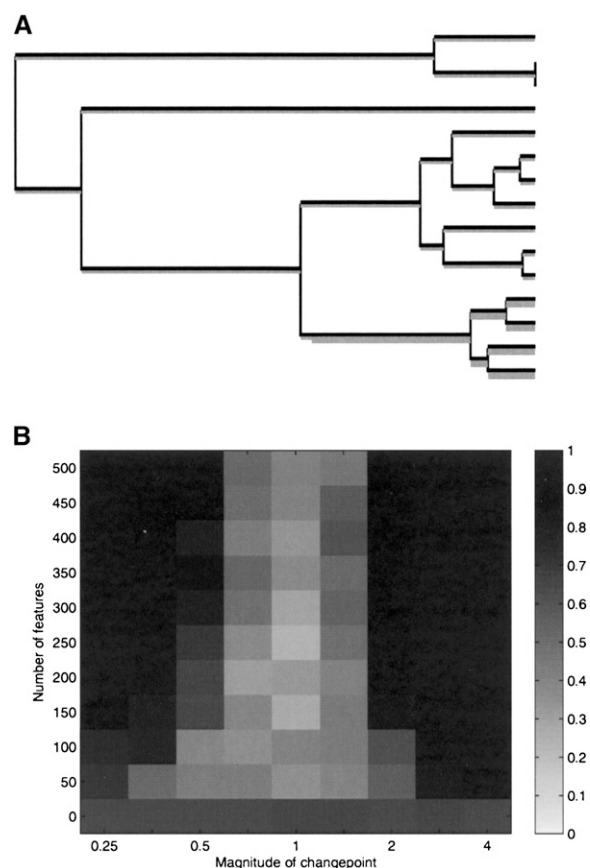
### Simulation study

We simulated a genealogy from the coalescent model (Kingman 1982a,b) for a sample of 15 individuals. Feature gain and loss was simulated on this genealogy under the assumption that  $r$  is constant throughout, and that  $r^+$  contained a single changepoint, as shown in Figure 2A. Several such datasets were simulated with different values for the number of features  $f$  and for the magnitude  $m$  of the change in  $r^+$  at the unique changepoint (i.e., the ratio of values of  $r^+$  after and before the changepoint).

Inference was performed for each of these simulated datasets by running our MCMC algorithm for 20,000 iterations. Figure 2B shows the results for a grid of values of  $f$  from 0 to 500, and of  $m$  from 0.25 to 4.0. Each datapoint in Figure 2B is the average over 20 different simulations and inferences, of the posterior probability



**Figure 1.** Illustration of the model. The branches of the phylogeny  $T$  are in black. The width of the light gray line on the left of the branches is proportional to the value of  $r^+$  (frequency of feature gain). Similarly, the dark gray line on the right of each branch represents  $r^-$  (frequency of feature loss). Individual feature gain events are represented by light gray arrows, and individual feature loss events are represented by black arrows.



**Figure 2.** Simulation study. (A) Coalescent genealogy on which the power study is based, with no changepoint in  $r^-$  and a single changepoint in  $r^+$  on the branch above the last four isolates. (B) Intensity plot of the posterior probability of having exactly one changepoint in  $r^+$  on the correct branch, as a function of the number of independent features and the changepoint magnitude.

that there is exactly one changepoint in  $r^+$  on the correct branch. For  $m = 1$ , the changepoint has no effect in the simulation, and the results are as expected given the Poisson prior with parameter 1 that we used for the number of changepoints in  $r^+$  on the whole tree. Our power to detect the changepoint increases with the number of features  $f$  that we use, and also as the magnitude  $m$  increases above or decreases under 1.

The conditions shown on Figure 2A are in many ways optimal for the detection of the changepoint in  $r^+$ , with no other changepoint on the whole tree confounding its effect, and its location implying an effect on a large subtree. Figure 2B should therefore be interpreted as the maximum that our method can offer in terms of changepoint detection. Yet, it shows that even with a small amount of data (the number of features in actual datasets is much larger), our method is able to detect even subtle changes in the rates.

### Application to *Francisella tularensis*

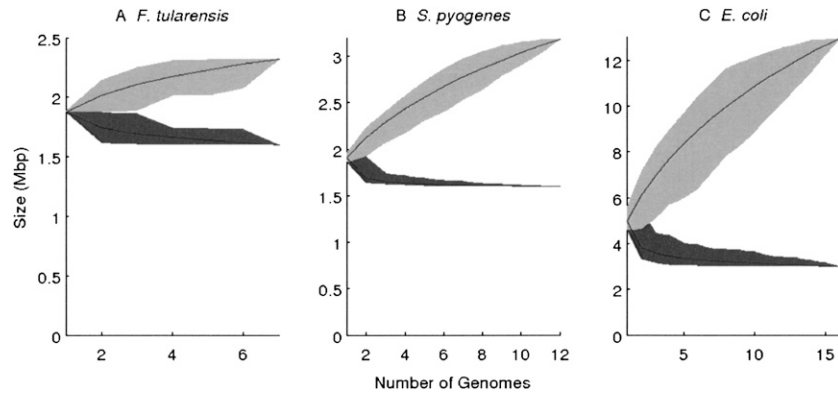
The  $\gamma$ -proteobacterium *Francisella tularensis* is composed of several phenotypically diverse subspecies. The most virulent one is sp. *tularensis*, which causes lethal pulmonary infections in humans and animals (Ellis et al. 2002). A first strain from subspecies *tularensis* was sequenced by Larsson et al. (2005),

and two subsequent sequencing projects showed that it exhibits little genomic diversity (Beckstrom-Sternberg et al. 2007; Chaudhuri et al. 2007). Subspecies *holarctica* is a highly infectious but rarely fatal lineage (Ellis et al. 2002), of which three strains have been sequenced (Petrosino et al. 2006; P. Chain, F. Larimer, M. Land, S. Stilwagen, P. Larsson, S. Bearden, M. Chu, P. Oyston, M. Forsman, S. Andersson, et al., unpubl.; S. Godbole, L. Zhou, D. Bruce, R. Crawford, C. Detter, M. Dempsey, C. Lion, C. Munk, J. Noronha, R. Scheuermann, et al., unpubl.). A sequence from subspecies *novicida*, which is rarely associated with human disease, has also been determined (Rohmer et al. 2007).

Table 1 summarizes the seven *F. tularensis* genomes. We aligned the genomes and determined their phylogeny based on the core genome as described in the Methods section. The average length of each genome is around 1.9 Mbp, of which 1.6 Mbp is found in all genomes (cf. Fig. 3A). The remaining genetic material (i.e., the dispensable genome) is made up of 7216 features of 100 bp present in one to six of the seven genomes.

We reconstructed the history of gain and loss of these features using the algorithm described in the Methods section. Figure 4A shows the inferred flux for the *Francisella* data set, Figure 4B shows the inferred reconstruction of the genome content for the nodes of the phylogeny, and Figure 4C shows the inferred probabilities of gain or loss for each branch and each feature. The genome content of the root node is ambiguous (last row of Fig. 4C), and, therefore, much uncertainty exists regarding the events that occurred on the two branches directly under the root (fifth and last rows of Fig. 4B). The inclusion of the outgroup *novicida* is, however, useful to reduce the uncertainty on the remainder of the branches, in particular the reconstruction of the genome content of the most recent common ancestor of *tularensis* and *holarctica* (cf. penultimate row of Fig. 4C).

We found that a large amount of genetic material was lost on the branch above the ancestor of *tularensis* and *holarctica* with an average of 172 kbp lost (last row on Fig. 4C). This is consistent with the observation that a high proportion (~10%) of their genes have degraded into pseudogenes (Larsson et al. 2005; Petrosino et al. 2006; Beckstrom-Sternberg et al. 2007) in contrast to the non-pathogenic *novicida* (Rohmer et al. 2007). Genome degradation and reduction both result in the disruption of pathways that may



**Figure 3.** Extent of the genomic regions found in all genomes (dark gray), or at least one genome (light gray), as a function of the number of genomes under study.

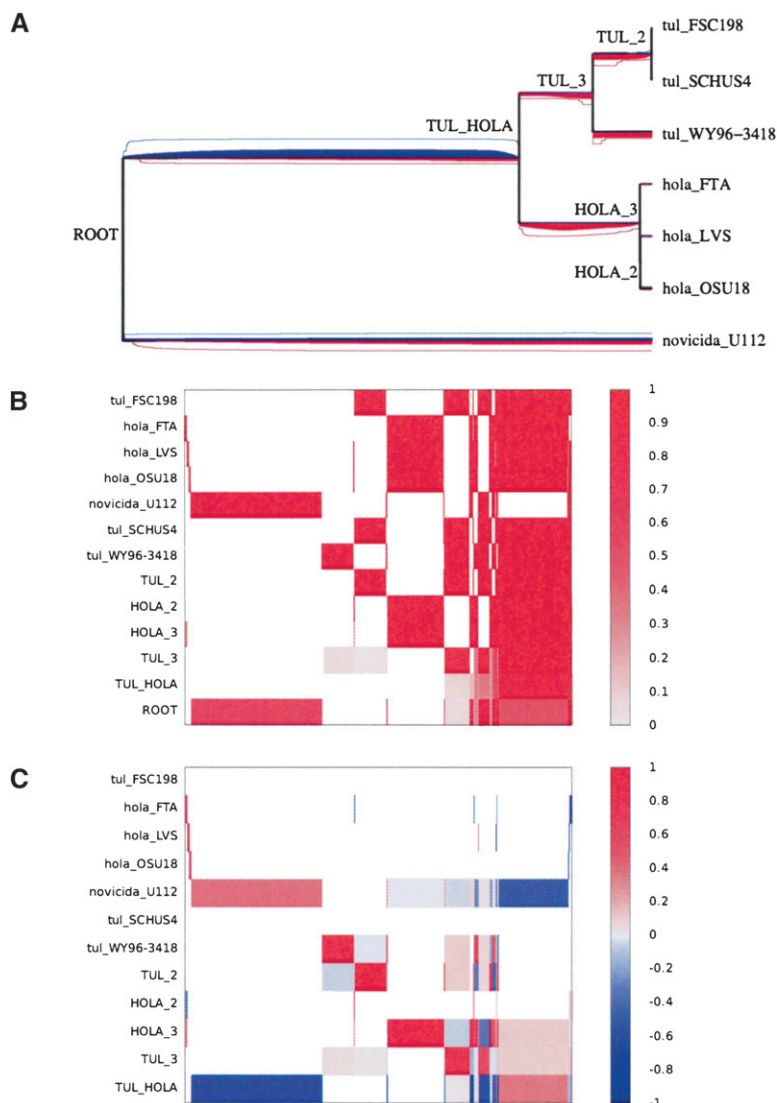
be redundant or even detrimental to the pathogenic lifestyle. Furthermore, a large amount of lateral gene transfer has taken place since the divergence of *holarctica* and *tularensis*, with an average of 144 and 70 kbp gained by each lineage, respectively. However, this flux seems to have stopped after the common ancestor of *holarctica*, with only 12 kbp gained since then by the three genomes of *holarctica* combined. This scenario is compatible with the observation of substantial chromosomal rearrangement within *tularensis* (Beckstrom-Sternberg et al. 2007) as well as between *holarctica* and *tularensis* (Petrosino et al. 2006), but little within *holarctica* (Petrosino et al. 2006).

**Application to *Streptococcus pyogenes***

*Streptococcus pyogenes* is a Gram-positive bacterium responsible for a wide range of human diseases such as bacteremia, tonsillitis, scarlet fever, or acute rheumatic fever (Cunningham 2000). The species is traditionally subdivided according to serologic differences in the M protein, which are strongly correlated with the frequency and type of infection caused. A total of 12 *S. pyogenes* genomes have been sequenced, spanning nine different M types, and we included all of them in this study (cf. Table 2). Previous genome comparisons revealed that the most noticeable difference between those genomes lies in the presence or absence of integrated prophages (Ferretti et al. 2001; Beres et al. 2002; Nakagawa et al. 2003; Banks et al. 2004; Holden et al. 2007). Those prophages contain a number of genes associated with virulence, so that the history of prophage gain and loss is likely to be pivotal to

**Table 1.** Genomes in the *Francisella tularensis* data set

Subspecies	Strain	Size	ORFs	Reference
<i>novicida</i>	U112	1910 kb	1719	Rohmer et al. 2007
<i>holarctica</i>	LVS	1895 kb	1754	P. Chain, F. Larimer, M. Land, S. Stilwagen, P. Larsson, S. Bearden, M. Chu, P. Oyston, M. Forsman, S. Andersson, et al., unpubl.
<i>holarctica</i>	OSU18	1895 kb	1555	Petrosino et al. 2006
<i>holarctica</i>	FTA	1890 kb	1603	S. Godbole, L. Zhou, D. Bruce, R. Crawford, C. Detter, M. Dempsey, C. Lion, C. Munk, J. Noronha, R. Scheuermann, et al., unpubl.
<i>tularensis</i>	SCHUS4	1892 kb	1603	Larsson et al. 2005
<i>tularensis</i>	WY96-3418	1898 kb	1634	Beckstrom Sternberg et al. 2007
<i>tularensis</i>	FSC198	1892 kb	1804	Chaudhuri et al. 2007



**Figure 4.** Results on the *Francisella turalensis* data set. (A) Phylogeny of the sample with the compound level  $r^+$  of feature gain in red below, and the compound level  $r^-$  of feature loss in blue above the branches of the phylogeny. The average level of  $r^+$  and  $r^-$  at each position on the tree is proportional to the quantity of red and blue at that position, and a 95% credibility interval for each rate at each position is shown by two lines of the corresponding color. (B) Probability of presence of each feature in the genome of each node of the phylogeny shown in A. The x-axis represents the different features (ordered according to their pattern of presence in the genomes), and the y-axis shows the different nodes, labeled as in A. (C) Probability of feature gain and loss for the branches of the phylogeny shown in A. The x-axis represents the different features (ordered as in B), and the y-axis shows the different branches, labeled by the name in A of the node directly under the branch.

explaining the different types of infection caused by different lineages.

The average length of the *S. pyogenes* genomes is around 1.9 Mbp, of which 1.6 Mbp is found in all 12 genomes (cf. Fig. 3B). There are 15,794 features of length 100 bp found in a strict subset of the genomes. The phylogeny estimated on the *S. pyogenes* core genome is highly star-like, in agreement with previous studies (Beres et al. 2006; Holden et al. 2007). The nine different M types are approximately equidistant, and the three pairs of genomes sharing the same M types are very closely related (MGAS5005 and SF370 sharing M type 1, SSI-1 and MGAS315 sharing M type 3, MGAS2096 and MGAS9429 sharing M type 12).

Our method avoids a common problem arising with analysis of prophage. The difficulty is that they almost all display some homology (Banks et al. 2004; Holden et al. 2007). Since prophage usually deteriorate faster than the core genome, it is difficult to definitely say whether homologous prophage were both vertically inherited or inherited via lateral transfer. The ambiguous orthology relationship in turn creates ambiguity for inference of prophage gain and loss. Furthermore, intragenomic recombination amongst resident prophages has been described (Nakagawa et al. 2003), which makes the reconstruction of flux even more tedious. Here we used Mauve as described in the Methods section to determine the orthologous regions of prophages. As Mauve is a synteny-based method, this results in a parsimonious evaluation of the gain and loss of prophage features, where all events can be safely assumed to have occurred, but one cannot exclude a more complex history obscured by the homology of different phages.

Our reconstruction of the genetic flux in *S. pyogenes* is shown in Figure 5. As expected, it is dominated by prophage gain and loss. The evolution of genome content seems to follow an approximate molecular clock with three clear exceptions. The first one is a clear increase in the rate of loss on the branch above the three genomes of type M5, M6, and M18, with 168 kbp lost on this relatively short branch. Furthermore, we found a clear increase in the rate of lateral gene transfer for the two genomes of type M1, and also for the two genomes of type M12. 415 kbp, 452 kbp, 430 kbp, and 459 kbp have been gained on the branches directly above SF370, MGAS315, MGAS5005, and SSI-1, respectively. The material gained by these four genomes since they diverged from one another is found in several locations around their genome (Supplemental Fig. 2), which indicates that several distinct insertion events happened on each branch. An increase in

the rate of gain was inferred for two of the three M types where we have two genomes. Such an increase would be impossible to infer for the M types, where we have only one genome, because of the long branches separating the different M types, which make it impossible to infer how recently gain occurred. It is also possible that an increase in the gain of M3 could not be spotted because of the very close relatedness of genomes MGAS315 and SSI-1 within this M type, which makes them virtually identical (Nakagawa et al. 2003).

These results therefore suggest that prophage integration has accelerated in recent times for the genomes of type M1 and M12, but also possibly for several other lineages of *S. pyogenes*. This is consistent with a previous study which found that maximum



**Table 2.** Genomes in the *Streptococcus pyogenes* data set

Strain	M type	Size	ORFs	Reference
MGAS5005	1	1838 kb	1865	Sumby et al. 2005
SF370	1	1852 kb	1697	Ferretti et al. 2001
MGAS10270	2	1928 kb	1987	Beres et al. 2006
SSI-1	3	1894 kb	1861	Nakagawa et al. 2003
MGAS315	3	1900 kb	1865	Beres et al. 2002
MGAS10750	4	1937 kb	1979	Beres et al. 2006
Manfredo	5	1840 kb	1745	Holden et al. 2007
MGAS10394	6	1899 kb	1886	Banks et al. 2004
MGAS2096	12	1860 kb	1898	Beres et al. 2006
MGAS9429	12	1836 kb	1877	Beres et al. 2006
MGAS8232	18	1895 kb	1845	Smoot et al. 2002
MGAS6180	28	1897 kb	1894	Green et al. 2005

likelihood estimates for the rate of genomic flux was higher on the external branches than on the internal branches of a phylogeny of *Streptococcus* (Marri et al. 2006). Another possibility is that prophage integration occurred at a constant rate, but is balanced by prophage excision or deletion. That is, most of the older phage insertions are not visible, as they have been removed, and only recent insertions are detected. A similar observation has been noted in *Salmonella enterica* (Vernikos et al. 2007). The sequencing of additional genomes sharing the same M types should shed more light on these hypotheses, which could reflect a more recent adaptation of lineages to specific niches than previously thought (Marri et al. 2006).

#### Application to *Escherichia coli* and *Shigella*

*E. coli* has long been considered an organism of choice for the study of bacterial pathogenicity due to the coexistence of various pathogenic and commensal lineages. A total of 10 genomes have so far been completely sequenced: three laboratory and commensal strains (MG1655, W3110, and HS), one avian pathogenic strain (APEC O1), two enterohemorrhagic strains (EHEC Sakai and EDL933), one enterotoxigenic strain (ETEC 24377A), and three uropathogenic strains (UPEC CFT073, 536, and UTI89).

We also included in our analysis the six sequenced genomes of the closely related genus *Shigella*: three from species *S. flexneri* (8401, 301, and 2457T) and one from each of the other three species (197, 227, and 046). Table 3 contains the list of these 16 genomes, with references to their original publications. All six strains of *Shigella* are causative agents of bacillary dysentery; hence, they have historically been classified in a separate genus, despite the fact that the *Shigella* phenotype has evolved multiple times from different clones of *E. coli* (Pupo et al. 2000; Jin et al. 2002; Wei et al. 2003). In agreement with this, the phylogeny we inferred for the 16 genomes shows the six strains of *Shigella* split into different phylogenetic groups.

The mean of the lengths of those 16 genomes is approximately equal to 5 Mbp, of which ~3 Mbp are found in all 16 genomes (Fig. 3C). The dispensable genome is made of 99,335 features of length 100 bp. The results of our analysis of genomic flux are shown in Figure 6. The branches on which the least gain and loss occurred are the ones above the three commensal and laboratory strains MG1655, W3110, and HS. These are therefore the closest in genomic content to the genome of the most recent common ancestor of all *E. coli* and *Shigella*.

All of the branches above the six *Shigella* genomes show important gains of genomic material (with an average of 977 kbp gained by each genome), comparable with the ones observed for

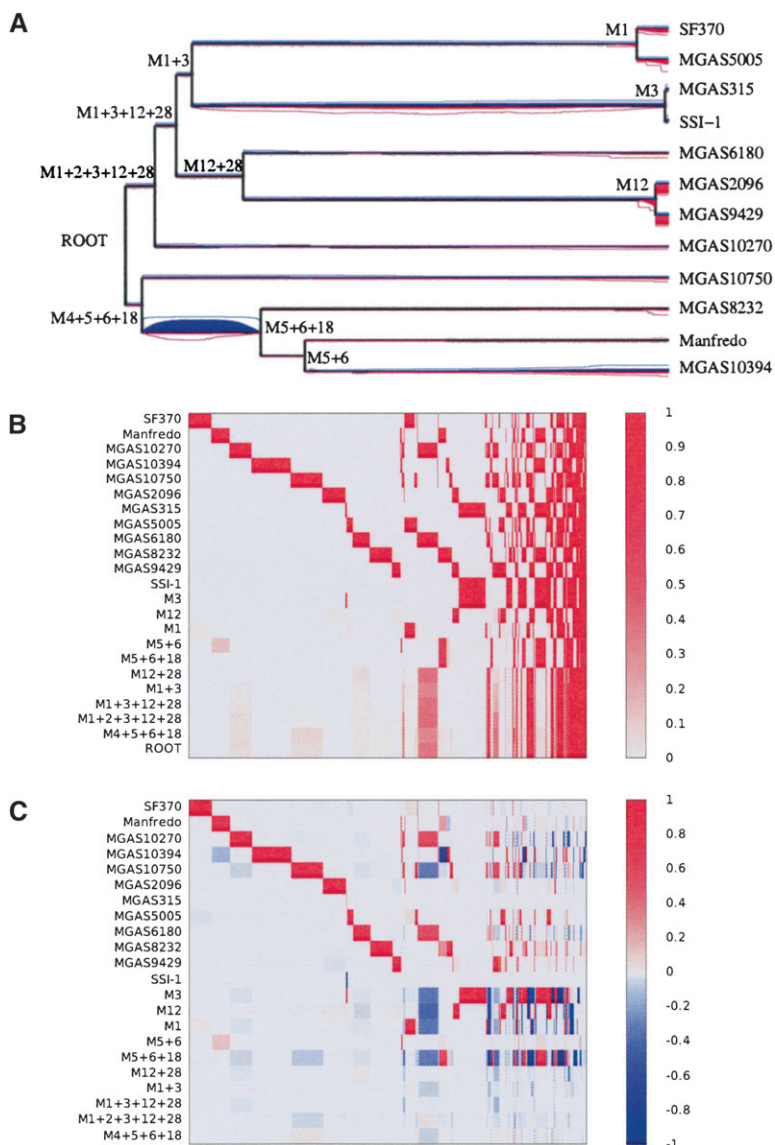
pathogenic strains of *E. coli* such as 24377A or EDL933 and Sakai. However, the *Shigella* genomes have lost many more features than any of the *E. coli* genomes, with an average of 569 kbp lost by each genome. This genomic reduction can be traced back to a higher presence of insertion sequences (IS) in the *Shigella* genomes (Yang et al. 2005). Furthermore, a larger number of pseudogenes is found in the genomes of *Shigella* than in those of *E. coli* (Nie et al. 2006). The fact that the pathogenic *E. coli* have not undergone such genome degradation and reduction (except for APEC O1, cf. below) may be a reflection of their larger host range (Cunningham 2000).

The APEC O1 genome is the only avian pathogenic (APEC) strain in our data set (Johnson et al. 2007). The phylogeny we inferred from the core genome indicates that it is a close relative of the three strains of uropathogenic *E. coli* (UPEC) in our data set, and especially of UTI89. This close relationship, as well as a comparison of the genome sequences and annotation for these four strains, suggest that *E. coli* strains from animals might be the source of uropathogenic *E. coli* infections (Johnson et al. 2007). However, our analysis has found a clear increase in the rate of gain and loss on the branch directly above strain APEC O1, resulting in a gain of 691 kbp on this branch. The increase in the rate of gain is comparable to that found for other branches of pathogenic *E. coli*, but the increased loss is unique to APEC O1 amongst all studied genomes of *E. coli*, and similar to the high rates described above for *Shigella*. This result hints that in spite of the close relationship of APEC O1 with the three UPEC strains, it may have already started to adapt to the avian host. This hypothesis may imply that the natural reservoir of human urinary tract pathogenic *E. coli* is not animals, and will require validation through the sequencing of additional APEC and UPEC strains.

The analysis above uses features of constant size 100 bp, as the unit of genomic flux as described in the Methods section. Our model and algorithm can, however, be applied for any other unit such as the gene, which has been the unit traditionally used in studies of genomic flux (Hao and Golding 2004, 2008; Marri et al. 2006). We therefore reanalyzed the *E. coli* and *Shigella* data set using gene presence/absence data in order to compare the two approaches.

We found a total of 14,752 genes to be present in one, but not all of the 16 genomes. Supplemental Figure 3 shows the result of our analysis of genomic flux based on gene data. The overall inferred history of gene flux is the same as the one described above based on these features: The commensals and laboratory strains have endured little flux, pathogenic strains of *E. coli* have gained some material, and *Shigella* lineages have gained and lost a large amount of genes.

Table 4 contrasts the number of features and genes found to be gained and lost, on average, by both analyses on all the branches of the phylogeny. The gene-based and feature-based analyses are in good agreement, which is not surprising, since the regions of the genomes identified as having been gained and lost are roughly the same in both analyses. For this reason, features and genes are gained and lost in approximately the same proportion on the branches as illustrated in Table 4. Small differences between the two analysis could be caused, for example, by variation in the density of coding genes from one region of the dispensable genome to another, or genome degradation causing a loss of genes (turned into pseudogenes), but not features. The largest difference between the two analyses is found for the amount of loss on the branch above APEC and UPEC, but being directly under the root of the tree, the uncertainty is strong for that branch, with a 95% credibility interval of [1.8;12.1] for the feature-based analysis and [1.1;6.8] for the gene-based analysis (cf. Supplemental Table 1). All



**Figure 5.** Results on the *Streptococcus pyogenes* data set. (A) Phylogeny of the sample with the compound level  $r^+$  of feature gain in red below, and the compound level  $r^-$  of feature loss in blue above the branches of the phylogeny. The average level of  $r^+$  and  $r^-$  at each position on the tree is proportional to the quantity of red and blue at that position, and a 95% credibility interval for each rate at each position is shown by two lines of the corresponding color. (B) Probability of presence of each feature in the genome of each node of the phylogeny shown in A. The x-axis represents the different features (ordered according to their pattern of presence in the genomes), and the y-axis shows the different nodes, labeled as in A. (C) Probability of feature gain and loss for the branches of the phylogeny shown in A. The x-axis represents the different features (ordered as in B), and the y-axis shows the different branches, labeled by the name in A of the node directly under the branch.

the credibility intervals for the amount gained or lost on the different branches are in good agreement when using features and genes.

## Discussion

We have presented a novel method to reconstruct genome content evolution based on whole-genome alignments. Our method is based on a model of genomic evolution that has the essential property of allowing deviations from a molecular clock in acquisition and loss of genetic material. Our use of a relaxed clock is important for two reasons. First, when a lot of material is gained or

lost in a single event (e.g., during phage integration), then we expect high variance in the amount of material flux on branches of the phylogeny, even if the events themselves follow a molecular clock. Second, accumulating evidence suggests that adaptation of an organism to a new niche is accompanied by increased rates of lateral gene transfer (Reid et al. 2000; Marri et al. 2006; Didelot et al. 2007) and/or gene loss (Maurelli et al. 1998; Cole et al. 2001; Welch et al. 2002; Cummings et al. 2004), so that the events themselves do not necessarily occur according to a molecular clock. As such, inferred changes in the rate of gene flux can provide a general means to capture changes in population dynamics or microbial lifestyle.

The methodology we described in order to perform inference under this model of genomic evolution makes use of Bayesian statistics, allowing for a complete quantification of the uncertainty in the reconstruction of material flux. This uncertainty is often large, especially on the branches directly under the genealogy root for which the data at the leaves is not very informative (e.g., Supplemental Table 1). The results can be graphically summarized as illustrated in Figures 4, 5, and 6 with datasets from *F. turalensis*, *S. pyogenes*, and *E. coli*. Our method gives a complete overview of the genomic flux, subsequently fixed in different parts of the phylogeny. As expected, we observed in all three examples a shifting distribution of this flux, which justifies our effort to model a relaxed clock for genomic flux. It is also clear in all three examples that genomic flux is strongly correlated with adaptation to a new lifestyle.

One innovation in the approach taken here is that we do not rely on gene identification. For this reason, the basic unit of our method is the feature (i.e., a sequence fragment of small size) rather than the gene. Clearly, this presents a number of advantages: Gene identification is a laborious process, the quality of existing annotations varies between genomes, and genes are not an indivisible unit of flux. Furthermore, it is always possible, after having found the list of features gained by a genome, to look into its annotation to find the genes (or gene fragments) affected, so that we do not lose the ability to identify gene gains or losses. However, our method can also be applied to gene presence/absence data. The choice of whether to use features or genes depends ultimately on which question is being asked: A gene-based view makes sense if one is interested in differences in functionality, whereas a feature-based view should be favored if one wants to study the mechanism of genomic flux.

**Table 3.** Genomes in the *E. coli* and *Shigella* data set

Species	Strain	Pathogenicity	Size	ORFs	Reference
<i>E. coli</i>	MG1655	None	4639 kb	4243	Blattner et al. 1997
<i>E. coli</i>	W3110	None	4646 kb	4227	Riley et al. 2006
<i>E. coli</i>	HS	None	4643 kb		D. Rasko, M. Rosovitz, C. Brinkley, G. Myers, R. Seshadri, R. Cer, L. Jiang, and J. Ravel, unpubl.
<i>E. coli</i>	O1	APEC	5082 kb	4384	Johnson et al. 2007
<i>E. coli</i>	Sakai	EHEC	5598 kb	4467	Hayashi et al. 2001
<i>E. coli</i>	EDL933	EHEC	5528 kb	5253	Perna et al. 2001
<i>E. coli</i>	E24377A	EPEC	4979 kb	5324	D. Rasko, M. Rosovitz, C. Brinkley, G. Myers, R. Seshadri, R. Cer, L. Jiang, and J. Ravel, unpubl.
<i>E. coli</i>	CFT073	UPEC	5231 kb	4755	Welch et al. 2002
<i>E. coli</i>	536	UPEC	4938 kb	5379	Brzuszkiewicz et al. 2006
<i>E. coli</i>	UTI89	UPEC	5065 kb	4629	Chen et al. 2006
<i>S. flexneri</i>	8401	Dysentery	4574 kb	5044	Chen et al. 2006
<i>S. flexneri</i>	301	Dysentery	4607 kb	4116	Nie et al. 2006
<i>S. flexneri</i>	2457T	Dysentery	4599 kb	4182	Jin et al. 2002
<i>S. dysenteriae</i>	197	Dysentery	4369 kb	4068	Wei et al. 2003
<i>S. boydii</i>	227	Dysentery	4519 kb	4274	Yang et al. 2005
<i>S. sonnei</i>	046	Dysentery	5039 kb	4353	Yang et al. 2005
				4461	Yang et al. 2005

(APEC) Avian pathogenic *E. coli*; (EHEC) enterohemorrhagic *E. coli*; (EPEC) enteropathogenic *E. coli*; (UPEC) uropathogenic *E. coli*.

Breaking down alignment blocks into features or genes as we do in the present work is useful in order to deal with rates of gain and loss in absolute terms (i.e., proportionally with the number of sites being gained or lost). However, we still fall short of a fully event-based reconstruction of history. Since each alignment block is found either in a contiguous region or not at all in each of the genomes, it is likely that each one was gained or lost in a single evolutionary event. By using alignment blocks as the unit of gain and loss instead of genes or features, one might therefore hope to reconstruct events. Unfortunately, alignment blocks often do not correspond to evolutionary units because of events occurring in different parts of the tree. For example, a region that was gained as a single unit in one or more branches of the phylogeny, but is broken up elsewhere, will appear to be two blocks throughout the phylogeny, wherever it is gained. The division into alignment blocks is highly dependent on exactly which genomes are in the sample, with poorly sampled lineages having larger blocks, which can, in turn, mislead inferences based on the rate of gain or loss of blocks.

Reconstructing the full history of evolutionary events that gave rise to the patterns of mosaicism in an observed sample of genome would therefore require a model of genome evolution that includes the possibility for a genome to gain a sequence of an arbitrary size at any position, to lose any subset of its sequence, and to move any subset to a different point (with the possibility of inversion and/or duplication). The inherent complexity of such a model would pose a serious challenge in trying to use it in an inferential setup. The approach we took in the present work avoids those difficulties, at the cost of being less evolutionary oriented.

## Methods

### Alignment

We start with a sample of  $n$  genome sequences from a single bacterial species or a few closely related species. We first produce a multiple alignment of those genomes using the Progressive

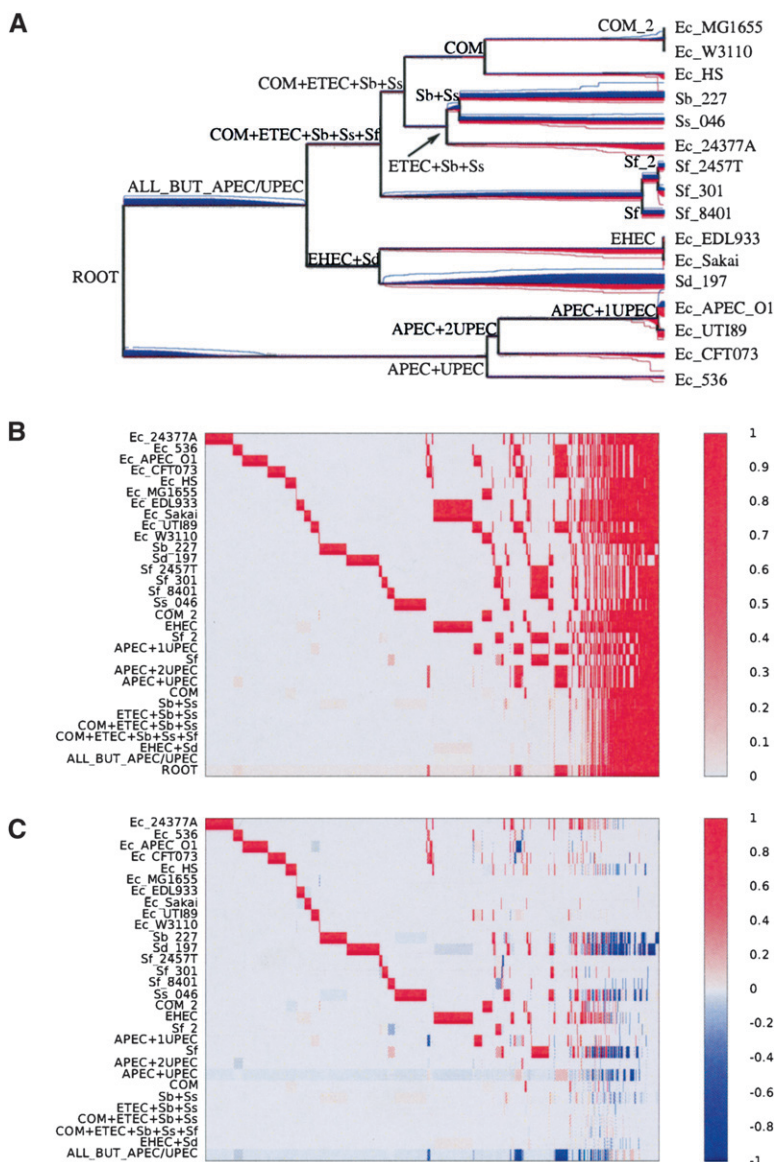
Mauve algorithm (Darling et al. 2004; <http://gel.ahabs.wisc.edu/mauve/index.php>). The Progressive Mauve alignment algorithm identifies and aligns all conserved orthologous segments and all positionally conserved repeat elements. The resulting alignments represent a mosaic of rearranged segments conserved among all genomes, segments conserved among subsets of genomes, and segments unique to a particular genome.

The gaps in a multiple genome alignment can be removed to define the “core genome” of a group of organisms. Gaps in the alignment occur when one or more genomes contain a subsequence not present in remaining genomes. Small alignment gaps are typically caused by mutational processes such as slipped-strand mispairing, whereas large gaps typically result from recombination processes involving gene gain and loss. By excluding alignment columns that participate in gaps larger than some fixed size threshold, for example 20 nt, we can precisely define a set of alignment columns participating in the “core genome.” The core genome can then be used to robustly infer the phylogeny  $T$  of the sample. Here, we used the UPGMA algorithm to do so, but Supplemental Figure 1 shows that neighbor joining, maximum parsimony, and minimum evolution all agree with the UPGMA algorithm, except for one branching order in the *E. coli* data set. Using the other branching order does not affect the results of our genetic flux analysis though.

The remainder of the alignment represents the blocks that have been lost or gained at least once during the evolution of the sample from a common ancestor (also known as the dispensable genome). We consider that each such block is made of small genetic regions of fixed length (i.e., 100 bp) called features. Let  $f$  denote the number of the features thus defined. The dispensable genome can thus be summarized by the binary matrix  $D = \{d_{ij}\} i \in [1..n], j \in [1..f]$ , where  $d_{ij} = 1$  if, and only if, individual  $i$  has the feature  $j$  in its genome.

The reason for choosing a feature size of 100 bp is as follows. Choosing a very small value (e.g., 10 bp) would increase the risk that some of the features do not represent real homologous material in all genomes. On the other hand, choosing a very





**Figure 6.** Results on the *Escherichia coli* and *Shigella* data set. (A) Phylogeny of the sample with the compound level  $r^+$  of feature gain in red below, and the compound level  $r^-$  of feature loss in blue above the branches of the phylogeny. The average level of  $r^+$  and  $r^-$  at each position on the tree is proportional to the quantity of red and blue at that position, and a 95% credibility interval for each rate at each position is shown by two lines of the corresponding color. (B) Probability of presence of each feature in the genome of each node of the phylogeny shown in A. The x-axis represents the different features (ordered according to their pattern of presence in the genomes), and the y-axis shows the different nodes, labeled as in A. (C) Probability of feature gain and loss for the branches of the phylogeny shown in A. The x-axis represents the different features (ordered as in B), and the y-axis shows the different branches, labeled by the name in A of the node directly under the branch.

high value (e.g., 10,000 bp) reduces our power to infer rate changes, since smaller elements are not taken into account, and, under the influence of rearrangements, even a large import can often be split into small subfragments. We consider that a value of 100 bp represents a good middle ground between these two potential issues, but our results are robust to slightly different choices.

### Model of genome evolution

Our model assumes that feature gain follows a compound Poisson process (Huelsenbeck et al. 2000). This means that acquisition of

features follows a Poisson process, whose rate  $r^+$  can vary along the branches of the phylogeny  $T$ . A number  $c^+$  of changes in  $r^+$  are uniformly distributed on  $T$ , and the different values taken by  $r^+$  are independent from one another. The loss of genetic features follows a similar (but fully independent) compound Poisson process with compound rate  $r^-$  containing  $c^-$  changes. All symbolic notations are summarized in Table 5 and an illustration of the model is given in Figure 1.

The likelihood of the compound rates  $r^+$  and  $r^-$  of our model can be decomposed feature-by-feature:

$$\mathbb{L}(r^+, r^-) = \mathbb{P}(D|r^+, r^-) = \prod_{j \in [1..f]} \mathbb{L}_j(r^+, r^-),$$

$$\text{where } \mathbb{L}_j(r^+, r^-) = \mathbb{P}(d_{-j}|r^+, r^-) \quad (1)$$

To calculate this likelihood, let us first consider the probability  $g(u|v, r^+, r^-, l)$  of observing state  $u \in [0, 1]$  at the bottom of a branch of length  $l$  when state  $v \in [0, 1]$  is at the top, and  $r^+$  and  $r^-$  are constant throughout the branch (i.e., there is no changepoint on the branch). This can be calculated by considering a two-state continuous time Markov chain with the transition matrix  $A = [1 - r^+, r^+; r^-, 1 - r^-]$ . Solving the Chapman-Kolmogorov equations for this process (Dynkin 1989) yields:

$$g(u|v, r^+, r^-, l) = \frac{r^+}{r^+ + r^-} \left( \left( \frac{r^-}{r^+} \right)^{1-u} - (-1)^{\delta_{uv}} \times \left( \frac{r^-}{r^+} \right)^v \exp(-l(r^+ + r^-)) \right) \quad (2)$$

Let us now consider the probability  $h(u|v, r^+, r^-, i)$  that a feature is in state  $u$  at node  $i$  in  $T$ , given that it is in state  $v$  at the parent node, and the values of  $r^+$  and  $r^-$ . If there is no changepoint on the branch above the node  $i$ , then  $h(u|v, r^+, r^-, i) = g(u|v, r^+, r^-, l)$ , where  $l$  is the length of the branch above node  $i$ . Otherwise, let  $c_i^+$  and  $c_i^-$  denote the number of changepoints on that branch for  $r^+$  and  $r^-$  respectively. This branch can then be decomposed into  $1 + c_i^+ + c_i^-$  successive segments of lengths  $\{l_k\}_{k \in [1..1+c_i^++c_i^-]}$ , on each of which both  $r^+$  and  $r^-$  are constant.  $h(u|v, r^+, r^-, i)$  can therefore be calculated using the following dynamic programming procedure:

1. Start with  $h(v) = 1$  and  $h(1 - v) = 0$ ;
2. For each consecutive segment  $k \in [1..1 + c_i^+ + c_i^-]$  of length  $l_k$  on which both  $r^+$  and  $r^-$  are constant, repeat the following recursion:

$$\begin{cases} h(0) := g(0|0, r^+, r^-, l_k)h(0) + g(0|1, r^+, r^-, l_k)h(1) \\ h(1) := g(1|0, r^+, r^-, l_k)h(0) + g(1|1, r^+, r^-, l_k)h(1) \end{cases} \quad (3)$$



**Table 4.** Comparison of a feature-based and a gene-based analyses of the *Escherichia coli* and *Shigella* data set

Branch	Features gained	Features lost	Gene gained	Gene lost
Ec_24377A	9.92	0.34	9.12	0.65
Ec_536	3.18	0.33	4.37	0.57
Ec_APEC_O1	6.95	1.32	5.49	2.83
Ec_CFT073	6.24	0.59	8.11	0.86
Ec_HS	3.90	0.95	4.79	0.88
Ec_MG1655	0.06	0.05	0.64	0.08
Ec_EDL933	1.81	0.05	3.69	0.04
Ec_Sakai	1.54	0.07	3.79	0.13
Ec_UT189	2.64	0.08	4.37	0.22
Ec_W3110	0.12	0.04	0.91	0.10
Sb_227	8.83	5.29	8.27	3.72
Sd_197	10.90	7.76	11.44	6.12
Sf_2457T	0.64	0.40	1.62	1.07
Sf_301	2.82	0.28	3.69	0.60
Sf_8401	1.63	0.83	3.17	1.12
Ss_046	11.04	3.43	8.37	2.58
COM_2	3.53	0.62	3.05	0.33
EHEC	12.71	0.80	8.83	0.83
Sf_2	1.43	0.60	2.12	0.91
APEC + 1UPEC	3.30	0.74	2.61	0.49
Sf	6.93	4.64	4.22	3.91
APEC + 2UPEC	0.38	0.24	0.60	0.19
APEC+UPEC	1.05	4.69	1.52	1.83
COM	1.12	0.38	0.63	0.27
Sb+Ss	1.09	0.92	0.15	0.89
ETEC + Sb + Ss	0.25	0.20	0.04	0.21
COM + ETEC + Sb + Ss	0.30	0.23	0.09	0.16
COM + ETEC + Sb + Ss + Sf	0.32	0.34	0.08	0.28
EHEC + Sd	0.42	0.52	0.08	0.56
ALL_BUT_APEC/UPEC	0.50	4.05	0.12	2.67

Each row represents a branch of the phylogeny as labeled on Figure 6, and shows the percentage of features and genes gained and lost on that branch according to the two analyses.

3.  $h(0|v, r^+, r^-, i)$  is equal to  $h(0)$  and  $h(1|v, r^+, r^-, i)$  is equal to  $h(1)$ .

Given this method to calculate  $h(u|v, r^+, r^-, i)$ , it is now possible to apply Felsenstein's pruning (Felsenstein 1973, 1981) to calculate the likelihood component  $\mathbb{L}_j$ :

- For all leaves  $i \in [1..n]$ , set  $f_0(x) := 1-d$  and  $f_1(x) := d$  where  $d = 1$  if the isolate represented by the leaf  $i$  has feature  $j$  and  $d = 0$  otherwise;
- For each internal node  $x$  with children  $y$  and  $z$  taken in increasing order of age, calculate:

$$\begin{cases} f_0(x) := (f_0(y)h(0|0, r^+, r^-, y) + f_1(y)h(1|0, r^+, r^-, y))x... \\ \quad (f_0(z)h(0|0, r^+, r^-, z) + f_1(z)h(1|0, r^+, r^-, z)) \\ f_1(x) := (f_0(y)h(0|1, r^+, r^-, y) + f_1(y)h(1|1, r^+, r^-, y))x... \\ \quad (f_0(z)h(0|1, r^+, r^-, z) + f_1(z)h(1|1, r^+, r^-, z)) \end{cases} \quad (4)$$

3. The likelihood component  $\mathbb{L}_j$  is equal to:

$$\mathbb{L}_j = \mathbb{P}(d_{-j}|r^+, r^-) = f_0(\text{MRCA})(1 - \sigma) + f_1(\text{MRCA})\sigma \quad (5)$$

where  $\sigma$  represents the prior probability that a feature belongs to the genome of the most recent common ancestor of the sample. We estimate  $\sigma$  as the average length of a genome minus the length of the core genome and divided by the total length of the dispensable genome. In doing so, we encode a prior expectation that the ancestral genome size can be approximated as the average of modern genome sizes. The full likelihood follows from Equation 5 using Equation 1. Note that the calculation can be greatly sim-

plified by noticing that any two features with the same pattern of occurrence in the  $n$  genomes contribute equal likelihood components to the overall likelihood, so that each pattern needs to be calculated only once.

### Bayesian inference

We perform Bayesian inference under the model of genome evolution described above. This requires introduction of a prior  $\pi_r$  for each of the different values taken by either  $r^+$  and  $r^-$ , and a prior  $\pi_c$  for the numbers  $c^+$  and  $c^-$  of changepoints in  $r^+$  and  $r^-$ . Using Bayes theorem, the posterior distribution  $\mathbb{P}(r^+, r^- | D)$  can then be decomposed as follows:

$$\mathbb{P}(r^+, r^- | D) = \pi_c(c^+) \pi_c(c^-) \prod_{j \in [1..f]} \mathbb{L}_j \prod_{i \in [1..c^+ + 1]} \pi_r(r_i^+) \prod_{i \in [1..c^- + 1]} \pi_r(r_i^-) \quad (6)$$

where  $r_i^+$  and  $r_i^-$  represent the different values taken by  $r^+$  and  $r^-$ .

We use a MCMC in order to sample from the posterior distribution (Metropolis et al. 1953; Hastings 1970). However, because the dimensionality of  $r^+$  and  $r^-$  depend on the number of changepoints, the dimensionality of the parameter space is not constant. We therefore use a reversible-jump MCMC (Green 1995, 2003). Our updating scheme uses two transdimensional jumps that propose to add and remove a changepoint to either  $r^+$  or  $r^-$ . We also use a move to update the location of a changepoint on a branch, and a move to update the value associated with a changepoint in either  $r^+$  or  $r^-$ . These moves are described in further detail in the Appendix.

Different uninformative priors for  $\pi_r$  and  $\pi_c$  were tested and found to have little effect on the posterior distributions for all three datasets. The results shown used  $\pi_r = \text{Exp}(1)$  and  $\pi_c = \text{Poisson}(1)$ . For each data set, five occurrences of the MCMC were started at different points on the parameter space, chosen according to the prior distribution. Each MCMC was run for

**Table 5.** Table of symbols

Symbol	Description
<b>Data variables</b>	
$n$	No. of isolates
$f$	No. of features
$D = \{d_{i,j}\} i \in [1..n], j \in [1..f]$	Binary feature data
$T$	Phylogeny of the $n$ isolates
$ T $	Sum of branch lengths of $T$
$\sigma$	Probability of presence of a feature in the MRCA genome
<b>Parameters of the model</b>	
$r^+$	Compound rate of feature gain on $T$
$r^-$	Compound rate of feature loss on $T$
$c^+$	No. of change points in $r^+$
$c^-$	No. of change points in $r^-$
<b>Other variables</b>	
$\pi_r$	Prior of the values in $r^+$ and $r^-$
$\pi_c$	Prior on the number of change points in $r^+$ and $r^-$
$h(u v, r^+, r^-, i)$	Probability that a feature goes from state $v$ to state $u$
$g(u v, r^+, r^-, l)$	On the branch above the $i$ -th node; cf. Equation 2 probability that a feature goes from state $v$ to state $u$ on a branch of length $l$ without changepoints; cf. Equation 3
$c_{x,j}$	Indicator of the presence of feature $j$ in the internal node $x$

200,000 iterations, the first half of which was discarded to avoid the influence of the starting point. Each iteration consists of an attempt at each of the moves described in the Appendix. Convergence of the MCMC was judged satisfactory in each case by manual comparison for the five runs of the trajectories of the likelihood,  $c^+$  and  $c^-$ , as well as application of the Gelman-Rubin test (Gelman and Rubin 1992) for  $c^+$ ,  $c^-$ , and the values taken by  $r^+$  and  $r^-$  at the top, middle, and bottom of each branch in the phylogeny. The results presented below for each data set are based on a concatenation of the five instances of the MCMC for maximum robustness.

### Sampling internal states

The location at which features are gained or lost is not explicitly included in our parametrization of the model in order to improve convergence and mixing rates of the MCMC. It is, however, often interesting to know which features have been gained or lost at different points on the phylogeny, and with which posterior probability. Here we show how this can be done by adding a few steps to the dynamic programming algorithm described above for the calculation of the likelihood. Note that this does not interfere in any way with the likelihood calculation, and does not represent a change of parametrization.

In summary, after using a pruning algorithm in a first pass from bottom to top of  $T$  to calculate the likelihood as described above, it is possible to pass again through  $T$  from top to bottom in order to sample the state of each internal node (Hein 1989). This procedure is similar to the forward-backward algorithm of hidden Markov models (Rabiner 1989).

For each node  $x$  of  $T$ , let  $c_{x,j}$  be equal to one if node  $x$  has feature  $j$  and to zero otherwise. The following steps 4 and 5 are added in order to sample  $c_{x,j}$  for all nodes:

4. Draw  $u \sim \text{Unif}([0, 1])$ , and set:

$$c_{\text{MRCA},j} := \left[ u < \frac{f_1(\text{MRCA})}{f_0(\text{MRCA}) + f_1(\text{MRCA})} \right] \quad (7)$$

5. For each internal node  $x$  taken in decreasing order of age, let  $y$  denote the father node of  $x$  in  $T$ , draw  $u \sim \text{Unif}([0, 1])$ , and set:

$$c_{x,j} := \left[ u < \frac{f_1(x)h(1|c_{y,j},r^+,r^-,x)}{f_0(x)h(0|c_{y,j},r^+,r^-,x) + f_1(x)h(1|c_{y,j},r^+,r^-,x)} \right] \quad (8)$$

### Acknowledgments

We thank Bob Mau and Nicole T. Perna for key insights that inspired this work. We also thank three anonymous reviewers for useful comments, ideas, and discussion. This work was funded in part by Wellcome Trust Grant WT082930MA. X.D. was supported by a research fellowship from the Centre for Research in Statistical Methodology (CRiSM). A.E.D. was supported by NSF grant DBI-063075. D.F. was supported by the Science Foundation of Ireland, grant no. 05/FE1/B882.

### Appendix

#### Markov chain Monte Carlo moves

The moves presented below are accepted according to the Metropolis-Hastings-Green ratio:

$$\alpha = \min \left[ 1, (\text{prior ratio PR}) \times (\text{proposal ratio QR}) \times (\text{likelihood ratio LR}) \times (\text{Jacobian J}) \right] \quad (9)$$

LR is equal to the ratio of likelihoods after and before the proposed move and can be calculated using Equation 1. The values of PR, QR, and J are given in each of the move descriptions below.

#### Move an existing changepoint in $r^+$ along a branch of $T$

In this move, one of the  $c^+$  changepoints of  $r^+$  is uniformly chosen. We propose to update the age  $t$  of the changepoint to  $t'$ , which is drawn uniformly on the branch to which the changepoint belongs. This proposal distribution ensures that proposing to move the age of the changepoint from  $t$  to  $t'$  is equally likely to propose a move from  $t'$  to  $t$ , so that QR = 1. Furthermore, the model assumes a uniform distribution of the changepoints on  $T$ , so that PR = 1. Finally, since this jump does not change the dimensionality of the parameter, we have J = 1.

#### Update a value in $r^+$

In this move, one of the  $(c^+ + 1)$  values taken by  $r^+$  is uniformly chosen and proposed to be updated by adding  $u$  to it, where  $u \sim \text{Unif}([- \in Z; \in])$ . If the new value is out of the domain of definition of  $r^+$ , the move is automatically rejected. Proposing to move from the old to the new value is equally likely than proposing to move from new to the old value, so that QR = 1. Furthermore, PR =  $\pi_r(r + u)/\pi_r(r)$  and J = 1.

#### Add/remove a changepoint in $r^+$

This move first decides to add or remove a changepoint, each with probability a half. To add a changepoint, a point  $x$  is chosen uniformly on the branches of  $T$ , and the value  $t$  of  $r^+$  associated with the new changepoint is drawn from  $\pi_r$ . To remove a changepoint, one of the  $c^+$  existing changepoints is chosen uniformly and removed. If no changepoint exists, the removing update is always rejected. Since the age of a new changepoint and its associated value are drawn from a proposal distribution, the Jacobian J is equal to one, even though this move is transdimensional (Troughton and Godsill 1998; Dellaportas et al. 2002; Lopes and West 2004).

If the move proposes to add a new changepoint at  $x$  with associated value  $t$ , we have:

$$\text{PR} = \frac{\pi_c(c^+ + 1)(c^+ + 1)\pi_r(t)}{\pi_c(c^+)|T|} \quad \text{and} \quad \text{QR} = \frac{|T|}{(c^+ + 1)\pi_r(t)} \quad (10)$$

If the move proposes to remove an existing changepoint  $x$ , we have:

$$\text{PR} = \frac{\pi_c(c^+ - 1)|T|}{\pi_c(c^+)c^+\pi_r(t)} \quad \text{and} \quad \text{QR} = \frac{c^+\pi_r(t)}{|T|} \quad (11)$$

### References

- Banks, D.J., Porcella, S.F., Barbian, K.D., Beres, S.B., Philips, L.E., Voyich, J.M., DeLeo, F.R., Martin, J.M., Somerville, G.A., Musser, J.M., et al. 2004. Progress toward characterization of the group A *Streptococcus* metagenome: Complete genome sequence of a macrolide-resistant serotype M6 strain. *J. Infect. Dis.* **190**: 727–738.

- Beckstrom-Sternberg, S.M., Auerbach, R.K., Godbole, S., Pearson, J.V., Beckstrom-Sternberg, J.S., Deng, Z., Munk, C., Kubota, K., Zhou, Y., Bruce, D., et al. 2007. Complete genomic characterization of a pathogenic A.II strain of *Francisella tularensis* subspecies *tularensis*. *PLoS One* **2**: e947. doi: 10.1371/journal.pone.0000947.
- Beres, S.B., Sylva, G.L., Barbian, K.D., Lei, B., Hoff, J.S., Mammarella, N.D., Liu, M.Y., Smoot, J.C., Porcella, S.F., Parkins, L.D., et al. 2002. Genome sequence of a serotype M3 strain of group A *Streptococcus*: Phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc. Natl. Acad. Sci.* **99**: 10078–10083.
- Beres, S.B., Richter, E.W., Nagiec, M.J., Sumbly, P., Porcella, S.F., Deleo, F.R., and Musser, J.M. 2006. Molecular genetic anatomy of inter- and intraserotype variation in the human bacterial pathogen group A *Streptococcus*. *Proc. Natl. Acad. Sci.* **103**: 7059–7064.
- Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**: 1453–1474.
- Bruzskiewicz, E., Brüggemann, H., Liesegang, H., Emmerth, M., Olschlager, T., Nagy, G., Albermann, K., Wagner, C., Buchrieser, C., Emody, L., et al. 2006. How to become a uropathogen: Comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains. *Proc. Natl. Acad. Sci.* **103**: 12879–12884.
- Chaudhuri, R.R., Ren, C.P., Desmond, L., Vincent, G.A., Silman, N.J., Brehm, J.K., Elmore, M.J., Hudson, M.J., Forsman, M., Isherwood, K.E., et al. 2007. Genome sequencing shows that European isolates of *Francisella tularensis* subspecies *tularensis* are almost identical to US laboratory strain Schu S4. *PLoS One* **2**: e352. doi: 10.1371/journal.pone.0000352.
- Chen, S.L., Hung, C.S., Xu, J., Reigstad, C.S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R.R., Ozersky, P., et al. 2006. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach. *Proc. Natl. Acad. Sci.* **103**: 5977–5982.
- Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honore, N., Garnier, T., Churcher, C., Harris, D., et al. 2001. Massive gene decay in the leprosy bacillus. *Nature* **409**: 1007–1011.
- Cummings, C.A., Brinig, M.M., Lepp, P.W., van de Pas, S., and Relman, D.A. 2004. *Bordetella* species are distinguished by patterns of substantial gene loss and host adaptation. *J. Bacteriol.* **186**: 1484–1492.
- Cunningham, M.W. 2000. Pathogenesis of group A streptococcal infections. *Clin. Microbiol. Rev.* **13**: 470–511.
- Darling, A.C., Mau, B., Blattner, F.R., and Perna, N.T. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**: 1394–1403.
- Dellaportas, P., Forster, J., and Ntzoufras, I. 2002. On Bayesian model and variable selection using MCMC. *Stat. Comput.* **12**: 27–36.
- Didelot, X., Achtman, M., Parkhill, J., Thomson, N.R., and Falush, D. 2007. A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: Convergence or divergence by homologous recombination? *Genome Res.* **17**: 61–68.
- Dobrindt, U. and Hacker, J. 2001. Whole genome plasticity in pathogenic bacteria. *Curr. Opin. Microbiol.* **4**: 550–557.
- Dynkin, E.B. 1989. Kolmogorov and the theory of Markov processes. *Ann. Probab.* **17**: 822–832.
- Ellis, J., Oyston, P.C., Green, M., and Titball, R.W. 2002. Tularemia. *Clin. Microbiol. Rev.* **15**: 631–646.
- Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* **22**: 240–249.
- Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively mis-leading. *Syst. Zool.* **27**: 1–4.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**: 368–376.
- Ferretti, J.J., McShan, W.M., Ajdic, D., Savic, D.J., Savic, G., Lyon, K., Primeaux, C., Sezate, S., Suvorov, A.N., Kenton, S., et al. 2001. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl. Acad. Sci.* **98**: 4658–4663.
- Gelman, A. and Rubin, D.B. 1992. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**: 457–511.
- Green, P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–732.
- Green, P. 2003. Trans-dimensional Markov chain Monte Carlo. *Highly Structured Stochastic Systems* **27**: 179–198.
- Green, N.M., Zhang, S., Porcella, S.F., Nagiec, M.J., Barbian, K.D., Beres, S.B., LeFebvre, R.B., and Musser, J.M. 2005. Genome sequence of a serotype M28 strain of group A *Streptococcus*: Potential new insights into puerperal sepsis and bacterial disease specificity. *J. Infect. Dis.* **192**: 760–770.
- Hao, W. and Golding, G. 2004. Patterns of bacterial gene movement. *Mol. Biol. Evol.* **21**: 1294–1307.
- Hao, W. and Golding, G. 2008. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. *BMC Genomics* **9**: 235. doi: 10.1186/1471-2164-9-235.
- Hastings, W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.
- Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.G., Ohtsubo, E., Nakayama, K., Murata, T., et al. 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **8**: 11–22.
- Hein, J. 1989. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. *Mol. Biol. Evol.* **6**: 649–668.
- Hershberg, R., Tang, H., and Petrov, D.A. 2007. Reduced selection leads to accelerated gene loss in *Shigella*. *Genome Biol.* **8**: R164. doi: 10.1186/gb-2007-8-8-r164.
- Holden, M.T., Scott, A., Cherevach, I., Chillingworth, T., Churcher, C., Cronin, A., Dowd, L., Feltwell, T., Hamlin, N., Holroyd, S., et al. 2007. Complete genome of acute rheumatic fever-associated serotype M5 *Streptococcus pyogenes* strain manfredo. *J. Bacteriol.* **189**: 1473–1477.
- Huelsenbeck, J.P., Larget, B., and Swofford, D. 2000. A compound poisson process for relaxing the molecular clock. *Genetics* **154**: 1879–1892.
- Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., Wang, J., Liu, H., Yang, J., Yang, F., et al. 2002. Genome sequence of *Shigella flexneri* 2a: Insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res.* **30**: 4432–4441.
- Johnson, T.J., Kariyawasam, S., Wannemuehler, Y., Mangiamale, P., Johnson, S.J., Doetkott, C., Skyberg, J.A., Lynne, A.M., Johnson, J.R., Nolan, L.K., et al. 2007. The genome sequence of avian pathogenic *Escherichia coli* strain O1:K1:H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes. *J. Bacteriol.* **189**: 3228–3236.
- Kingman, J.F.C. 1982a. On the genealogy of large populations. *J. Appl. Probab.* **19**: 27–43.
- Kingman, J.F.C. 1982b. The coalescent. *Stochastic Process. Appl.* **13**: 235–248.
- Larsson, P., Oyston, P.C., Chain, P., Chu, M.C., Duffield, M., Fuxelius, H.H., Garcia, E., Halltorp, G., Johansson, D., Isherwood, K.E., et al. 2005. The complete genome sequence of *Francisella tularensis*, the causative agent of tularemia. *Nat. Genet.* **37**: 153–159.
- Lawrence, J.G. and Roth, J.R. 1996. Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**: 1843–1860.
- Linz, S., Radtke, A., and von Haeseler, A. 2007. A likelihood framework to measure horizontal gene transfer. *Mol. Biol. Evol.* **24**: 1312–1319.
- Lopes, H.F. and West, M. 2004. Model uncertainty in factor analysis. *Statistica Sinica* **14**: 41–67.
- Marri, P.R., Hao, W., and Golding, G.B. 2006. Gene gain and gene loss in *Streptococcus*: Is it driven by habitat? *Mol. Biol. Evol.* **23**: 2379–2391.
- Maurelli, A.T., Fernández, R.E., Bloch, C.A., Rode, C.K., and Fasano, A. 1998. “Black holes” and bacterial pathogenicity: A large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proc. Natl. Acad. Sci.* **95**: 3943–3948.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. 1953. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**: 1087–1091.
- Nakagawa, I., Kurokawa, K., Yamashita, A., Nakata, M., Tomiyasu, Y., Okahashi, N., Kawabata, S., Yamazaki, K., Shiba, T., Yasunaga, T., et al. 2003. Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res.* **13**: 1042–1055.
- Nie, H., Yang, F., Zhang, X., Yang, J., Chen, L., Wang, J., Xiong, Z., Peng, J., Sun, L., Dong, J., et al. 2006. Complete genome sequence of *Shigella flexneri* 5b and comparison with *Shigella flexneri* 2a. *BMC Genomics* **7**: 173. doi: 10.1186/1471-2164-7-173.
- Ochman, H., Lawrence, J.G., and Groisman, E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299–304.
- Perna, N.T., Plunkett, G., Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., et al. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**: 529–533.
- Petrosino, J.F., Xiang, Q., Karpathy, S.E., Jiang, H., Yerrapragada, S., Liu, Y., Gioia, J., Hemphill, L., Gonzalez, A., Raghavan, T.M., et al. 2006. Chromosome rearrangement and diversification of *Francisella tularensis* revealed by the type B (OSU18) genome sequence. *J. Bacteriol.* **188**: 6977–6985.
- Pol, D. and Siddall, M. 2001. Biases in maximum likelihood and parsimony: A simulation approach to a 10-Taxon Case. *Cladistics* **17**: 266–281.



- Pupo, G.M., Lan, R., and Reeves, P.R. 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc. Natl. Acad. Sci.* **97**: 10567–10572.
- Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**: 257–286.
- Reid, S.D., Herbelin, C.J., Bumbaugh, A.C., Selander, R.K., and Whittam, T.S. 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406**: 64–67.
- Riley, M. and Labeledan, B. 1997. Protein evolution viewed through *Escherichia coli* protein sequences: Introducing the notion of a structural segment of homology, the module. *J. Mol. Biol.* **268**: 857–868.
- Riley, M., Abe, T., Arnaud, M.B., Berlyn, M.K., Blattner, F.R., Chaudhuri, R.R., Glasner, J.D., Horiuchi, T., Keseler, I.M., Kosuge, T., et al. 2006. *Escherichia coli* K-12: A cooperatively developed annotation snapshot. *Nucleic Acids Res.* **34**: 1–9.
- Rohmer, L., Fong, C., Abmayr, S., Wasnick, M., Larson-Freeman, T.J., Radey, M., Guina, T., Svensson, K., Hayden, H.S., Jacobs, M., et al. 2007. Comparison of *Francisella tularensis* genomes reveals evolutionary events associated with the emergence of human-pathogenic strains. *Genome Biol.* **8**: R102. doi: 10.1186/gb-2007-8-6-r102.
- Silva, F.J., Latorre, A., and Moya, A. 2001. Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. *Trends Genet.* **17**: 615–618.
- Smoot, J.C., Barbian, K.D., Van Gompel, J.J., Smoot, L.M., Chaussee, M.S., Sylva, G.L., Sturdevant, D.E., Ricklefs, S.M., Porcella, S.F., Parkins, L.D., et al. 2002. Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. *Proc. Natl. Acad. Sci.* **99**: 4668–4673.
- Spratt, B.G. 1988. Hybrid penicillin-binding proteins in penicillin-resistant strains of *Neisseria gonorrhoeae*. *Nature* **332**: 173–176.
- Sumby, P., Porcella, S.F., Madrigal, A.G., Barbian, K.D., Virtaneva, K., Ricklefs, S.M., Sturde-Vant, D.E., Graham, M.R., Vuopio-Varkila, J., Hoe, N.P., et al. 2005. Evolutionary origin and emergence of a highly successful clone of serotype M1 group A *Streptococcus* involved multiple horizontal gene transfer events. *J. Infect. Dis.* **192**: 771–782.
- Swofford, D., Waddell, P., Huelsenbeck, J., Foster, P., Lewis, P., and Rogers, J. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* **50**: 525–539.
- Troughton, P.T. and Godsill, S.J. 1998. A reversible jump sampler for autoregressive time series. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* **4**: 2257–2260.
- Vernikos, G., Thomson, N., and Parkhill, J. 2007. Genetic flux over time in the *Salmonella* lineage. *Genome Biol.* **8**: R100. doi: 10.1186/gb-2007-8-6-r100.
- Wei, J., Goldberg, M.B., Burland, V., Venkatesan, M.M., Deng, W., Fournier, G., Mayhew, G.F., Plunkett, G., Rose, D.J., Darling, A., et al. 2003. Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect. Immun.* **71**: 2775–2786.
- Welch, R.A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci.* **99**: 17020–17024.
- Wren, B.W. 2000. Microbial genome analysis: Insights into virulence, host adaptation and evolution. *Nat. Rev. Genet.* **1**: 30–39.
- Yang, F., Yang, J., Zhang, X., Chen, L., Jiang, Y., Yan, Y., Tang, X., Wang, J., Xiong, Z., Dong, J., et al. 2005. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res.* **33**: 6445–6458.

Received June 18, 2008; accepted in revised form October 29, 2008.