DOI: 10.1534/genetics.110.120121

Inference of Homologous Recombination in Bacteria Using Whole-Genome Sequences

Xavier Didelot,*,[1] Daniel Lawson,† Aaron Darling‡ and Daniel Falush§,[2]

**Department of Statistics, University of Oxford, Oxford, OX1 3TG, United Kingdom, †Department of Mathematics, University of Bristol, Bristol, BS8 1TW, United Kingdom, ‡Genome Center, University of California, Davis, California 95616 and §Environmental Research Institute, University College, Cork, Ireland*

Manuscript received June 22, 2010
Accepted for publication October 1, 2010

ABSTRACT

Bacteria and archaea reproduce clonally, but sporadically import DNA into their chromosomes from other organisms. In many of these events, the imported DNA replaces an homologous segment in the recipient genome. Here we present a new method to reconstruct the history of recombination events that affected a given sample of bacterial genomes. We introduce a mathematical model that represents both the donor and the recipient of each DNA import as an ancestor of the genomes in the sample. The model represents a simplification of the previously described coalescent with gene conversion. We implement a Monte Carlo Markov chain algorithm to perform inference under this model from sequence data alignments and show that inference is feasible for whole-genome alignments through parallelization. Using simulated data, we demonstrate accurate and reliable identification of individual recombination events and global recombination rate parameters. We applied our approach to an alignment of 13 whole genomes from the *Bacillus cereus* group. We find, as expected from laboratory experiments, that the recombination rate is higher between closely related organisms and also that the genome contains several broad regions of elevated levels of recombination. Application of the method to the genomic data sets that are becoming available should reveal the evolutionary history and private lives of populations of bacteria and archaea. The methods described in this article have been implemented in a computer software package, ClonalOrigin, which is freely available from http://code.google.com/p/clonalorigin/.

BACTERIA and their distant relatives the archaea make up the majority of cellular living organisms. Short generation times combine with enormous population sizes to create tremendous evolutionary potential. It is currently not feasible to track individual organisms in natural conditions to directly observe their evolution. Instead, genomic deoxyribonucleic acid (DNA) sequencing provides a window onto how bacteria disperse, diversify, and adapt because DNA contains information of how organisms are related. In bacteria and archaea, genomic DNA is replicated as part of reproduction by binary fission. Changes in genomic DNA can accumulate because replication is unfaithful or due to DNA damage, but might also be introduced by recombining a segment of foreign DNA into the chromosome. Three mechanisms can lead to the introduction of foreign DNA into a bacterial or an archaeal cell: transduction, conjugation, and transformation. The transduction process transfers DNA via phage infection (Canchaya *et al.* 2003). Conjugation requires two cells to come in contact for DNA to be transmitted from one to the other (Chen *et al.* 2005). Transformation is the uptake of naked DNA from the environment and is regulated by a complex machinery (Claverys *et al.* 2009). Recombination through these three processes has been found to occur frequently in many groups of bacteria and to be a driving force in their evolution and adaptation (Didelot and Maiden 2010).

Recombination in bacteria is analogous to gene conversion rather than crossing over in sexually reproducing organisms (McVean *et al.* 2002), in the sense that the recipient and donor cells make asymmetric contributions to the genetic makeup of the resulting bacterium: typically the donor contributes only a small contiguous segment of DNA whereas the recipient contributes the rest of the genome. For a given set of bacterial isolates, it is thus possible to define its clonal genealogy (Guttman 1997) irrespective of how frequently recombination happened, by tracing back in time the ancestry of the isolates following the line of ancestry of the recipient (and not that of the donor) whenever recombination took place. The clonal genealogy is a bifurcating tree where each leaf is an isolate and each internal node represents the most recent common ancestor of the samples below it. Homologous

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.110.120121/DC1.

[1] *Corresponding author:* Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom. E-mail: xavier.didelot@gmail.com

[2] *Present address:* Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany.

DNA that has been inherited by strict vertical descent evolved according to this clonal tree [this is the so-called clonal frame (Milkman and Bridges 1990)]. However, recombination leads to different parts of the genome having different relationships, each of which can be represented by their own "local tree." Parts of each local tree may be identical to the clonal tree, reflecting vertical descent of DNA, while other parts of the tree can look entirely different due to recombination events bringing in DNA from a different source. Direct evidence for this phenomenon can be found in multilocus sequence typing studies (MLST) (Maiden 2006), where the phylogenies reconstructed at the various loci can be very different from one another, for example in Helicobacter (Achtman *et al.* 1999), Bacillus (Priest *et al.* 2004), or Salmonella (Falush *et al.* 2006).

In previous work, we developed a method to infer the clonal genealogy of a group of organisms, while simultaneously identifying for each branch of that genealogy the genomic locations where recombination occurred. The implementation of that method in software is called ClonalFrame (Didelot and Falush 2007). ClonalFrame has proved useful to identify interesting patterns of recombination in a wide variety of organisms including Campylobacter (Sheppard *et al.* 2008), Neisseria (Didelot *et al.* 2009d), and Francisella (Larsson *et al.* 2009).

To perform efficient inference, ClonalFrame does not model the source of specific recombination events (Didelot and Falush 2007). However, this simplification has two important drawbacks. First, ClonalFrame can identify only recombination events that introduced a number of substitutions higher than expected through mutation alone and will miss events that introduce fewer changes (Didelot and Falush 2007). It does not explicitly account for other signals of recombination, most importantly homoplasy, which occurs when segregating nucleotides at pairs of sites are not consistent with a single tree (Maynard Smith and Smith 1998). The signal of homoplasy could be correctly interpreted if the source of recombination events was modeled. Second, because ClonalFrame does not provide any information about the source of the recombination events it identifies, it cannot be used to infer patterns of gene flow between groups of bacteria. One solution is to post-process the output of ClonalFrame by giving each recombination event a likely origin (Didelot *et al.* 2009a), but this is not as accurate as detecting events and origins at the same time.

Here we introduce a model similar to ClonalFrame, but where the origin of each recombination event is explicitly modeled as a point on the clonal genealogy. The model can therefore be described informally as a tree representing the clonal genealogy, with some additional "recombinant edges" going from one point of the tree to another (Figure 1A) and affecting a subset of the genome. A recombinant edge "arrives" on the tree at the time that recombination occurred from an (unsampled) contemporary bacterium. The ancestry of the unsampled bacterium is followed back in time to its most recent common ancestor with an isolate in our sample giving its "departure" time. The local tree at any given site can be traced back by considering only the recombinant edges affecting the site (Figure 1B).

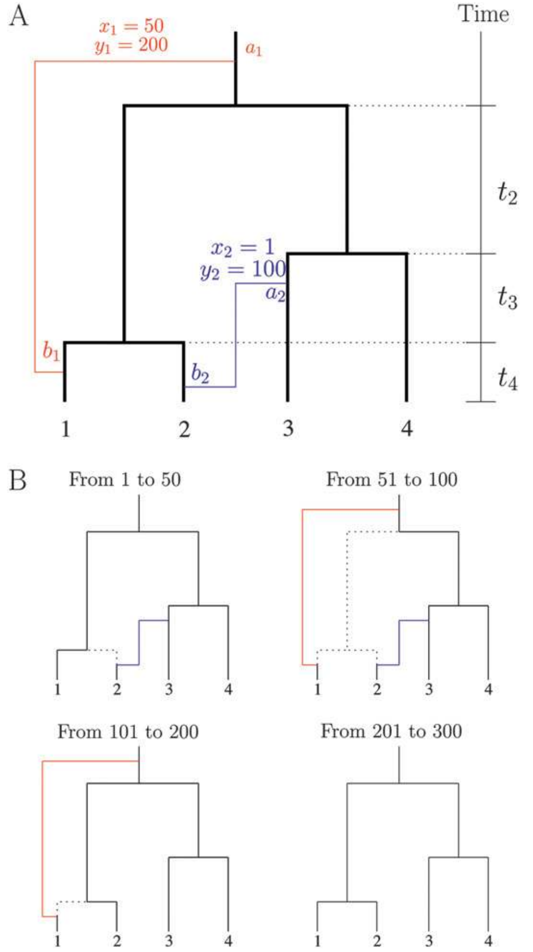

Figure 1.—Illustration of our model for a single region of 300 bp and a sample of four isolates. (A) The full graph of ancestry, with the clonal genealogy shown in thick black lines and two recombination events shown in red and blue. The red event, for example, affected the positions 50–200 of an ancestor of the first isolate at the point b_1, and the donor last shared clonal ancestry with the sample at the point a_1. (B) The local trees for each site. Points a_i are denoted as "departures" of recombinant edges from the tree and b_i are "arrivals," with b_i occurring closer to the observed sequences at the tips.

We show how inference can be performed under this new model and demonstrate that it outperforms ClonalFrame in detecting recombination events in simulated data sets of a closed recombining population. We also illustrate the use of our new model on a data set containing 13 whole genomes of Bacillus. The software we developed to perform inference under our new model is called ClonalOrigin and is freely available from http://code.google.com/p/clonalorigin/.

TABLE 1

Table of symbols

Symbol	Description
	Symbols used for the data
$\mathcal{D}$	Aligned sequence data
N	Number of isolates
L	Total length of the alignment
B	Number of blocks in the alignment
D_b	Data contained in block b
D_s	Data at site s
	Symbols used for the clonal genealogy
$\mathcal{T}$	Clonal genealogy
t_i	Length of time during which there were i ancestors in the clonal genealogy
T	Sum of branch lengths of the clonal genealogy
	Symbols used for the recombination events
$\mathcal{R}$	Set of recombination events and their properties
R	Number of recombination events
R_b	Recombination events affecting block b
a_i	Departure point of the recombination event i
b_i	Arrival point of the recombination event i
x_i	First site affected by the recombination event i
y_i	Last site affected by the recombination event i
T_s	Local tree at site s
	Symbols used for the remaining parameters
$\theta/2$	Rate of mutation on the branches of the clonal genealogy and the recombinant edges
$\theta_s/2$	Per-site rate of mutation
$\rho/2$	Rate of recombination on the branches of the clonal genealogy
$\rho_s/2$	Per-site rate of recombination
δ	Mean of the geometric distribution modeling the length of recombinant segments

MODEL AND METHODS

Model: We use a tree to model the clonal genealogy of organisms and consider recombination events as localized changes to this tree affecting a small region of DNA, resulting in differing local trees for each site. A conceptual representation of this model is given in Figure 1, and the meanings of the mathematical symbols used in the description below are summarized in Table 1.

The tree $\mathcal{T}$ represents the clonal genealogy of the sample of N bacteria under study. We assume a coalescent prior for $\mathcal{T}$ (Kingman 1982), which means that if $t_2, \ldots, t_N$ denote the length of time during which the sample has 2, . . . , N ancestors, respectively, then the probability of the entire genealogy (*i.e.*, the coalescence times plus the tree topology) is given by

$$\mathbf{P}(\mathcal{T}) = \prod_{i=2}^{N} \exp\left(-\binom{i}{2} t_i\right). \qquad (1)$$

The tree $\mathcal{T}$ has total branch length $T = \sum_{i=2}^{N} i t_i$, and along the branches of the clonal genealogy, recombination events occur independently at a constant rate $\rho/2$. Therefore the distribution of the total number R of recombination events is

$$R \mid \mathcal{T}, \rho \sim \text{Poisson}\left(\frac{\rho T}{2}\right). \qquad (2)$$

Each of the $i = 1, \ldots, R$ recombination events is characterized by four variables: (1) an arrival point b_i on the clonal genealogy, (2) a departure point a_i on the clonal genealogy, (3) the site x_i where the recombination starts along the observed genetic material, and (4) the site y_i where the recombination ends along the observed genetic material.

The pair (a_i, b_i) can be represented as a recombinant edge linking two points of the clonal genealogy with b_i occurring closer to the observed sequences at the tips than a_i (Figure 1A). Since recombination happens at a constant rate on the clonal genealogy, the arrival points are independent and identically distributed uniformly on the clonal genealogy; *i.e.*,

$$\mathbf{P}(b_i \mid \mathcal{T}) = \frac{1}{T}. \qquad (3)$$

Given an arrival point, the recombinant edge reconnects with the clonal genealogy at a rate equal to the number of ancestors in the clonal genealogy, as expected under the coalescent model. a_i is therefore distributed as

$$\mathbf{P}(a_i \mid b_i, \mathcal{T}) = \exp(-L(a_i, b_i)), \qquad (4)$$

where $L(a_i, b_i)$ is the sum of the branch lengths of $\mathcal{T}$ found between the time of a_i and that of b_i.

We assume that when recombination occurs, it affects a region that is uniformly distributed along the genome and of length geometrically distributed with mean δ. Therefore when B blocks of the genome are under study for a total sequence length of L, the priors for x_i and y_i are given by

$$\mathbf{P}(x_i = s \mid \delta) = \begin{cases} \frac{\delta}{B\delta + L - B} & \text{if } s \text{ is at the beginning of a block} \\ \frac{1}{B\delta + L - B} & \text{otherwise} \end{cases} \qquad (5)$$

and:

$$\mathbf{P}(y_i = s \mid x_i, \delta) = \begin{cases} \delta^{-1}(1-\delta^{-1})^{s-x_i} & \text{if } s \text{ is before the end of the block} \\ (1-\delta^{-1})^{s-x_i+1} & \text{if } s \text{ is the end of the block} \end{cases} \qquad (6)$$

(Didelot and Falush 2007). Let $\mathcal{R}$ denote the (unordered) list of all recombination events including all their properties. Combining Equations 2–6, we get the complete distribution of $\mathcal{R}$:

$$\begin{aligned}\mathbf{P}(\mathcal{R}\,|\,\mathcal{T},\rho,\delta)\\ &= \mathbf{P}(R\,|\,\rho,\mathcal{T})R!\prod_{i=1}^{R}\mathbf{P}(x_i\,|\,\delta)\mathbf{P}(y_i\,|\,x_i,\delta)\mathbf{P}(b_i\,|\,\mathcal{T})\mathbf{P}(a_i\,|\,b_i,\mathcal{T})\\ &= \exp(-\rho T/2)(\rho/2)^R\prod_{i=1}^{R}\mathbf{P}(x_i\,|\,\delta)\mathbf{P}(y_i\,|\,x_i,\delta)\exp(-L(a_i,b_i)).\end{aligned} \tag{7}$$

On each branch of the clonal genealogy and each recombinant edge mutation events occur at rate $\theta/2$. For simplicity we assume the model of Jukes and Cantor (1969), where all substitutions are equally likely, but our model can equally be used with other mutational processes (Whelan *et al.* 2001).

In this framework, time is measured in nondimensional "coalescent units," with per-generation rates given by $\theta_g = \theta/2N_e$ for mutation and $\rho_g = \rho/2N_e$ for recombination, where N_e is the effective population size (which does not need to be known). It is also useful to define the per-site mutation rate $\theta_s = \theta/L$ and per-site recombination rate $\rho_s = \rho/[(\delta - 1)B + L]$.

Bayesian inference: Let $\mathcal{D}$ denote the set of sequences for which we want to perform inference, related by a known clonal genealogy $\mathcal{T}$. We assume for the moment that the values of the parameters θ, ρ, and δ are also known. We want to perform inference on the posterior distribution:

$$\mathbf{P}(\mathcal{R}\,|\,\mathcal{D},\mathcal{T},\theta,\rho,\delta) \propto \mathbf{P}(\mathcal{R}\,|\,\mathcal{T},\rho,\delta)\mathbf{P}(\mathcal{D}\,|\,\mathcal{T},\mathcal{R},\theta). \tag{8}$$

The first term (the prior) is given by Equation 7. To compute the second term (the likelihood), we define the local tree T_s of each site $s = 1,\ldots, L$ as the tree obtained by following the recombinant edges for which $x_i \leq s \leq y_i$ (*cf.* Figure 1B) and the clonal genealogy otherwise. The data D_s observed at site s depend on the ancestry graph only through the local tree T_s and therefore the likelihood can be decomposed as

$$\mathbf{P}(\mathcal{D}\,|\,\mathcal{T},\mathcal{R},\theta) = \prod_{s=1}^{L}\mathbf{P}(D_s\,|\,T_s,\theta), \tag{9}$$

where each of the terms in the product can be computed using the pruning algorithm of Felsenstein (1981). This algorithm provides a natural way of dealing with gaps in the alignment by treating them as missing data.

To perform inference, we use a Monte Carlo Markov chain (MCMC) with reversible jumps (Green 1995). Briefly, our update scheme is made of two reversible-jump moves: a "remove" move, which proposes to remove an existing recombinant edge chosen uniformly at random, and an "add" move, which proposes to add a recombination event with properties proposed according to their priors as defined in Equations 3–6. These two moves are accepted according to their Metropolis–Hastings–Green ratio as described in appendix a. We also use nontransdimensional moves proposing to update the departure point, arrival point, starting site, and finishing site of an existing recombinant edge, as described in appendix a.

Inference using whole genomes: The previous section described how to infer the recombination events $\mathcal{R}$ from some data $\mathcal{D}$, assuming knowledge of $\mathcal{T}$, θ, ρ, and δ. Direct inference could in principle be done when those quantities are unknown by adding MCMC moves for those parameters, including phylogenetic updates as originally proposed by Yang and Rannala (1997) and by Mau and Newton (1997). However, because we are primarily interested in inference using whole genomes, such a scheme would be unable to converge because the combined parameter space is extremely large and a parallelization scheme is difficult to implement efficiently. When $\mathcal{T}$, θ, ρ, and δ are known, inference can be greatly simplified by noting that the recombination events affecting the various alignment blocks $b = 1, \ldots, B$ are independent. In other words, if D_b denotes the subset of the data corresponding to the block b and R_b denotes the subset of recombination events affecting the block b, then we have

$$\mathbf{P}(\mathcal{R}\,|\,\mathcal{D},\mathcal{T},\theta,\rho,\delta) = \prod_{b=1}^{B}\mathbf{P}(R_b\,|\,D_b,\mathcal{T},\theta,\rho,\delta). \tag{10}$$

Thus inference when $\mathcal{T}$, θ, ρ, and δ are known can be done even for a large genomic alignment by parallelization of the inference of the recombination events for each alignment region. Alignment regions are induced by genomic rearrangement processes (Darling *et al.* 2004) and it is convenient to treat them as independent (given $\mathcal{T}$, θ, ρ, and δ) as previously proposed (Didelot and Falush 2007). Furthermore, when whole genomes are being used, the statistical uncertainty on $\mathcal{T}$, θ, ρ, and δ is likely to be small. We therefore decompose the inference for whole-genome alignments into a three-step process:

Step 1. Infer the clonal genealogy $\mathcal{T}$ given the data $\mathcal{D}$.

Step 2. Infer the mutation rate θ, recombination rate ρ, and average tract length of recombination δ given the data $\mathcal{D}$ and the clonal genealogy $\mathcal{T}$ inferred in step 1.

Step 3. Infer independently for each alignment block b the recombination events R_b affecting b given the data D_b, the clonal genealogy $\mathcal{T}$ inferred in the first step, and the parameters θ, ρ, and δ inferred in the second step.

In practice, we perform step 1 using the ClonalFrame algorithm (Didelot and Falush 2007). Step 2 is performed by running the inference under our model for each alignment block independently, with θ, ρ, and δ treated as additional parameters (*cf.* appendix b for the corresponding MCMC moves). The median value inferred for all blocks is then used as a constant value of θ, ρ, and δ when performing step 3.

TABLE 2

Comparison with the ancestral recombination graph

	S		*H*		r^2		*A*	
ρ/θ	Full	Approx	Full	Approx	Full	Approx	Full	Approx
0.0	8.7	8.8	7.0	7.1	0.19	0.18	0.0	0.0
0.5	8.7	8.8	7.5	7.5	0.14	0.14	0.9	0.9
1.0	8.7	8.9	7.8	7.9	0.11	0.12	1.7	1.7
1.5	8.8	8.9	8.1	8.1	0.10	0.10	2.2	2.4
2.0	8.7	9.0	8.3	8.4	0.08	0.09	2.6	2.9
2.5	8.7	9.1	8.4	8.6	0.08	0.08	2.9	3.3
3.0	8.8	9.2	8.6	8.8	0.07	0.07	3.3	3.7
3.5	8.7	9.3	8.7	8.9	0.07	0.07	3.4	4.1
4.0	8.7	9.3	8.8	9.0	0.06	0.07	3.6	4.3
4.5	8.7	9.5	8.9	9.2	0.06	0.06	3.8	4.8
5.0	8.8	9.5	9.0	9.3	0.06	0.06	4.0	5.1

Approx, approximate.

RESULTS

Relationship with the ancestral recombination graph: Although we have described our model independently, it is natural to think about it as a simplification of the ancestral recombination graph (ARG) with gene conversion (Wiuf and Hein 2000; Didelot *et al.* 2009c). Our model is in fact equivalent to an ARG model in which nonclonal lines of ancestry are not allowed to either recombine or coalesce with each other. These two simplifications can be justified if we consider that the recombination rate (ρ) is relatively low. Nonclonal lines carry little ancestral material (of order δ/L) and therefore have a low effective recombination rate, so that they are unlikely to recombine in the full ARG model into two ancestors with nonempty ancestral material. Furthermore, two nonclonal lines that coalesce in the ARG are unlikely to carry overlapping ancestral material and ignoring such events has been shown to have little effect in the crossover ARG (McVean and Cardin 2005; Marjoram and Wall 2006).

The simplifications in our model relative to the ARG are motivated by our desire to perform inference under the model for very large data sets. Inference under the full ARG process is difficult for data sets of nontrivial size (Stumpf and McVean 2003), but our simplification implies that each recombinant edge can be added and removed independently in the MCMC, which greatly simplifies inference. Furthermore, the blockwise-independence property of our model in Equation 10 allows inference to be performed independently for each region of an alignment, but this property does not hold for the full ARG model since two recombinant edges affecting different regions may coalesce with each other.

To test our approximation, we simulated data under both our model and the ARG using SimMLST (Didelot *et al.* 2009c). In each simulation we considered two loci, with a per-site mutation rate $\theta_s = 0.01$ and a recombination tract length $\delta = 500$ bp. Table 2 shows the average value under both models of the following data statistics often used to study recombination: number of segregating sites *S*, number of unique alleles *H* (Wall 2000), measure of linkage between the loci r^2 (Hill and Robertson 1968), and measure of homoplasy between the loci *A* (Maynard Smith and Smith 1998). When $\rho = 0$ (first row of Table 2), the two models reduce to a tree model and are formally equivalent. For low values of the recombination rate ($\rho/\theta = 0.5$ or 1) the two models are indistinguishable on the basis of the summary statistics considered. As the recombination rate increases, we find that the approximate model generates systematically higher values of *S*, *H*, and *A* (Table 2). This explains why when inference is performed under our model using data from the ARG, the recombination rate tends to be overestimated (*cf.* next section). The measure r^2 of the linkage remains the same between the two models even for higher values of the recombination rate (Table 2).

Application to simulated data: We used SimMLST to simulate sequence data under the ARG model for a representative range of parameters. We then applied our algorithm to infer the recombination events and rate ρ given the tree, the mutation rate θ, and the recombination tract length δ. We consider sequences of length 10,000 bp, which is characteristic of genomic alignment block sizes.

Inference on an ARG with $N = 10$ sequences, $\theta = 300$, $\rho = 50$, and $\delta = 236$ bp is considered in Figure 2. There are no instances of confidently inferred but incorrect recombination events in this (typical) example, with false-positive recombination intensity being limited to two types. First, the boundary of the recombination region is sometimes imperfectly found (*e.g.*, on branch 1 around 5200 bp), and second, the origin may be incorrect (*e.g.*, parent of sequences 2 and 8, 100 bp). In both of these cases the error is "small" in the sense that the prediction is close to the true value. Several kinds of uncertainty are captured: the event itself may be uncertain, the arrival branch may be unclear (*e.g.*, an arrival at sequences 9 and 10 or their parent at 5500 bp), the recombination may have poorly defined boundaries, or the origin may be poorly determined. In the older part of the ancestry, the inference becomes less certain because the data become less informative. In addition to making no false-positive claims about recombination event arrival in this example, ClonalOrigin captures a much larger set of the recombined regions than does ClonalFrame. Many events in the full ARG do not change the tree topology, or contain no mutations, and are therefore undetectable. The inferred recombination rate in the ClonalOrigin model has mean $\rho = 62.5$ [95% confidence interval (45.5, 83.6)] and in the ClonalFrame model has mean $\rho = 25.4$ [95% confidence interval (16.0, 37.6)].

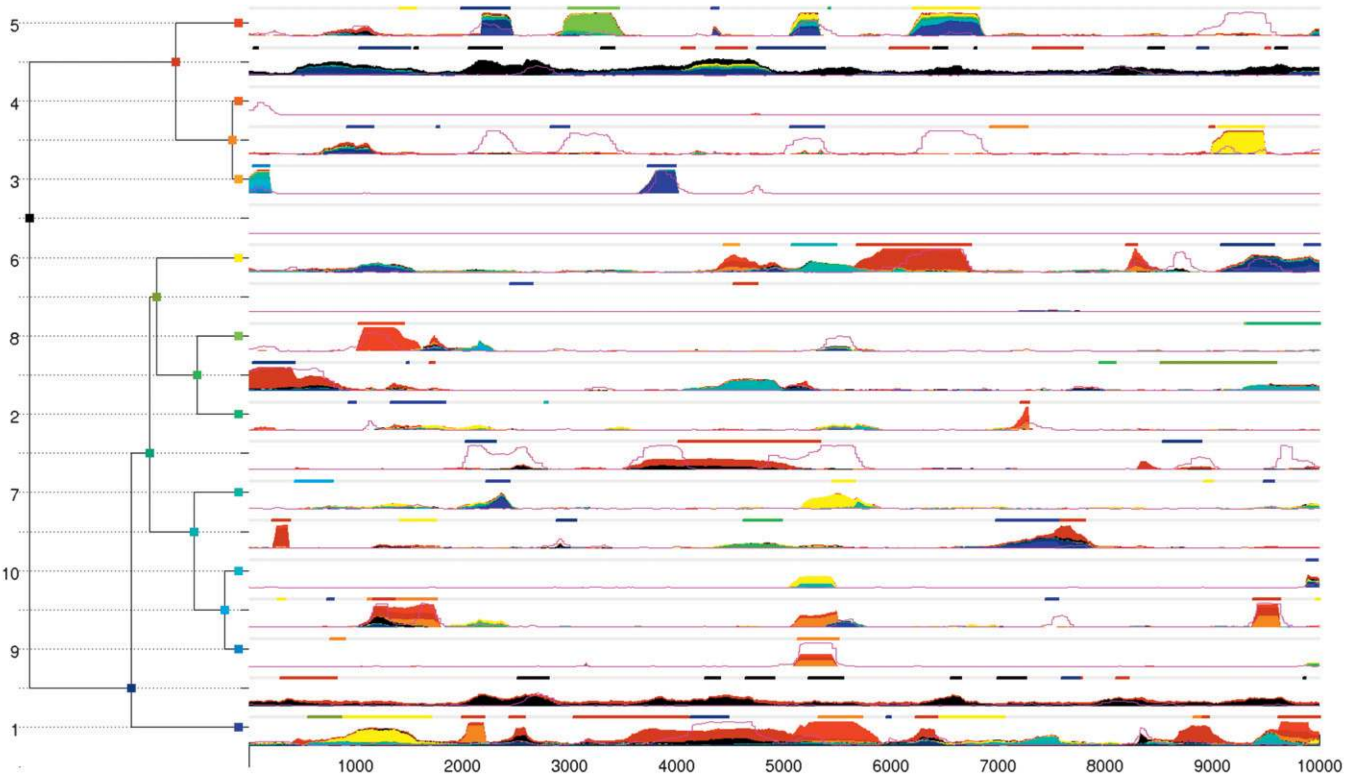

FIGURE 2.—Results on simulated data for a single simulation. The clonal genealogy is shown on the left, and each node is given a color. Each horizontal row on the right represents the arrival of recombination on the branch of the clonal genealogy it is aligned with. For each row, the *x*-axis represents the sequence measured in base pairs and the *y*-axis represents the probability of recombination on a scale from 0 (where the magenta line is most of the time) to 1 (just below the light gray line). ClonalFrame inference is represented by a thin magenta line. ClonalOrigin inference is shown in solid colors according to their reconstructed origin. Small bars above each row correspond to the true recombined regions in the ARG and are colored according to their origin (or in very light gray to represent absence of recombination). For example, on the branch above genome 9, two real events have occurred, both from an "orange" origin. The first one (around position 900) was fairly short and therefore stayed undetected. The second one (around position 5200) was detected by ClonalFrame with posterior probability close to 100% and by ClonalOrigin with posterior probability ~50% and an origin very likely to be orange but that could also be brown or red.

Having established that our algorithm can correctly recover simulated recombination events, we consider how many events we capture as we vary other parameters. In Figure 3 we consider the inferred ρ for a range of ARGs simulated with $N = 20$, $\delta = 236$, and varying $\rho = (25, 50, 75, 100, 150, 200, 250, 300, 400)$ and $\theta = (50, 100, 200, 300, 400, 500)$. We average over 10 ARGs for each set of parameters to reduce variability, which can be very large under the ARG model. ClonalOrigin infers ρ much closer to the true value than does ClonalFrame, which tends to underestimate ρ by roughly a factor of 2 because it misses events that have origins close to the departure point on the tree. ClonalOrigin infers the correct recombination rate for low ρ and overestimates ρ when the mutation rate θ is high. We conjecture that this happens because the full ARG model allows recombination events to recombine and coalesce, for which ClonalOrigin infers additional events to represent the resulting mosaic of origins. Such mosaic imports to the clonal lineage become more common as recombination rate increases and are easier to detect as mutation rate increases. Therefore the recombination rate inferred by ClonalOrigin corresponds to the true recombination rate in the limit of small ρ (and large L), but should be interpreted in terms of the number of distinct recombined tracts (rather than recombination events) as ρ increases.

Application to a Bacillus genomic data set: Bacteria from the *Bacillus cereus* group live predominantly in the soil, feeding from dead organic matter, but occasionally infect humans where they can inflict diseases ranging from food poisoning to deadly anthrax (STENFORS ARNESEN *et al.* 2008). MLST has been applied to the *B. cereus* group to investigate its population structure and history (PRIEST *et al.* 2004; SOROKIN *et al.* 2006). Three major phylogenetic clades have been found, which do not agree with species designations (PRIEST *et al.* 2004; SOROKIN *et al.* 2006; DIDELOT *et al.* 2009a). Analysis of MLST data using ClonalFrame found that recombination occurs at a rate approximately one-fifth of that of mutation ($\rho/\theta \approx 0.2$) and results in a greater number of substitutions being introduced ($r/m \approx 1.5$) (DIDELOT and FALUSH 2007; DIDELOT *et al.* 2009a; VOS and DIDELOT 2009).

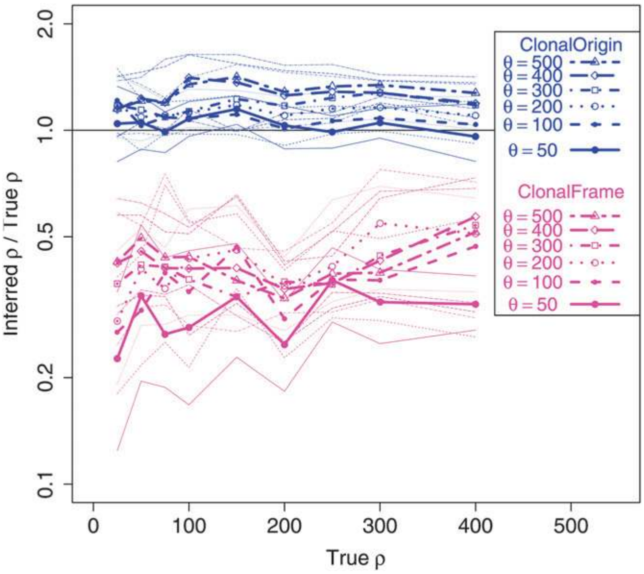

FIGURE 3.—Inferred values of ρ relative to true values for many simulated data sets across various parameter values. Shown are values for ClonalFrame (magenta) and ClonalOrigin (blue). For each of the six values of θ we plot the median (thick line) and interquartile range (thin line) of the ratio of inferred ρ/true ρ, considering the combined results for 10 different instances of the ARG. Lines are labeled by the order they appear at ρ = 400. The true value of 1 is shown as a horizontal line for comparison.

Since the sequencing of the first genome of *B. cereus* by IVANOVA *et al.* (2003), several more isolates have been fully sequenced (RASKO *et al.* 2004; HAN *et al.* 2006; CHALLACOMBE *et al.* 2007; RAVEL *et al.* 2009; XIONG *et al.* 2009). We collected 13 such genomes (9 from clade 1, 3 from clade 2, and 1 from clade 3), summarized in Table 3, and aligned them using progressiveMauve (DARLING *et al.* 2004, 2010). We found $B = 1218$ blocks of homologous sequence shared between all genomes, with lengths ranging from 502 to 55,619 bp and combined length $L = 3{,}636{,}155$ alignment columns. Cumulative alignments of subsets of those 13 genomes indicated that most of the material shared between them is likely to be part of the core genome shared by all members of the group (supporting information, Figure S1). We applied GenoPlast (DIDELOT *et al.* 2009b) to the material not shared by all genomes and found that the rates of gain and loss of material have been approximately constant during the evolution of the sample, except for a recent acceleration of the rate of gain for the genomes in clade 1 (Figure S2).

Application of the step-by-step methodology: The first step of our analysis was to reconstruct the clonal genealogy of the sample using ClonalFrame (DIDELOT and FALUSH 2007). A unique tree topology was inferred, with little uncertainty in the branch lengths (Figure S3). The same topology was also found when using UPGMA, neighbor joining, maximum likelihood, or minimum evolution (Figure S4). ClonalFrame found that many recombination events happened during the evolution of the sample, as shown in Figure S3. These were found to happen at rate $\rho/\theta \approx 0.21$ with an interquartile range (IQR) of (0.20; 0.23) relative to mutation, to be of average length $\delta = 171$ bp [IQR (168; 175)], and to result in $r/m = 2.41$ [IQR (2.37; 2.45)] more substitutions introduced by recombination than by mutation.

TABLE 3

Genomes of the *Bacillus cereus* group used in this study

Genome	Clade	Length	GenBank
1 *B. anthracis* Ames 0581	1	5503926	NC_007530
2 *B. cereus* 03BB102	1	5449308	NC_012472
3 *B. cereus* AH187	1	5599857	NC_011658
4 *B. cereus* AH820	1	5588834	NC_011773
5 *B. cereus* ATCC 10987	2	5432652	NC_003909
6 *B. cereus* ATCC14579	2	5427083	NC_004722
7 *B. cereus* B4264	2	5419036	NC_011725
8 *B. cereus* G9842	1	5736823	NC_011772
9 *B. cereus* Q1	1	5506207	NC_011969
10 *B. cereus* ZK	1	5843235	NC_006274
11 *B. thuringiensis* Al Hakam	1	5313030	NC_008600
12 *B. thuringiensis* konkukian	1	5314794	NC_005957
13 *B. weihenstephanensis* KBAB4	3	5872743	NC_010184

We then applied our new model to infer the recombination events (R_b), mutation rate, recombination rate, and recombination tract length independently for each block of the alignment. The values inferred for each block are shown in Figure 4. The median value for the recombination tract length δ was 236 bp [IQR (131; 537)]. A few blocks, however, took extremely low or high values, reflecting the limited information available on δ when working with a single block. The median value for the per-site mutation rate ($\theta_s/2$) was 0.0219 [IQR (0.0171; 0.0277)]. This was found to be fairly constant throughout the blocks. The median value for the per-site recombination rate ($\rho_s/2$) was 0.0087 [IQR (0.0047; 0.0173)]. Higher rates of recombination were found in three regions of the genome (Figure 4). The median inferred value of ρ/θ was 0.4051, which is almost twice as high as found by ClonalFrame. This reflects the higher sensitivity of our new model to detect recombination.

Finally we completed the analysis by applying our new model again with values of δ, θ_s, and ρ_s fixed to the median values of the previous paragraph.

Patterns of recombination inferred across the genomes: We estimated that ~240,000 recombination events occurred since the 13 genomes shared a common clonal ancestor, but most of these events affected the deep branches of the clonal genealogy, where the statistical uncertainty about each event is very high. Figure 5 shows the numbers of recombination events found by our analysis for any recipient/donor combination of branches, relative to their expectation under the inferred recombination rate using Equation 7. The main pattern in Figure 5 is that genomes recombine

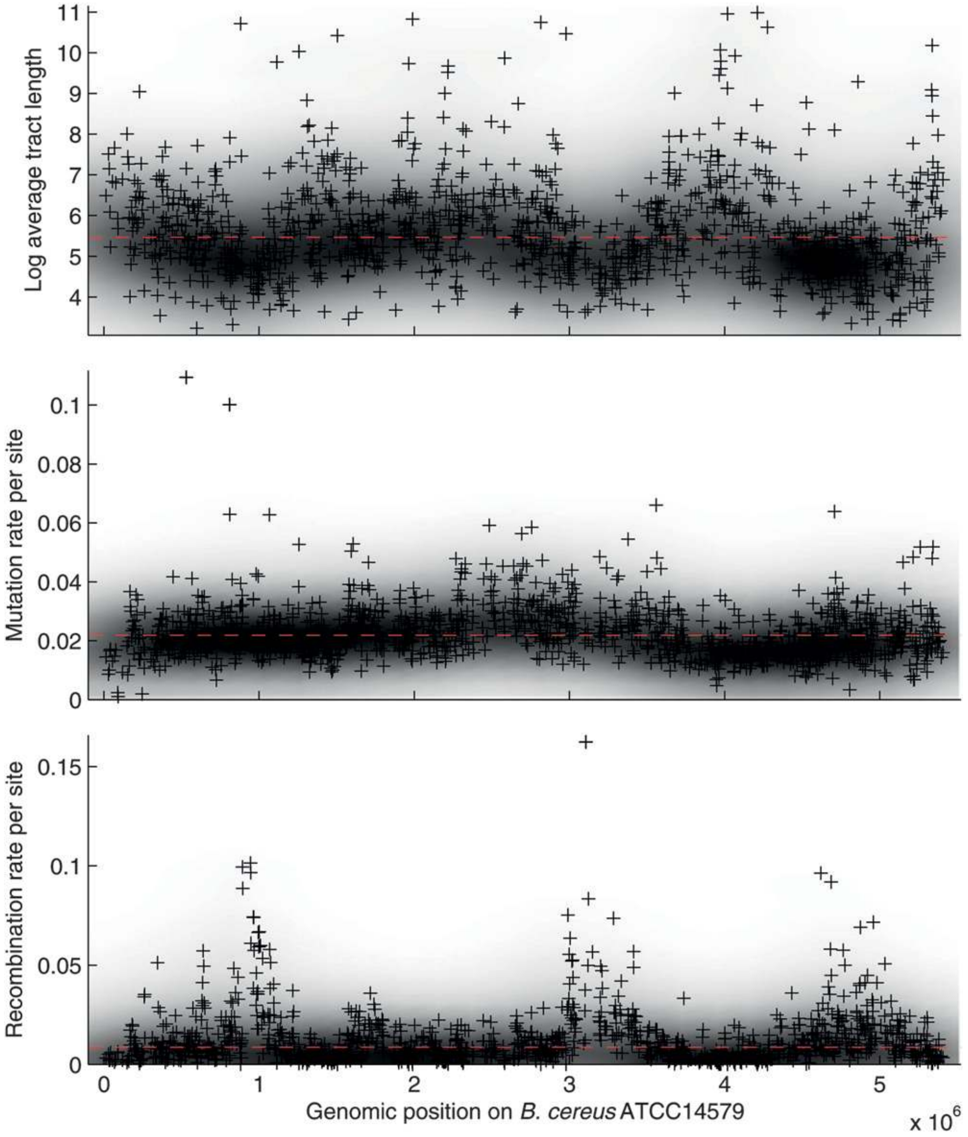

FIGURE 4.—Scatterplots for all blocks of the stage 2 analysis of Bacillus, showing the inferred values of the log-average tract length ($\log(\delta)$), the mutation rate per site ($\theta_s/2$), and recombination rate per site ($\rho_s/2$). A density plot of the scatterplot is shown using gray shading. The median for all blocks is shown in red.

more within clades than between clades. The pattern is particularly visible in clade 1 and exists despite our algorithm having increased power to detect recombination between more divergent sequences. This result may not be surprising considering that recombination in bacteria is sequence identity dependent in the laboratory (*e.g.*, MAJEWSKI 2001). In the *B. cereus* group, genetic exchanges have previously been found to occur more often within than between clades using MLST data (DIDELOT *et al.* 2009a). Figure 5 also contains evidence for a weaker sexual isolation between the two subclades of clade 1, consisting, respectively, of genomes (1, 2, 4, 10, 11, 12) and (3, 5, 9).

The imports inferred on the branch ancestral to genomes 2 and 11 are atypical for two reasons. First, this branch has imported approximately two times more material from external sources (*i.e.*, origins above the most recent common ancestor of our sample) than expected under the prior, even though the dependency of recombination to sequence identity should limit the frequency of such imports. Second, many imports have been detected coming from genomes 1 and 4. Even though such imports represent within-clade recombination, their increased number relative to the prior seems to go beyond the general increase caused by homology dependency for the rest of the genomes in this small data set. This increase is evenly distributed along the genome (Figure S5).

Figure 6 shows the number of recombination event boundaries found for each region of the alignment. The number of recombination events is higher than average in four regions at positions 0.8, 1.8, 3.2, and 4.9 Mbp along the genome of ATCC14579. This result confirms that the variation observed when inferring ρ in step 2 (*cf.* Figure 4) was not just caused by a lack of information. Such hotspots of recombination have previously been described in genomic regions under strong positive selection, for example in Streptococcus (LEFEBURE and STANHOPE 2007; MUZZI *et al.* 2008) and in *Escherichia coli* (MILKMAN *et al.* 2003; TOUCHON *et al.* 2009). Here the

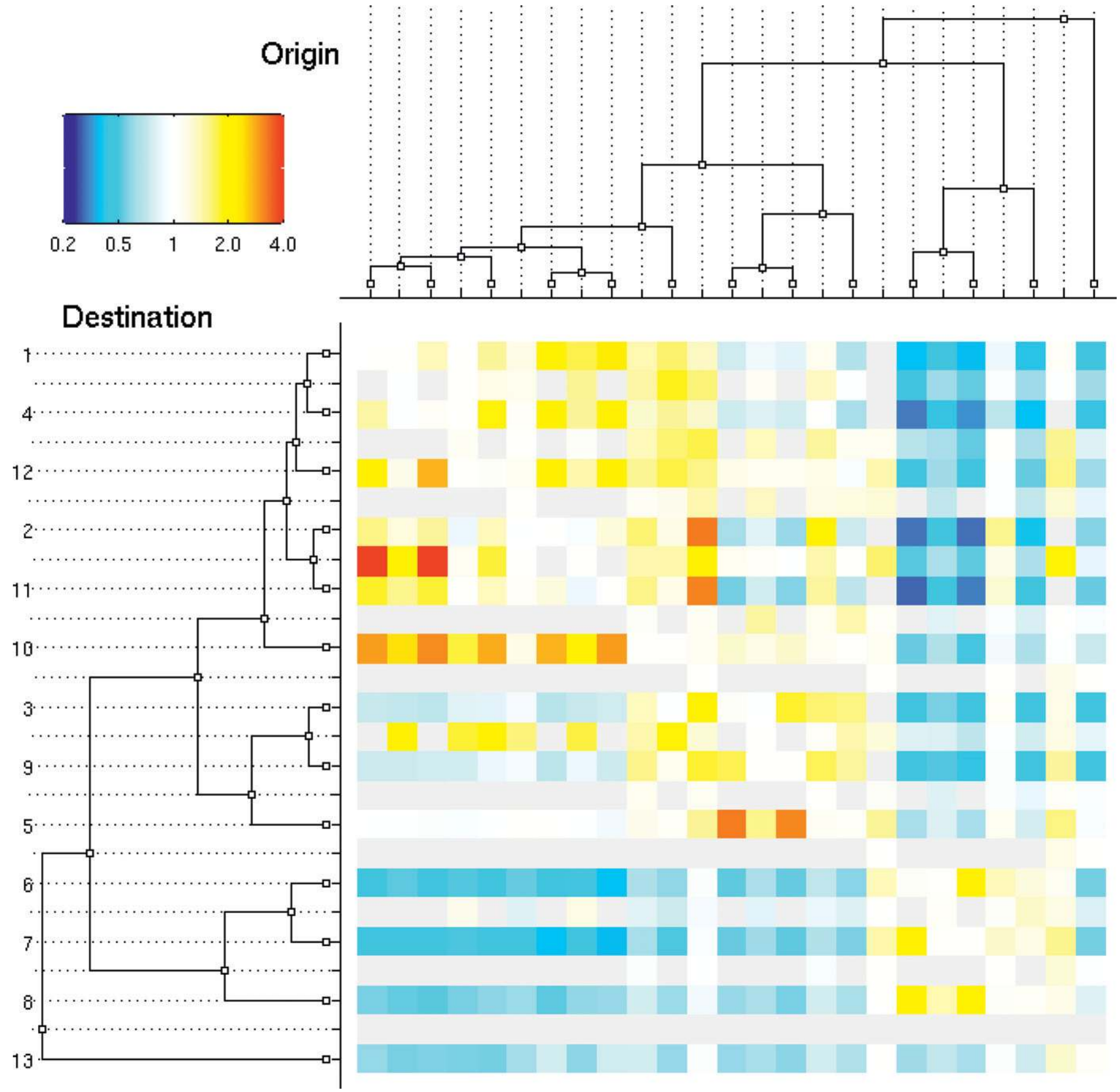

Figure 5.—Heat map for the Bacillus stage 3 analysis showing the number of recombination events inferred relative to its expectation under our prior model given the stage 2 inferred recombination rate, for each donor/recipient pair of branches. The cells in very light gray are the ones for which the ratio would be meaningless because there are less than three observed and expected events.

two peaks at 0.8 and 1.8 Mbp correspond to regions of important change in GC content (Ivanova *et al.* 2003). The peak at 3.2 Mbp contains a large number of genes annotated with antibiotic and other drug resistance (Ivanova *et al.* 2003) that may be under positive selection. The peaks at 0.8 and 4.8 Mbp are also located near *rrn* operons (Ivanova *et al.* 2003) and a tRNA gene array that harbors the integration site for a Bacillus site-specific integrative conjugative element (Grohmann 2010).

Recombination events inferred in specific regions: Figure 7 shows the recombination events found in the first 2000 bp of the two blocks shown by a blue and a green dot (respectively) in Figure 6. The first region is located right at the beginning of the sequence of genome ATCC14579, which corresponds to the origin of replication, where recombination is not particularly prevalent (Figure 6). This region contains the *dnaA* gene and the beginning of the BC0002 gene (Ivanova *et al.* 2003), which both play essential roles in DNA replication. One of

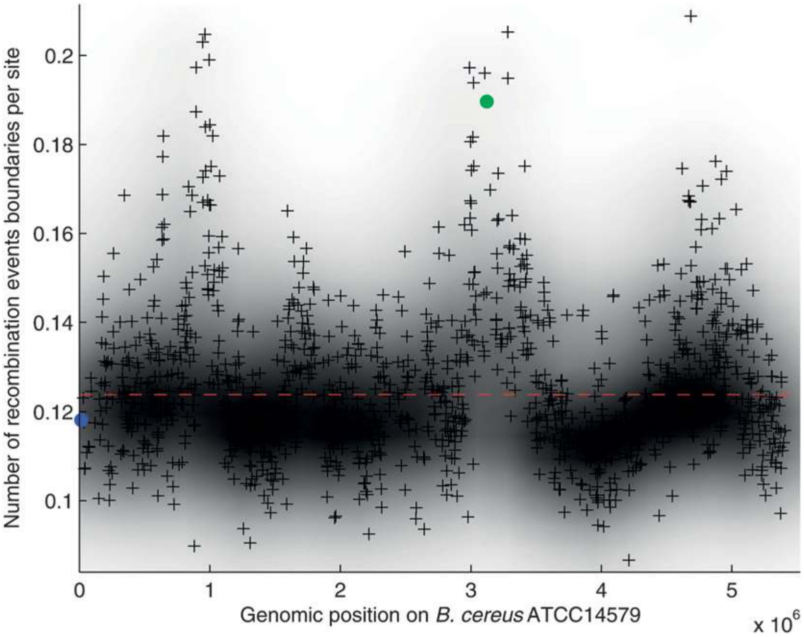

FIGURE 6.—Scatterplot of the stage 2 analysis of Bacillus showing the number of recombination event boundaries per site for each block in the alignment of Bacillus. Details of the two blocks shown by a blue and a green dot are shown in Figure 7.

the clearest recombination events detected in this region is an import into genome 11 from the cluster containing genomes 1, 4, and 12 that spans the entire region shown in Figure 7. In another clear event, genome 10 imported the first ∼1000 bp also from the cluster 1, 4, 12. The ancestor of genomes 2 and 11 imported the second half of the region from an ancestor of clade 1. The first ∼100 bp (*i.e.*, before the start of gene *dnaA*) may have been imported by any of the three genomes of clade 2 from a member of clade 1. With this single (unclear) exception, there have been no interclade events in this region. There are, however, many branches and genomic locations for which no recombination was found at all, for example on the branches above genomes 1, 3, 4, and 9.

The second region shown in Figure 7 is located in the third hotspot of recombination at 3.2 Mbp (Figure 6). This region contains the end of the BC3152 gene that produces an arsenate reductase, the BC3153 gene that produces an arsenic-resistance protein, the BC3154 gene that produces a lactoylglutathione lyase, and the beginning of the BC3155 gene that produces an arsenic-resistance operon repressor (IVANOVA *et al.* 2003). We found many recombination events in this region, and no single branch of the clonal genealogy remains

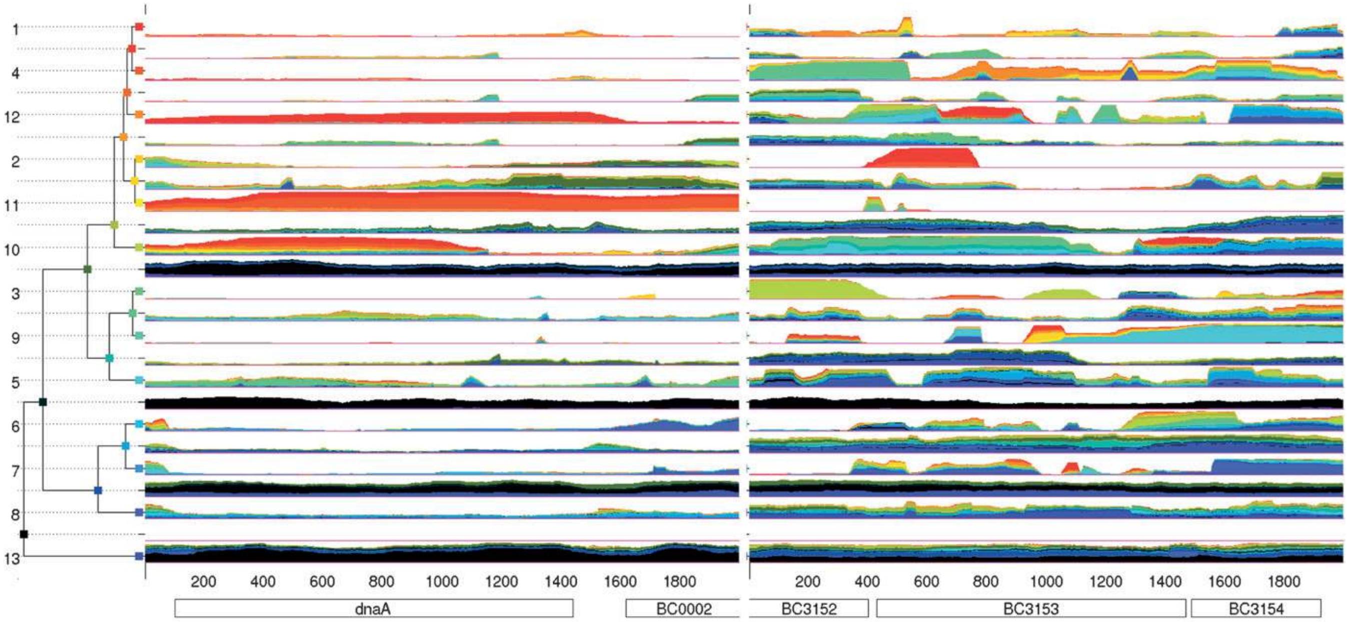

FIGURE 7.—Results of our stage 3 analysis for two example regions of the Bacillus alignment. The representation is the same as in Figure 2. The two regions are shown by a blue and a green dot, respectively, in Figure 6.

unaffected (Figure 7). Some of these events represent interclade recombination, for example the import of the last ~400 bp of genome 12 (which belongs to clade 1) from clade 2 or the import centered on position 1400 bp of genome 6 (which belongs to clade 2) from clade 1. Some events have a very clearly defined origin, for example the import of the first ~600 bp of genome 4 from the ancestor of genomes 3 and 9, whereas others present more uncertainty, for example the import at position 100 bp on genome 5 that could come from at least six branches. This second region contrasts with the first one shown in Figure 7 in several respects: the number of recombination events is higher, their tract lengths are on average shorter, and interclade events are more frequent. Since the genes in this second region are involved in resistance to arsenic and its compounds (which are often used as pesticides, herbicides, or insecticides), these genes are likely to be under positive selective pressure (Petersen *et al.* 2007), which often implies a higher rate of recombination (Lefebure and Stanhope 2007; Muzzi *et al.* 2008; Orsi *et al.* 2008; Touchon *et al.* 2009).

DISCUSSION

Recombination and its consequences have previously been detected and quantified in many different ways. The standard population genetic approach starts with the assumption of a randomly mating population and inference of the rates of mutation and recombination. Information on recombination comes in particular from the patterns of linkage disequilibrium, which have been used to build detailed maps of recombination rates in humans and other eukaryotes (McVean *et al.* 2002, 2004; Myers *et al.* 2005; Winckler *et al.* 2005). This technique has also been applied to bacteria and archaea (Jolley *et al.* 2005; Wirth *et al.* 2007; Tanabe *et al.* 2009; Touchon *et al.* 2009).

In bacteria and archaea the standard population genetic framework is problematic because of the absence of a mating pool with defined boundaries or homogeneous rates of exchange (Didelot and Maiden 2010). Recombination occurs in a clonal context, due to the asymmetry of the contributions of donor and recipient cells. These features of prokaryote reproductive biology have motivated us to develop specialized methods of inference. Building on Didelot and Falush (2007), we discuss a new method, ClonalOrigin, to infer recombination from an alignment of whole bacterial genome sequences. We presented an application to 13 genomes of *B. cereus*, which revealed interesting variation in recombination rates both across lineages (Figure 5) and across the genome (Figure 6).

The size of genomic data sets means that statistical inference can be computationally challenging. ClonalOrigin is based on a model under which recombination patterns in different alignment blocks can be analyzed independently, facilitating parallelization of most calculations. Inference is decomposed into a three-step process that infers successively the clonal genealogy, global parameters, and recombination events. For the first step we used ClonalFrame (Didelot and Falush 2007), which is not strictly statistically correct since it is based on a different model. There is, however, typically little ambiguity about the clonal genealogy when working with whole genomes, as shown here by the similarity between the clonal genealogy reconstructed by ClonalFrame and the results of a variety of simpler phylogenetic algorithms (Figure S4). Furthermore, small differences in clonal genealogies should not affect the results of the second and third step significantly. For the second step we used the median of the global parameters found by each alignment block when unconditioned, which once again is not strictly correct but likely to be close to the truth given the large amount of genomic sequence considered (>3 Mbp) and the relative stability in their inferred values across blocks (Figure 4). Step 3 introduces no further approximation.

ClonalOrigin follows the standard population genetic approach in estimating values of θ and ρ, which are assumed to be constant across the genome. However, the inferred number and size of events on specific branches of the clonal genealogy may differ substantially from the expectation given by θ and ρ. Our model assumes that recombination events are distributed as if the isolates in the sample were uniformly drawn from a randomly mating population (Equation 4), but the inferred events can follow a very different pattern, which provides considerable information on the diverse biological processes influencing recombination rates. Furthermore, although we do not attempt it here, it is possible in principle to infer the DNA substitutions introduced by each recombination event and hence to study its biological consequences, for example in facilitating the spread of beneficial alleles.

Instead of looking at the output for patterns of deviation from prior expectation, it would be more statistically powerful to account for such possibilities in our model. Since our model is based on the coalescent (Kingman 1982), it can easily be extended to account for a number of additional biological processes, such as population dynamics (Griffiths and Tavare 1994) or population structure (Nielsen and Wakeley 2001; Wilson *et al.* 2003). Such extensions would introduce new parameters that would appear in both the prior for the clonal genealogy (Equation 1) and the prior for recombinant edges (Equation 7) and would therefore make the decomposition into three inference steps problematic. An interesting alternative would be to leave inference as it is and investigate extensions of the model in a postprocessing step using importance sampling (Meligkotsidou and Fearnhead 2007).

The model we described assumed a constant rate of mutation and recombination along the genome. This assumption held approximately in our example but may not always be appropriate. The model could therefore be extended, for example using a changepoint process for the mutation and recombination rates along the genome, as used for example in LDhat (McVean *et al.* 2004) or DualBrothers (Minin *et al.* 2005). This would be computationally feasible under the current three-step process by fitting a changepoint model to the rates instead of taking the median for the third step.

A key difference of the model underlying ClonalOrigin in comparison with our previous effort ClonalFrame (Didelot and Falush 2007) is that the origin of recombination events is explicitly modeled. We showed that this difference makes ClonalOrigin more accurate than ClonalFrame (Figures 2 and 3) when detecting recombination from the ancestral recombination graph model (Wiuf and Hein 2000; Didelot *et al.* 2009c). Furthermore, it allows a quantification of the genetic flux between lineages (Figure 5) that would not be possible otherwise. Although ClonalOrigin can still detect external imports, ClonalFrame is a more appropriate model when most DNA imports are from an external source into the sampled population. In such a scenario, the attempts made by ClonalOrigin at inferring the origin of the imports may be detrimental to the detection of these recombination events compared to ClonalFrame, which makes no such attempt (Didelot and Falush 2007). A cross between the two models could therefore be envisaged, where some events would have an origin as defined by ClonalOrigin and others would introduce novel polymorphism as in ClonalFrame.

Mark Achtman, Sylvain Brisse, Paul Fearnhead, Peter Green, Eduardo Rocha, and three anonymous reviewers provided useful comments, ideas, and discussion. This work was funded in part by Wellcome Trust grant WT082930MA. A. Darling was supported by National Science Foundation grant DBI-0630765. D. Falush was supported by Science Foundation of Ireland grant 05/FE1/B882.

LITERATURE CITED

Achtman, M., T. Azuma, D. E. Berg, Y. Ito, G. Morelli *et al.*, 1999 Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. Mol. Microbiol. **32:** 459–470.

Canchaya, C., G. Fournous, S. Chibani-Chennoufi, M. L. Dillmann and H. Brüssow, 2003 Phage as agents of lateral gene transfer. Curr. Opin. Microbiol. **6:** 417–424.

Challacombe, J. F., M. R. Altherr, G. Xie, S. S. Bhotika, N. Brown *et al.*, 2007 The complete genome sequence of *Bacillus thuringiensis* Al Hakam. J. Bacteriol. **189:** 3680–3681.

Chen, I., P. J. Christie and D. Dubnau, 2005 The ins and outs of DNA transfer in bacteria. Science **310:** 1456–1460.

Claverys, J. P., B. Martin and P. Polard, 2009 The genetic transformation machinery: composition, localization, and mechanism. FEMS Microbiol. Rev. **33:** 643–656.

Darling, A. C., B. Mau, F. R. Blattner and N. T. Perna, 2004 Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. **14:** 1394–1403.

Darling, A., B. Mau and N. Perna, 2010 ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One **5:** e11147.

Didelot, X., and D. Falush, 2007 Inference of bacterial microevolution using multilocus sequence data. Genetics **175:** 1251–1266.

Didelot, X., and M. C. Maiden, 2010 Impact of recombination on bacterial evolution. Trends Microbiol. **18:** 315–322.

Didelot, X., M. Barker, D. Falush and F. Priest, 2009a Evolution of pathogenicity in the *Bacillus cereus* group. Syst. Appl. Microbiol. **32:** 81–90.

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

Didelot, X., D. Lawson and D. Falush, 2009c SimMLST: simulation of multi-locus sequence typing data under a neutral model. Bioinformatics **25:** 1442–1444.

Didelot, X., R. Urwin, M. C. J. Maiden and D. Falush, 2009d Genealogical typing of *Neisseria meningitidis*. Microbiology **155:** 3176–3186.

Falush, D., M. Torpdahl, X. Didelot, D. F. Conrad, D. J. Wilson *et al.*, 2006 Mismatch induced speciation in *Salmonella*: model and data. Philos. Trans. R. Soc. Lond. B Biol. Sci. **361:** 2045–2053.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

Green, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711–732.

Griffiths, R., and S. Tavare, 1994 Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. B Biol. Sci. **344:** 403–410.

Grohmann, E., 2010 Conjugative transfer of the integrative and conjugative element ICEBs1 from *Bacillus subtilis* likely initiates at the donor cell pole. J. Bacteriol. **192:** 23–25.

Guttman, D., 1997 Recombination and clonality in natural populations of *Escherichia coli*. Trends Ecol. Evol. **12:** 16–22.

Han, C. S., G. Xie, J. F. Challacombe, M. R. Altherr, S. S. Bhotika *et al.*, 2006 Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* isolates closely related to *Bacillus anthracis*. J. Bacteriol. **188:** 3382–3390.

Hill, W., and A. R. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 183–201.

Ivanova, N., A. Sorokin, I. Anderson, N. Galleron, B. Candelon *et al.*, 2003 Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. Nature **423:** 87–91.

Jolley, K. A., D. J. Wilson, P. Kriz, G. Mcvean and M. C. J. Maiden, 2005 The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. Mol. Biol. Evol. **22:** 562–569.

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

Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Larsson, P., D. Elfsmark, K. Svensson, P. Wikstrm, M. Forsman *et al.*, 2009 Molecular evolutionary consequences of niche restriction in *Francisella tularensis*, a facultative intracellular pathogen. PLoS Pathog. **5:** e1000472.

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

Maiden, M., 2006 Multilocus sequence typing of bacteria. Annu. Rev. Microbiol. **60:** 561–588.

Majewski, J., 2001 Sexual isolation in bacteria. FEMS Microbiol. Lett. **199:** 161–169.

Marjoram, P., and J. D. Wall, 2006 Fast "coalescent" simulation. BMC Genet. **7:** 16.

Mau, B., and M. Newton, 1997 Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. J. Comput. Graph. Stat. **6:** 122–131.

Maynard Smith, J., and N. H. Smith, 1998 Detecting recombination from gene trees. Mol. Biol. Evol. **15:** 590–599.

McVean, G., and N. Cardin, 2005 Approximating the coalescent with recombination. Philos. Trans. R. Soc. B **360:** 1387–1393.

McVean, G., P. Awadalla and P. Fearnhead, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:** 1231–1241.

McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581–584.

Meligkotsidou, L., and P. Fearnhead, 2007 Postprocessing of genealogical trees. Genetics **177:** 347–358.

Milkman, R., and M. M. Bridges, 1990 Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics **126:** 505–517.

Milkman, R., E. Jaeger and R. D. McBride, 2003 Molecular evolution of the *Escherichia coli* chromosome. VI. Two regions of high effective recombination. Genetics **163:** 475–483.

Minin, V. N., K. S. Dorman, F. Fang and M. A. Suchard, 2005 Dual multiple change-point model leads to more accurate recombination detection. Bioinformatics **21:** 3034–3042.

Muzzi, A., M. Moschioni, A. Covacci, R. Rappuoli and C. Donati, 2008 Pilus operon evolution in *Streptococcus pneumoniae* is driven by positive selection and recombination. PLoS One **3:** e3660.

Myers, S., L. Bottolo, C. Freeman, G. McVean and P. Donnelly, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310:** 321–324.

Nielsen, R., and J. Wakeley, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics **158:** 885–896.

Orsi, R., Q. Sun and M. Wiedmann, 2008 Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of *Listeria monocytogenes*. BMC Evol. Biol. **8:** 233.

Petersen, L., J. P. Bollback, M. Dimmic, M. Hubisz and R. Nielsen, 2007 Genes under positive selection in *Escherichia coli*. Genome Res. **17:** 1336–1343.

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

Rasko, D. A., J. Ravel, O. A. Økstad, E. Helgason, R. Z. Cer *et al.*, 2004 The genome sequence of *Bacillus cereus* ATCC 10987 reveals metabolic adaptations and a large plasmid related to *Bacillus anthracis* pXO1. Nucleic Acids Res. **32:** 977–988.

Ravel, J., L. Jiang, S. T. Stanley, M. R. Wilson, R. S. Decker *et al.*, 2009 The complete genome sequence of *Bacillus anthracis* Ames "Ancestor". J. Bacteriol. **191:** 445–446.

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

Sorokin, A., B. Candelon, K. Guilloux, N. Galleron, N. Wackerow-Kouzova *et al.*, 2006 Multiple-locus sequence typing analysis of *Bacillus cereus* and *Bacillus thuringiensis* reveals separate clustering and a distinct population structure of psychrotrophic strains. Appl. Environ. Microbiol. **72:** 1569–1578.

Stenfors Arnesen, L. P., A. Fagerlund and P. E. Granum, 2008 From soil to gut: *Bacillus cereus* and its food poisoning toxins. FEMS Microbiol. Rev. **32:** 579–606.

Stumpf, M., and G. McVean, 2003 Estimating recombination rates from population-genetic data. Nat. Rev. Genet. **4:** 959–968.

Tanabe, Y., T. Sano, F. Kasai and M. Watanabe, 2009 Recombination, cryptic clades and neutral molecular divergence of the microcystin synthetase (mcy) genes of toxic cyanobacterium *Microcystis aeruginosa*. BMC Evol. Biol. **9:** 115.

Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl *et al.*, 2009 Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. **5:** e1000344.

Vos, M., and X. Didelot, 2009 A comparison of homologous recombination rates in bacteria and archaea. ISME J. **3:** 199–208.

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

Whelan, S., P. Li and N. Goldman, 2001 Molecular phylogenetics: state-of-the art methods for looking into the past. Trends Genet. **17:** 262–272.

Wilson, I., M. Weale and D. Balding, 2003 Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J. R. Stat. Soc. Ser. A (Stat. Soc.) **166:** 155–201.

Winckler, W., S. R. Myers, D. J. Richter, R. C. Onofrio, G. J. McDonald *et al.*, 2005 Comparison of fine-scale recombination rates in humans and chimpanzees. Science **308:** 107–111.

Wirth, T., G. Morelli, B. Kusecek, A. van Belkum, C. van der Schee *et al.*, 2007 The rise and spread of a new pathogen: seroresistant *Moraxella catarrhalis*. Genome Res. **17:** 1647–1656.

Wiuf, C., and J. Hein, 2000 The coalescent with gene conversion. Genetics **155:** 451–462.

Xiong, Z., Y. Jiang, D. Qi, H. Lu, F. Yang *et al.*, 2009 Complete genome sequence of the extremophilic *Bacillus cereus* strain Q1 with industrial applications. J. Bacteriol. **191:** 1120–1121.

Yang, Z., and B. Rannala, 1997 Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. **14:** 717–724.

Communicating editor: Y. S. Song

APPENDIX A: DETAILS OF THE MCMC MOVES

Reversible-jump moves: We use two reversible-jump moves: a "remove" move that proposes removal of an existing recombination event chosen uniformly at random and an "add" move that proposes adding a recombination event with properties a^*, b^*, x^*, y^* proposed according to their priors as described in Equations 3–6.

These two moves are accepted according to the Metropolis–Hastings–Green ratio

$$\alpha = \min\left(1, \frac{\mathbf{P}(\mathcal{D}\,|\,\mathcal{R}')}{\mathbf{P}(\mathcal{D}\,|\,\mathcal{R})}\frac{\mathbf{P}(\mathcal{R}')}{\mathbf{P}(\mathcal{R})}\frac{Q(\mathcal{R}'\rightarrow\mathcal{R})}{Q(\mathcal{R}\rightarrow\mathcal{R}')}\ \text{Jacobian}\right), \tag{A1}$$

where $\mathcal{R}$ is the old value of the parameter and $\mathcal{R}'$ is the proposed value. The Jacobian is equal to one because no transformation of parameter is being done. The first term of the product is the ratio of likelihoods that is calculated using Equation 9. The second term is the ratio of priors that is calculated using Equation 7. This leaves only to calculate the third term, *i.e.*, the ratio of proposal distributions. For the add move we have (*cf.* Equation 7)

$$Q(\mathcal{R}'\rightarrow\mathcal{R}) = \frac{1}{R+1} \quad \text{and} \quad Q(\mathcal{R}\rightarrow\mathcal{R}') = \frac{\mathbf{P}(x^*)\mathbf{P}(y^*\,|\,x^*)\exp(-L(a^*,\ b^*,))}{T} \tag{A2}$$

so that the acceptance ratio is

$$\alpha = \min\left(1,\ \frac{\mathbf{P}(\mathcal{D}\,|\,\mathcal{R}')}{\mathbf{P}(\mathcal{D}\,|\,\mathcal{R})}\frac{\rho T}{2(R+1)}\right). \tag{A3}$$

For the remove move, if we note the recombinant edge proposed to be removed with a^*, we have

$$Q(\mathcal{R}' \rightarrow \mathcal{R}) = \mathbf{P}(x^*)\mathbf{P}(y^*\,|\,x^*)\frac{\exp(-L(a^*, b^*))}{T} \quad \text{and} \quad Q(\mathcal{R} \rightarrow \mathcal{R}') = \frac{1}{R} \tag{A4}$$

so that the acceptance ratio is

$$\alpha = \min\left(1,\ \frac{\mathbf{P}(\mathcal{D}\,|\,\mathcal{R}')}{\mathbf{P}(\mathcal{D}\,|\,\mathcal{R})}\frac{2R}{\rho T}\right). \tag{A5}$$

Update of the start and end points of a recombination event: We describe the move we used to propose increasing the starting site x of a given recombination event. Equivalent moves were also used to propose decreasing x and to propose increasing or decreasing the ending site y. We update x by adding to it an amount $d \in [0, D]$ (where D is a fixed quantity; we used $D = 10$). The move can thus be written $x \rightarrow x + d$. d is proposed according to its relative posterior probability in $[0, D]$; *i.e.*,

$$Q(x \rightarrow x + d) = \frac{\mathbf{P}(\mathcal{D}\,|\,x + d)\mathbf{P}(x + d)}{\sum_{j=0}^{D}\mathbf{P}(\mathcal{D}\,|\,x + j)\mathbf{P}(x + j)}. \tag{A6}$$

The terms $\mathbf{P}(x + d)$ follow from Equations 5 and 6. The terms $\mathbf{P}(D\,|\,x + d)$ are calculated as follows. If the likelihood at site s is L_s with the recombinant edge and L'_s without it, the likelihood for moving $x \rightarrow x + d$ is $\mathbf{P}(\mathcal{D}\,|\,x + d) = \mathbf{P}(\mathcal{D}\,|\,x) \times \prod_{k=1}^{d}(L'_{x+k}/L_{x+k})$. We thus calculate L_{x+K} and L'_{x+K} for all $k \in [0, D]$ and all values of the terms $\mathbf{P}(D\,|\,x + d)$ follow.

We therefore sample from the proposal distribution $Q(x \rightarrow x + d)$ and accept with probability

$$\alpha = \min\left(1,\ \frac{\sum_{j=0}^{D}\mathbf{P}(\mathcal{D}\,|\,x + j)(\mathbf{P}(x + j))}{\sum_{j'=-D}^{0}\mathbf{P}(\mathcal{D}\,|\,x + d - j')\mathbf{P}(x + d - j')}\right). \tag{A7}$$

Update of departing and arrival points of a recombination event: We propose a new value of the departing or arrival point of a recombination event by adding a perturbation $\varepsilon \sim \text{Normal}(0, 0.01)$ to its age. If the age is decreased, at each bifurcation crossed one of the two daughter branches is followed so that the probability of choosing a given branch at the new time is 2^n, where n is the number of bifurcations between the old and the new point. If the age is increased, then let $-n$ denote the number of bifurcations of the clonal genealogy between the old and the new point. The Metropolis–Hastings acceptance probability of this move is therefore

$$\alpha = \min\left(1,\ 2^n\ \frac{\mathbf{P}(\mathcal{D}\,|\,\mathcal{R}')}{\mathbf{P}(\mathcal{D}\,|\,\mathcal{R})}\frac{\mathbf{P}(\mathcal{R}')}{\mathbf{P}(\mathcal{R})}\right), \tag{A8}$$

where the ratio of the prior can be calculated using Equation 7 and the ratio of likelihoods can be calculated by applying Equation 9 only for the sites affected by the recombination event.

APPENDIX B: ADDITIONAL MOVES FOR THE PARAMETERS θ, ρ, AND δ

Here we describe the additional moves of the MCMC required when θ, ρ, and δ are treated as additional parameters as is required in step 2 of the analysis.

Update of θ: We used an improper Uniform prior on $[0, \infty)$ for the mutation rate θ and updated its value by proposing the addition of a perturbation ε drawn from Uniform($[-5; 5]$). Since this proposal is symmetric and the prior is uniform, the Metropolis–Hastings acceptance ratio for this move reduces to a ratio of likelihoods that can be computed using Equation 9.

Update of ρ: We use a Gamma(α, β) prior for the recombination rate ρ. This has the advantage to be conjugate with the distribution of the number of recombination events R given ρ, which is Poisson$(\rho T/2)$. Thus we can deduce that the posterior distribution of ρ is Gamma$(R + \alpha, (1/\beta + T/2)^{-1})$. We update ρ by proposing from this distribution, which is a Gibbs move. In the examples shown we used an improper Uniform prior on $[0, \infty)$ for ρ that is obtained by taking $\alpha = 1$ and $\beta = \infty$ and thus the posterior distribution becomes Gamma$(R + 1, 2/T)$.

Update of δ: Given Equations 5 and 6, the likelihood of the mean length of imports δ is of the form

$$L(\delta) = \frac{\delta^X (1 - 1/\delta)^Y}{(b\delta + L - b)^R}, \tag{B1}$$

where R is the number of recombinant edges, $Y = \left(\sum_{i=1}^{R} y_i - x_i\right)$ plus the number of import ends falling on block ends, and X is the number of import starts falling on block starts, minus the number of import ends falling before block ends. We assumed an improper Uniform prior on $(0, \infty)$, so that the expression above is the posterior distribution for δ. The update can thus be done by proposing, adding a small perturbation ε drawn from Uniform$([-5; 5])$ and accepting according to the ratio $L(\delta')/L(\delta)$.

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.120121/DC1

Inference of Homologous Recombination in Bacteria Using Whole-Genome Sequences

Xavier Didelot, Daniel Lawson, Aaron Darling and Daniel Falush

DOI: 10.1534/genetics.110.120121

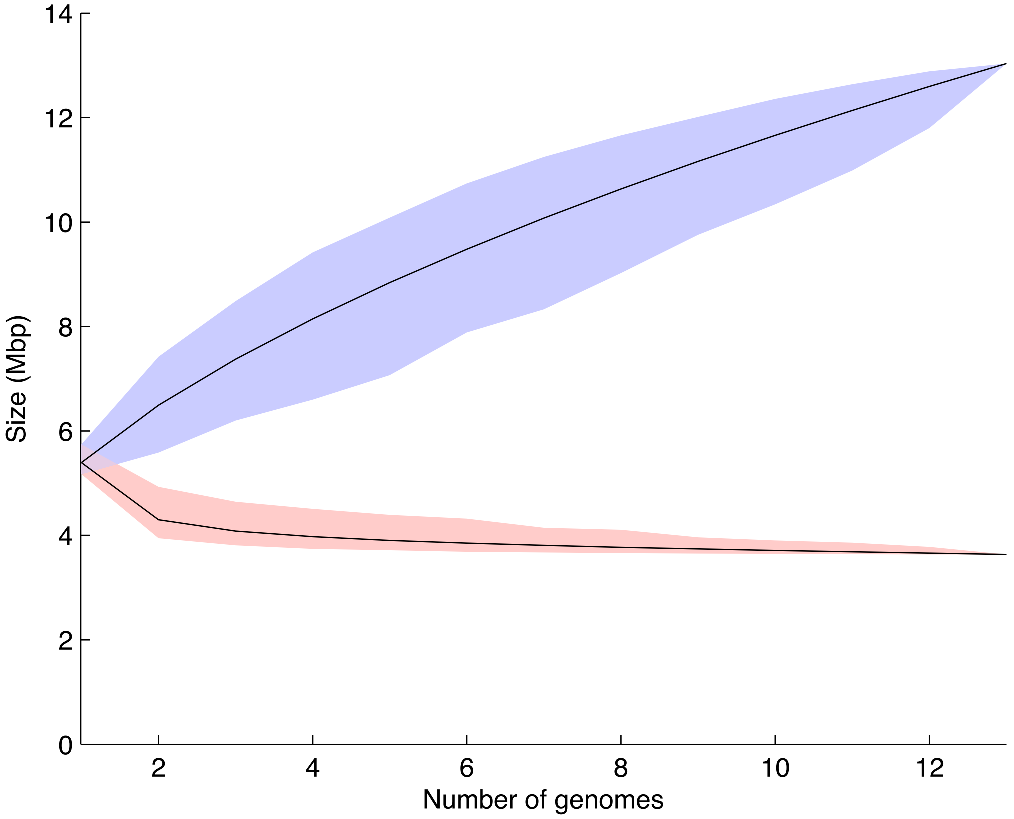

FIGURE S1.—Extent of the genomic regions found in all genomes of Bacillus (blue), or at least one genome (red), as a function of the number of genomes being aligned.

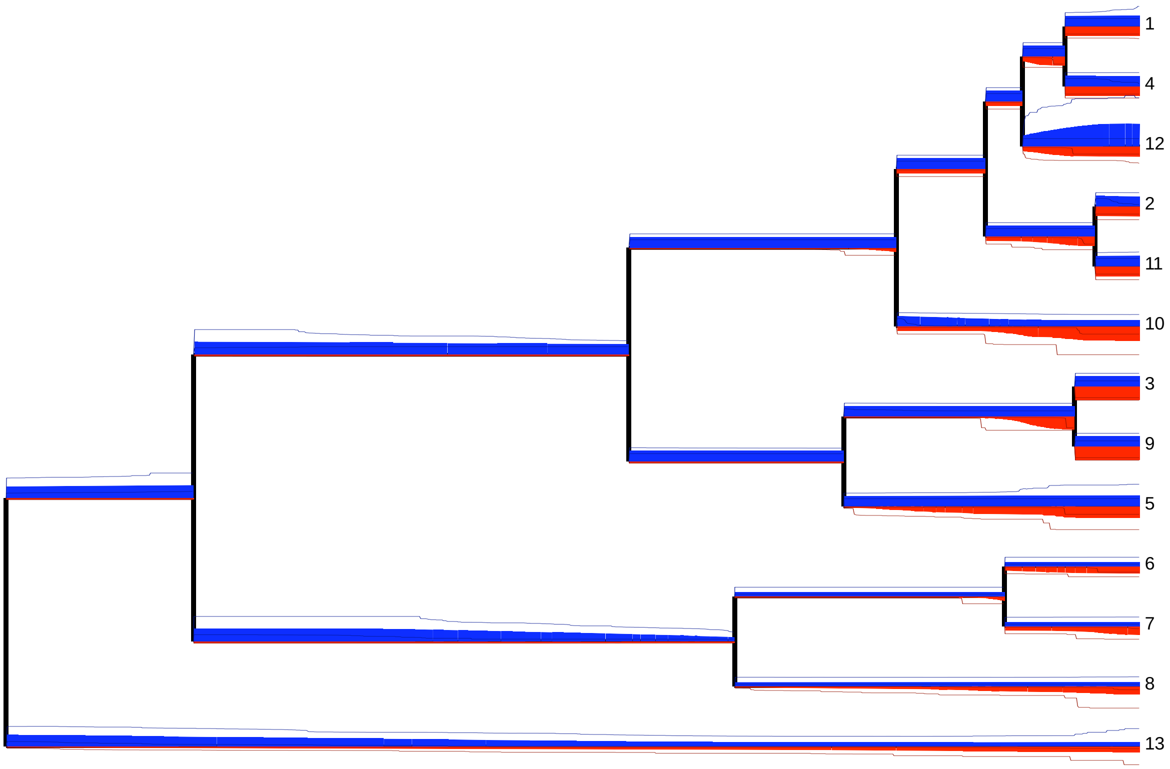

FIGURE S2.—Result of GenoPlast. The quantity of solid red and blue on the tree is proportional to the rates of gain and loss of material, with the 95% credibility interval for each rate shown by two lines of the corresponding color.

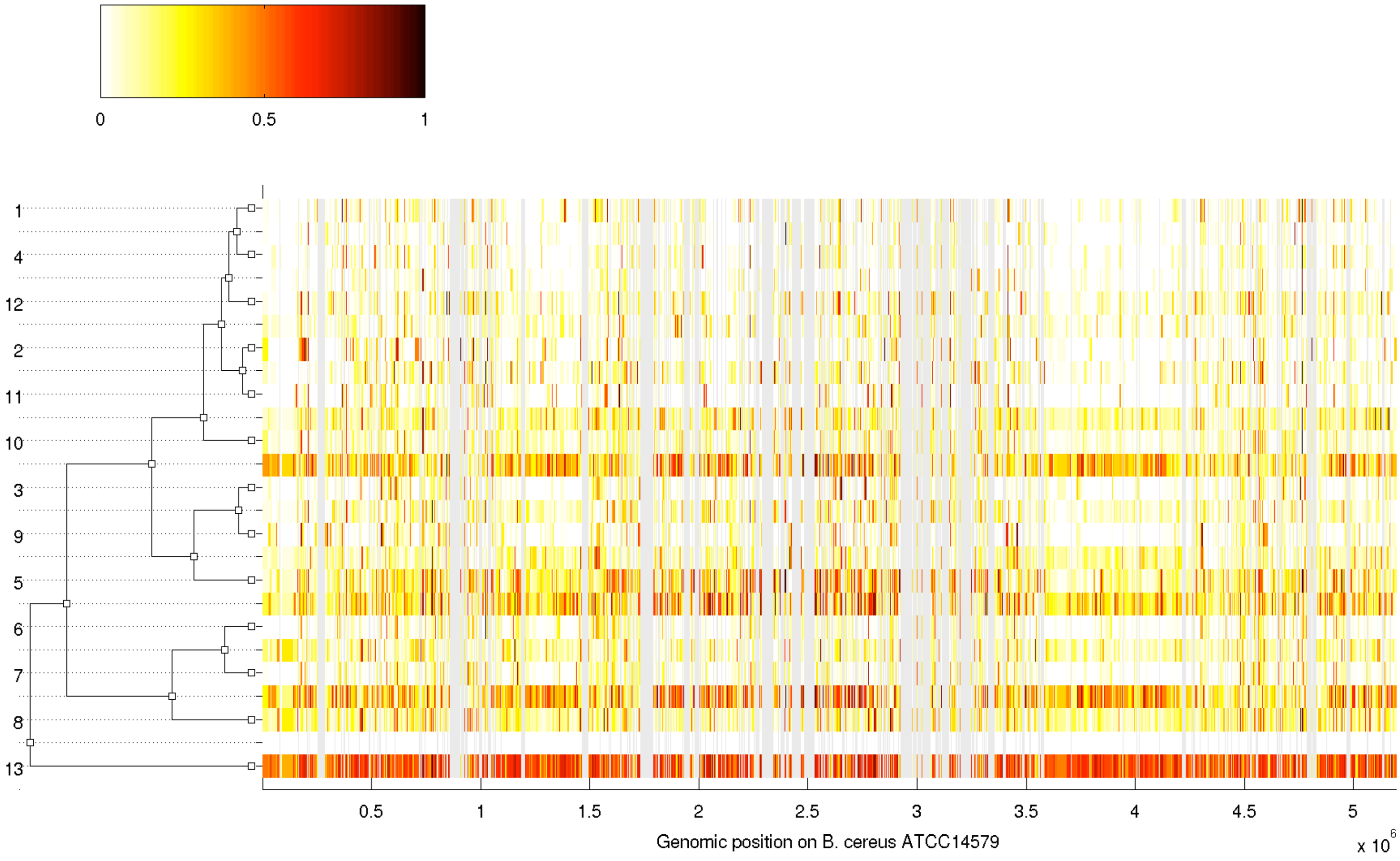

FIGURE S3.—Result of ClonalFrame. The clonal genealogy is shown on the left hand side, and the recombination events affecting each of its branches on the right hand side. The intensity of color is proportional to the probablility of recombination at a given position, and regions that are not in the alignment are shown in light gray.

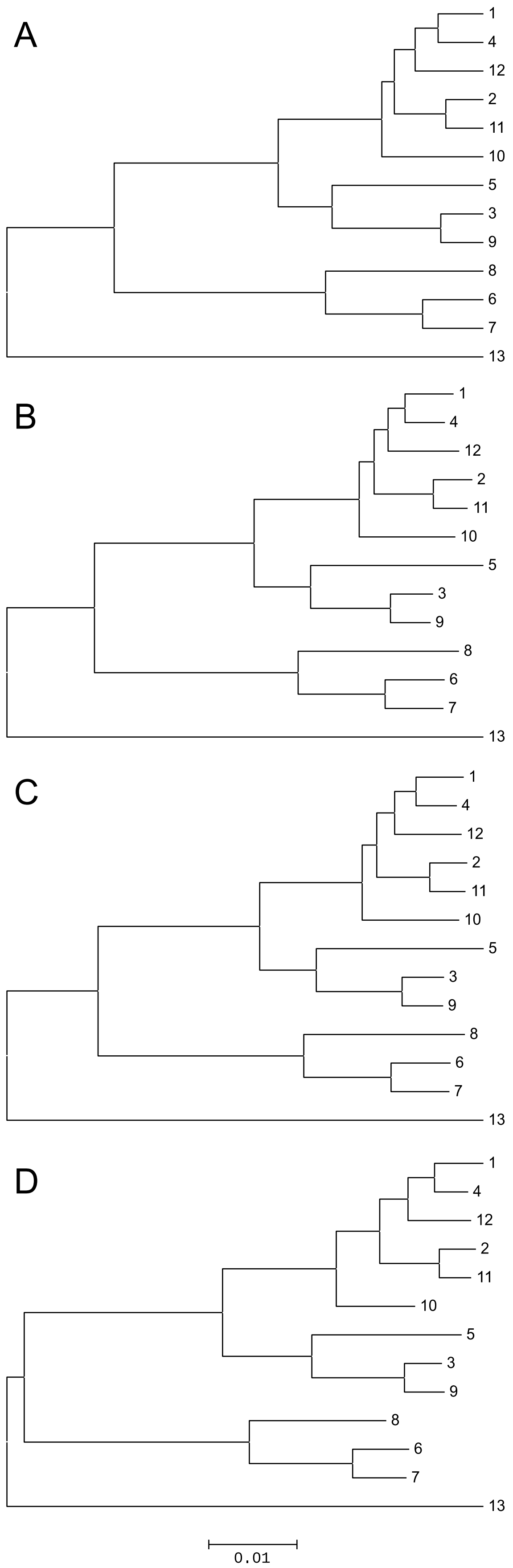

FIGURE S4.—Genealogies reconstructed from the Bacillus alignment using (A) UP-GMA, (B) Neighbor-Joining, (C) Minimum evolution and (D) Maximum parsimony.

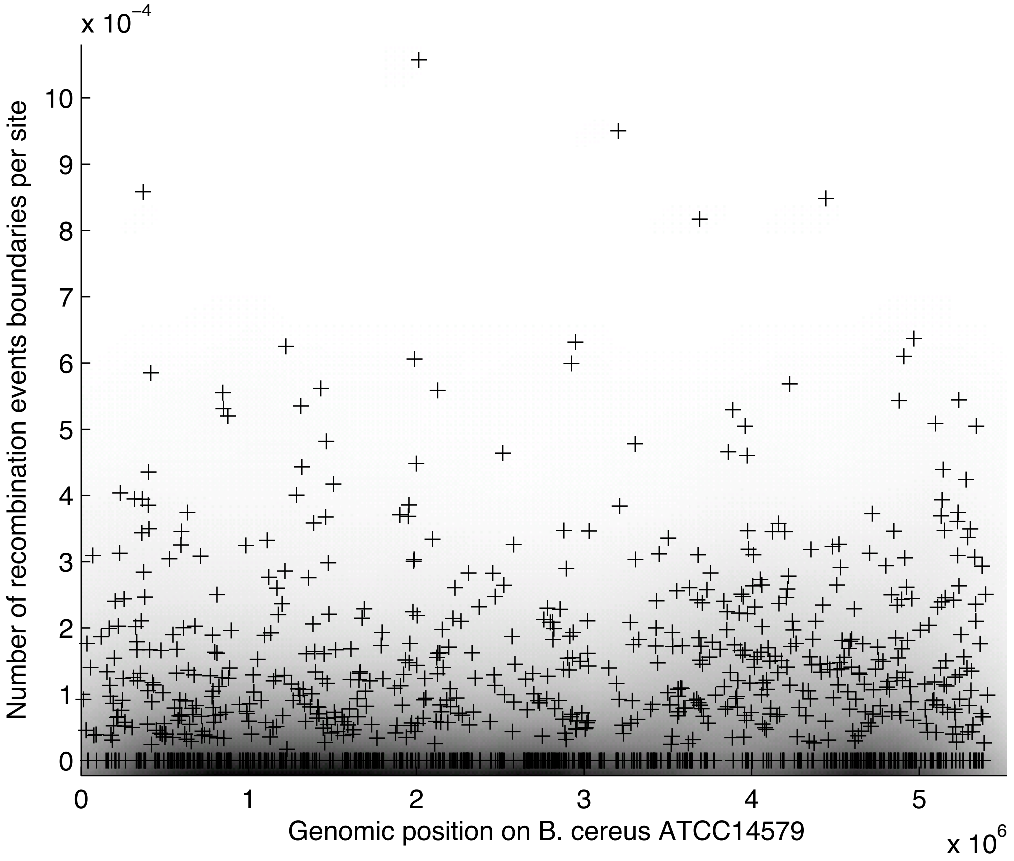

FIGURE S5.—Scatterplot of the number of recombination events boundaries per site for each block in the alignment of Bacillus when looking only at events from the branches above genomes 1 or 4 to the branch above the ancestor of 2 and 11.