Inference of homologous recombination in bacteria using whole genome sequences

Xavier Didelot¹,*, Daniel Lawson², Aaron Darling³, Daniel Falush⁴

1 Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
2 Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK
3 Genome Center, University of California-Davis, 451 Health Sciences Dr., Davis, CA, 95616, USA
4 Environmental Research Institute, University College Cork, Ireland

* Email: X.Didelot@warwick.ac.uk
Abstract

Bacteria and archaea reproduce clonally, but horizontal exchange of genetic material is known to occur in some organisms. In the majority of events, the imported DNA replaces an homologous segment in the recipient genome. We introduce a mathematical model that describes both the recipient and the origin of homologous DNA imports in terms of ancestors to the sequences 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 alignment through parallelization. Using simulated data, we demonstrate accurate and reliable identification of individual recombination events and recombination rate parameters. Our approach is applied to an alignment of 13 whole genomes from the Bacillus cereus group. We find, as expected from laboratory experiments, that recombination rate is higher between closely related organisms and also that the genome contains several broad regions of elevated levels of recombination. The methods described in this article have been implemented in a computer software package, ClonalOrigin, which is freely available. Application of the method to the genomic datasets that are becoming available should reveal the evolutionary history and private lives of populations of bacteria and archaea.
Author Summary

Bacteria, unlike us, reproduce on their own. They, do however have mechanisms that transfer DNA between organisms, a process more formally known as recombination. The mechanisms by which recombination takes place have been studied extensively in the laboratory but much remains to be understood concerning how, when and where recombination takes place within natural populations of bacteria and how it helps them to adapt to new environments. We have developed a new computer software called ClonalOrigin which analyses the sequences of a number of bacterial genomes in order to reconstruct the recombination events that have taken place in their ancestry. Applying the method to 13 complete genomes from a group of Bacillus bacteria, we identify a quarter of a million of recombination events in this data. We find there are large regions of the genome that have elevated recombination rates - “hot regions” and identify lineages that have atypical patterns of import. Application of the method to the genomic datasets that are becoming available should reveal the private life of bacteria and archaea - the most numerous and interesting organisms on the planet.
**Introduction**

Bacteria and their distant relatives the archaea make up the majority of living organisms. Short generation times combine with enormous population sizes to create tremendous evolutionary potential. Despite their ubiquity, however, these tiny organisms lead their lives largely in private. It is currently not feasible to track individual organisms in natural conditions which makes it difficult to observe the environmental perturbations experienced by particular lineages.

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 [1]. Changes in genomic DNA can accumulate because replication is unfaithful or due to DNA damage, but might also be introduced by recombining a homologous segment of foreign DNA into the chromosome. Three main mechanisms of recombination-mediated lateral gene transfer exist in bacteria and archaea: transduction, conjugation, and transformation. The transduction process transfers DNA via phage infection in bacteria (and by viruses in archaea) [2]. Conjugation encompasses the basic features of sexual processes, with specialized DNA transmission mechanisms and mating types allowing controlled DNA exchange [3]. Transformation of DNA from the environment relies on mechanisms for DNA uptake, homology recognition, and repair and is additionally regulated by mechanisms such as selective digestion of DNA by restriction modification systems [4]. 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 [5].

Recombination in bacteria is analogous to gene-conversion rather than crossing-over in sexually reproducing organisms [6], in the sense that the recipient and donor cells make asymmetric contributions to the genetic make-up 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 sample of bacteria, it is thus possible to define its clonal genealogy [7] 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 DNA that has been inherited by strict vertical descent evolved according to this clonal tree (this is the so-called clonal frame [8]). 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 will 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 Multi-Locus Sequence Typing studies (MLST; [9]) where the phylogenies reconstructed at the various loci can be very different from one another, for example in Helicobacter [10], Bacillus [11] or Salmonella [12].

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 [13]. ClonalFrame has proved useful to identify interesting patterns of recombination in a wide variety of organisms including Campylobacter [14], Neisseria [15] and Francisella [16].

In order to perform efficient inference, ClonalFrame does not model the source of specific recombination events [13]. However, that approach has two important drawbacks. Firstly ClonalFrame can only identify recombination events that introduced a number of substi-
tutions higher than expected through mutation alone, and will miss events that introduce fewer changes [13]. There are also other signals of recombination, most importantly homoplasy which occurs when segregating nucleotides at pairs of sites are not consistent with a single tree [17]. That signal could be correctly interpreted if the source of recombination events was modelled. Secondly, because ClonalFrame does not provide any information about the source of the recombination events it identifies, it can not be used to infer patterns of gene flow between groups of bacteria. One solution is to postprocess the output of ClonalFrame by giving each recombination event a likely origin [18], but that would not be 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 modelled as a point on the clonal genealogy. We show how this model outperforms ClonalFrame in detecting recombination events in simulated datasets. We illustrate the use of our new model on a dataset containing 12 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/.

**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 meaning of the mathematical symbols used in the description below are summarized in Table 1. The model is motivated as a simplification of the Ancestral Recombination Graph (ARG) with gene conversion [19,20], because in the limit of a low recombination
rate, recombined regions of DNA will not intersect and therefore will connect points on the clonal genealogy. Our model corresponds to an ARG in which non-clonal lines of ancestry are not allowed to recombine or coalesce with each other. The model does permit material to persist in hidden lineages prior to recombining back into the clonal genealogy, see for example the red lineage in Figure 1A. Inference under the full ARG process is not feasible for datasets of non-trivial size, but our simplification implies that each recombination edge can be treated as independent which greatly simplifies inference.

The tree $T$ represents the clonal genealogy of the sample of $N$ bacteria under study. We assume a coalescent prior for $T$ [21], which means that if $t_1, ..., t_N$ denote the length of time during which the sample has $2, ..., N$ ancestors respectively, then the probability of the entire genealogy is given by:

$$P(T) = \prod_{i=2}^{N} \exp \left( -\frac{i}{2} t_i \right)$$

(1)

The tree $T$ 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 affecting on the clonal genealogy is:

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

(2)

Each of the $i = 1, ..., R$ 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
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 $a_i$ occurring earlier than $b_i$ (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.:

$$\forall i \in [1..R], P(a_i|T) = \frac{1}{T}$$  \hspace{1cm} (3)$$

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

$$\forall i \in [1..R], P(b_i|a_i, T) = \exp(-L(a_i, b_i))$$  \hspace{1cm} (4)$$

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

We assume that when recombination happens, it affects a region which is uniformly distributed along the genome and of length geometrically distributed with mean $\delta$. When $B$ blocks of the genome are under study for a total sequence length of $L$, the prior for $x_i$ and $y_i$ is [13]:

$$P(x_i = s|\delta) = \begin{cases} 
\delta/(B\delta + L - B) & \text{if } s \text{ is at the beginning of a block} \\
1/(B\delta + L - B) & \text{otherwise.} 
\end{cases}$$  \hspace{1cm} (5)$$
\( P(y_i = s|x_i, \delta) = \begin{cases} \\
\delta^{-1}(1 - \delta^{-1})^{s-x_i} & \text{if } s \text{ is before the end of the block} \\
(1 - \delta^{-1})^{s-x_i+1} & \text{if } s \text{ is the end of the block} \\
\end{cases} \)  

Combining Equations 2 to 6, we finally get the complete distribution of the recombination events in \( R \):

\[
P(R|T, \rho, \delta) = \exp(-\rho T/2)(\rho/2)^R \prod_{i=1}^{R} P(x_i, \delta)P(y_i|x_i, \delta)\exp(-L(a_i, b_i))
\]

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 [22] where all substitutions are equally likely, but our model can equally be used with other mutational processes [23].

**Bayesian Inference**

Let \( D \) denote the dataset for which we want to perform inference, and let us assume for the moment that the clonal genealogy \( T \) is known as well as the values of the parameters \( \theta, \rho \) and \( \delta \). We want to perform inference on the posterior distribution:

\[
P(R|D, T, \rho, \delta) \propto P(R|T, \rho, \delta)P(D|T, R, \theta).
\]

The first term (the prior) is given by Equation 7. In order to compute the second term (the likelihood), we define the local tree \( T_s \) of each site \( s = 1, ..., L \) 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 \( s \) depends on the ancestry graph only through the local tree \( T_s \) and therefore the likelihood can be decomposed as:

\[
P(D|T, R, \theta) = \prod_{s=1}^{L} P(D_s|T_s, \theta)
\]
where each of the terms $P(D_s|T, \theta)$ can be computed using the pruning algorithm of [24].

In order to perform inference, we use a reversible-jump Monte-Carlo Markov Chain [25]. 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 a “add” move which proposes to add a recombination event with properties proposed according to their priors as defined in Equations 3 to 6. These two moves are accepted according to their Metropolis-Hastings-Green ratio as described in Appendix A. We also use non-transdimensional moves proposing to update the departure point, arrival point, starting site and finishing site of an existing recombination 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 $T, \theta, \rho$ and $\delta$. Direct inference could in principle be done when those quantities are unknown by adding MCMC moves for those parameters, with slight modifications to standard tree updates found in e.g. [26]. 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 $T, \theta, \rho$ and $\delta$ are known, inference can be greatly simplified by noticing that the recombination events affecting the various alignment blocks $b = 1, ..., B$ 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:

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

This is a key property of our model, which does not hold in the Ancestral Recombination
Graph model since two recombinant branches affecting different regions may coalesce with each other. Thus inference when $T$, $\theta$, $\rho$ and $\delta$ 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 and as such are natural biological entities. Furthermore, when whole genomes are being used, the statistical uncertainty on $T$, $\theta$, $\rho$ and $\delta$ 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 $T$ given the data $D$.

Step 2 : Infer the mutation rate $\theta$, recombination rate $\rho$ and average tract length of recombination $\delta$ given the data $D$ and the clonal genealogy $T$ 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 $T$ inferred in the first step, and the parameters $\theta$, $\rho$ and $\delta$ inferred in the second step.

In practice, we perform Step 1 using the ClonalFrame algorithm [13]. Step 2 is performed by running the inference under our model for each alignment block independently, with $\theta$, $\rho$ and $\delta$ 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 $\theta$, $\rho$ and $\delta$ when performing Step 3.

Applications to data

Application to simulated data

We used SimMLST [20] 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 $\rho$ in the corresponding graph, given the tree, the mutation rate $\theta$ and the recombination tract length $\delta$. We consider sequences of length 10000bp, which is characteristic of genomic alignment block sizes.

Inference on an ARG with $N=10$ sequences, $\theta = 300$, $\rho = 50$ and $\delta = 236$ is considered on Figure 2. There are no instances of confidently inferred but incorrect recombination events in this example, with false-positive recombination intensity being limited to two types. Firstly, the edge of recombination region is sometimes imperfectly found (e.g. on branch 1 around 5200bp), and secondly the origin may be incorrect (e.g. parent of branch 2 and 8, 100bp). 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 edge may be unclear (e.g. an arrival at branch 9, 10 or their parent at 5500bp); the recombination may have poorly defined edges, or the origin may be poorly determined. As the distance from the tips increases, the inference becomes less certain because the data becomes 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 the mean $\rho = 25.4$ (95% confidence interval $[16.0, 37.6]$).

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 $\rho$ for a range of ARGs simulated with $N = 20$, $\delta = 236$ and varying $\rho = (25, 50, 75, 100, 150, 200, 250, 300, 400)$ and $\theta = (50, 100, 2000, 300, 400, 500)$. We average over 10 ARGs for each set of parameters to reduce variability, which can be very
large under the ARG model. ClonalOrigin infers $\rho$ that is much closer to the true value than does ClonalFrame, which tends to underestimate $\rho$ by roughly a factor of 2 because it misses events that have origins close to the departure point on the tree (it primarily captures imports from outside the diversity of the sampled taxa). ClonalOrigin infers the correct recombination rate for low $\rho$, and overestimates $\rho$ when the mutation rate $\theta$ 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 $\rho$ (and large $L$), but should be interpreted in terms of the number of distinct recombined tracts (rather than recombination events) as $\rho$ increases.

**Application to a Bacillus genomic dataset**

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 [27]. MLST has been applied to the *B. cereus* group to investigate its population structure and history [11, 28]. Analysis of MLST data using ClonalFrame found that recombination occurs at a rate approximately a fifth of that of mutation ($\rho/\theta \approx 0.2$) and results in a greater number of substitutions being introduced ($r/m \approx 1.5$) [13, 18, 29].

Since the sequencing of the first genome of *B. cereus* by [30], several more isolates have been fully sequenced [31–35]. We collected 13 such genomes summarized in Table 2, and aligned them using **progressiveMauve** [36,37]. We found $B = 1218$ blocks of homologous
sequence shared between all genomes, with lengths ranging from 502bp to 55619bp, and combined length $L = 3636155$ 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 (Supplementary Figure 1). We applied GenoPlast [38] 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 (Supplementary Figure 2).

**Application of the step-by-step methodology**

For Step 1 of our process, we reconstructed the clonal genealogy of the sample using ClonalFrame [13]. A unique tree topology was inferred, with little uncertainty in the branch lengths (Supplementary Figure 3). The same topology was also found when using UPGMA, Neighbor-Joining, Maximum-likelihood or Minimum Evolution (Supplementary Figure 4). ClonalFrame also found that many recombination events happened during the evolution of the sample, as shown in Supplementary Figure 3. These were found to happen at rate $\rho/\theta = 0.21$ with an interquartile range (IQR) of $[0.20; 0.23]$ relative to mutation, to be of average length $\delta=171bp$ (IQR $[168; 175]$) and to result in $r/m = 2.41$ (IQR $[2.37; 2.45]$) more substitutions introduced by recombination than by mutation.

For Step 2 we 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 $\delta$ was 236bp (IQR $[131; 537]$). A few blocks however took extremely low or high values, reflecting the limited information available on $\delta$ 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]). Three regions of the genome seemed to have potentially higher rates of recombination (Figure 4). The median inferred value of $\rho/\theta$ 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 obtain our Step 3 results by applying our new model again with values of $\delta$, $\theta_s$ and $\rho_s$ fixed to the median values of the previous paragraph.

**Patterns of recombination inferred across the genomes**

We estimated that $\sim$240,000 recombination events occurred since the 13 genomes shared a clonal common ancestor, but most of these events affected the deep branches of the clonal genealogy, where uncertainty is important. 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 prior of Equation 7. The main pattern in this figure is that genomes recombine 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. [39]). In the *Bacillus cereus* group, genetic exchanges have previously been found to occur more often within than between clades using MLST data [18]. 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. Firstly, this branch has imported approximately two times more material from
external sources (ie. origins above the MRCA of our sample) than expected under the prior, even though the dependency of recombination to sequence identity should limit the frequency of such imports. Secondly, 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 dataset. This increase is evenly distributed along the genome (Supplementary Figure 5).

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 0.8Mbp, 1.8Mbp, 3.2Mbp and 4.9Mbp along the genome of ATCC14579. This result confirms that the variation observed when inferring $\rho$ 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* [40, 41] and in *E. coli* [42, 43]. Here the two peaks at 0.8Mbp and 1.8Mbp correspond to regions of important change in GC content [30]. The peak at 3.2Mbp contains a large number of genes annotated with antibiotic and other drug resistance [30] which may be under positive selection. The peaks at 0.8Mbp and 4.8Mbp are also located near to *rrn* operons [30] and a tRNA gene array that harbors the integration site for a *Bacillus* site-specific integrative conjugative element [44].

**Recombination events inferred in specific regions**

Figure 7 shows the recombination events found in the first 2000bp of the two blocks shown by a blue and a green dot (respectively) on 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 \textit{dnaA} gene and the beginning of the BC0002 gene \cite{30} which both play essential roles in DNA replication. One of the clearest recombination events detected in this region is an import into genome 11 from the cluster containing genomes 1, 4 and 12 which spans the entire region shown in Figure 7. In another clear event, genome 10 imported the first \(~1000\text{bp}\) 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 \(~100\text{bp}\) (ie. before the start of gene \textit{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 inter-clade 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 or 9.

The second region shown in Figure 7 is located in the third hotspot of recombination at 3.2Mbp (Figure 6). This region contains the end of the BC3152 gene which produces an arsenate reductase, the BC3153 gene which produces an arsenic-resistance protein, the BC3154 gene which produces a lactoylglutathione lyase and the beginning of the BC3155 gene which produces an arsenic resistance operon repressor \cite{30}. We found many recombination events in this region, and no single branch of the clonal genealogy remains unaffected (Figure 7). Some of these events represent inter-clade recombination, for example the import of the last \(~400\text{bp}\) of genome 12 (which belongs to clade 1) from clade 2, or the import of centered on position 1400bp 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 \(~600\text{bp}\) of genome 4 from the ancestor of genomes 3 and 9, whereas others present more uncertainty, for example the import at position 100bp on genome 5 which could come from at least 6 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 inter-clade 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 may be under positive selective pressure [45] which often implies a higher rate of recombination [40, 41, 43, 46].

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. The level of diversity can be used to estimate the mutation rate $\theta = 2mN_e$, where $m$ is the mutation frequency per individual per generation, and $N_e$ the effective population size. Recombination can be quantified by a second parameter $\rho = 2rN_e$, where $r$ is the recombination frequency per individual per generation. Information on the value of $\rho$ 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 [6, 47–49]. This technique has also been applied to bacteria and archaea [43, 50–52].

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 [5]. A further difference is that recombination occurs in a clonal context, due to the asymmetry of the contributions of donor and recipient cells. In eukaryotes apart from recombination events that take place within defined pedigrees there is typically little or no information on the organismal context in which recombination events occur. The peculiar reproductive biology of bacteria and archaea has motivated us to develop a new model of recombination that better matches the major features of such organisms.

We have described a new method, ClonalOrigin, to infer recombination from an alignment of whole bacterial genome sequences. ClonalOrigin follows the standard population ge-
netic approach in estimating values of $\theta$ and $\rho$, which are assumed to be constant across the genome (although see below). However, we seek to infer individual recombination events within the clonal context in which they occur. Indeed the inferred number and size of events on specific branches of the clonal genealogy may differ substantially from the expectation given by $\theta$ and $\rho$. Furthermore, although our model assumes that the source of recombination events is as expected if the individuals in the sample were drawn from a random mating population (Equation 4), the inferred events can follow a very different pattern, which provides considerable information on the diverse biological processes affecting recombination in bacteria.

The ability of ClonalOrigin to detect such patterns was illustrated by our 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). Although we do not attempt it here, it is possible in principal to infer the DNA substitutions introduced by each recombination event and hence, for example, to study to what extent recombination facilitates the spread of different types of advantageous DNA sequence variation through the population and the replacement of newly arisen mildly deleterious mutations with the wild type allele.

A key difference of the model underlying ClonalOrigin in comparison with our previous effort ClonalFrame [13], is that the origin of recombination events is explicitly modelled. We showed that this difference makes ClonalOrigin more accurate than ClonalFrame (Figures 2 and 3) when detecting recombination from the Ancestral Recombination Graph model [19,20]. Furthermore, it allows a quantification of the genetic flux between lineages (Figure 5) which would not be possible otherwise.

Another important difference with ClonalFrame is that here we focus on the analysis of whole genome data. With the recent improvements in cost and speed of sequencing, such
data is becoming increasingly frequent \cite{53,54}. But despite the huge potential that this represents for our understanding of microbial evolution, the analysis of such large data sets also raises a number of statistical challenges. ClonalOrigin avoids these issues by being based on a model under which alignment blocks can be analyzed independently given the clonal genealogy and global parameters $\theta$, $\rho$ and $\delta$ (Equation 10). Statistical inference is then decomposed into a three step process which infers successively the clonal genealogy, global parameters and recombination events. For the first step we used ClonalFrame \cite{13} which is not strictly 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 (Supplementary Figure 4). 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 very close to the truth given the large amount of genomic sequence considered ($>3$Mbp) and the little variation in their inferred values across blocks (Figure 4). Step 3 introduces no further approximation.

Since our model is based on the coalescent \cite{21}, it can easily be extended to account for a number of additional biological processes. For example the basic coalescent model we used here (Equation 1) assumes a constant population size, but any population dynamics model can be accounted for by simple rescaling of the coalescent time scale \cite{55}. Any such model could therefore be used in our inferential setup simply by changing our priors for the clonal genealogy (Equation 1) and for the departure time of a recombinant edge given its arrival time (Equation 4). It would also be possible to use existing extensions of the coalescent designed to tackle the effect of population structure \cite{56,57}. Such extensions
typically allow coalescence to only occur for two lineages sharing the same deme, but here we could also model the effect of population structure on the recombination process by allowing recombination to occur only within demes.

The model we described assumed a constant rate of mutation and recombination along the genome. This assumption held relatively well on the Bacillus dataset we analyzed (Figure 4), but may not always be realistic. 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 [47] or DualBrothers [58]. If inference was performed in a single step, this extension would require to add computationally expensive reversible jump moves [25]. However, because of the way we decompose inference in three steps, variation (on a scale of ~10 alignment blocks) could be detected by performing the first two steps as described, and then fitting a changepoint model in the scatterplots shown in Figure 4 instead of taking the median in the third step as we did.

One final limitation of the method is that the input alignment is assumed to contain only sites that are either orthologous or xenologous [59]. If the input alignment contains alignments of paralogous gene copies, our method may incorrectly attribute within-genome paralogous gene conversion events as imports from external lineages. We expect such scenarios to be uncommon for several reasons. First, gene families in bacteria and archaea are generally small. Second, the alignment method used here operates at the DNA level, yet many multi-gene families have homology recognizable only at the amino-acid level and so erroneous alignments among paralogs will not occur. Finally, the alignment method used here targets alignment of the positionally-conserved homolog in multi-gene families [37], which will result in alignment of the orthologous or xenologous copy except in the case of duplication and subsequent gene conversion among paralogs. The mere existence of duplicate gene copies does not pose a problem.
Acknowledgments

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.
Figure legends

**Figure 1**: Illustration of our model for a single region of 300bp and a sample of 4 isolates. Part A shows the full graph of ancestry, with the clonal genealogy shown in bold black and two recombination events shown in red and blue. Part B shows the local trees for each site. Points $a_i$ are denoted as “departures” of recombination edges from the tree and $b_i$ are “arrivals”, with $b_i$ occurring closer to the sequence tips.

**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 edge 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 probability of recombination. 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).

**Figure 3**: Inferred values of $\rho$ relative to true values using simulated data for many simulations and across various parameter values. Shown are values for ClonalFrame (magenta) and ClonalOrigin (blue). For each of the 6 values of $\theta$ we plot the median (thick line) and inter-quartile range (thin line) of the ratio of inferred $\rho$/true $\rho$, considering the combined results for 10 different instances of the ARG. Lines are labeled by the order they appear at $\rho = 400$. The true value of 1 is shown as a horizontal line for comparison.

**Figure 4**: Scatterplots for all blocks of the *Bacillus* alignment of the inferred values of the log-average tract length ($\log(\delta)$), the mutation rate per site ($\theta_s/2$) and recombination rate per site ($\rho_s/2$). The median for all blocks is shown in red.
**Figure 5:** Heatmap showing for each donor/recipient pair of branches the number of recombination events inferred relative to its expectation under our prior model. The cells in very light gray are the ones for which the ratio would be meaningless because there are less than 3 observed and expected events.

**Figure 6:** Scatterplot of the number of recombination events boundaries per site for each block in the alignment of *Bacillus*. Details of the two blocks shown by a blue and green dot are shown in Figure 7.

**Figure 7:** Results of our analysis for two regions of the *Bacillus* alignment. The representation is the same as for Figure 2. The two regions are shown by a blue and green dot respectively on Figure 6.
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D )</td>
<td>Aligned sequence data</td>
</tr>
<tr>
<td>( N )</td>
<td>Number of isolates</td>
</tr>
<tr>
<td>( L )</td>
<td>Total length of the alignment</td>
</tr>
<tr>
<td>( B )</td>
<td>Number of blocks in the alignment</td>
</tr>
<tr>
<td>( D_b )</td>
<td>Data contained in block ( b )</td>
</tr>
<tr>
<td>( D_s )</td>
<td>Data at site ( s )</td>
</tr>
</tbody>
</table>

Symbols used for the clonal genealogy

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T )</td>
<td>Clonal genealogy</td>
</tr>
<tr>
<td>( t_i )</td>
<td>Length of time during which there were ( i ) ancestors in the clonal genealogy</td>
</tr>
<tr>
<td>( T )</td>
<td>Sum of branch lengths of the clonal genealogy</td>
</tr>
</tbody>
</table>

Symbols used for the recombination events

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R )</td>
<td>Set of recombination events and their properties</td>
</tr>
<tr>
<td>( R )</td>
<td>Number of recombination events</td>
</tr>
<tr>
<td>( R_b )</td>
<td>Recombination events affecting block ( b )</td>
</tr>
<tr>
<td>( a_i )</td>
<td>Departure point of the recombination event ( i )</td>
</tr>
<tr>
<td>( b_i )</td>
<td>Arrival point of the recombination event ( i )</td>
</tr>
<tr>
<td>( x_i )</td>
<td>First site affected by the recombination event ( i )</td>
</tr>
<tr>
<td>( y_i )</td>
<td>Last site affected by the recombination event ( i )</td>
</tr>
<tr>
<td>( T_s )</td>
<td>Local tree at site ( s )</td>
</tr>
</tbody>
</table>

Symbols used for the remaining parameters

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta/2 )</td>
<td>Rate of mutation on the branches of the clonal genealogy and the recombination edges</td>
</tr>
<tr>
<td>( \theta_s/2 )</td>
<td>Per-site rate of mutation</td>
</tr>
<tr>
<td>( \rho/2 )</td>
<td>Rate of recombination on the branches of the clonal genealogy</td>
</tr>
<tr>
<td>( \rho_s/2 )</td>
<td>Per-site rate of recombination</td>
</tr>
<tr>
<td>( \delta )</td>
<td>Mean of the geometric distribution modelling the length of recombinant segments</td>
</tr>
</tbody>
</table>

Table 1. Table of symbols
<table>
<thead>
<tr>
<th>Index</th>
<th>Genome</th>
<th>Clade</th>
<th>Length</th>
<th>ORFs</th>
<th>GenBank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>B. anthracis</em> Ames 0581</td>
<td>1</td>
<td>5503926</td>
<td>2887</td>
<td>NC_007530</td>
</tr>
<tr>
<td>2</td>
<td><em>B. cereus</em> 03BB102</td>
<td>1</td>
<td>5449308</td>
<td>2918</td>
<td>NC_012472</td>
</tr>
<tr>
<td>3</td>
<td><em>B. cereus</em> AH187</td>
<td>1</td>
<td>5599857</td>
<td>2973</td>
<td>NC_011658</td>
</tr>
<tr>
<td>4</td>
<td><em>B. cereus</em> AH820</td>
<td>1</td>
<td>5588834</td>
<td>2977</td>
<td>NC_011773</td>
</tr>
<tr>
<td>5</td>
<td><em>B. cereus</em> ATCC 10987</td>
<td>2</td>
<td>5432652</td>
<td>3006</td>
<td>NC_003909</td>
</tr>
<tr>
<td>6</td>
<td><em>B. cereus</em> ATCC14579</td>
<td>2</td>
<td>5427083</td>
<td>2787</td>
<td>NC_004722</td>
</tr>
<tr>
<td>7</td>
<td><em>B. cereus</em> B4264</td>
<td>2</td>
<td>5419036</td>
<td>2783</td>
<td>NC_011725</td>
</tr>
<tr>
<td>8</td>
<td><em>B. cereus</em> G9842</td>
<td>1</td>
<td>5736823</td>
<td>3052</td>
<td>NC_011772</td>
</tr>
<tr>
<td>9</td>
<td><em>B. cereus</em> Q1</td>
<td>1</td>
<td>5506207</td>
<td>2750</td>
<td>NC_011969</td>
</tr>
<tr>
<td>10</td>
<td><em>B. cereus</em> ZK</td>
<td>1</td>
<td>5843235</td>
<td>2872</td>
<td>NC_006274</td>
</tr>
<tr>
<td>11</td>
<td><em>B. thuringiensis</em> Al Hakam</td>
<td>1</td>
<td>5313030</td>
<td>2449</td>
<td>NC_008600</td>
</tr>
<tr>
<td>12</td>
<td><em>B. thuringiensis</em> konkukian</td>
<td>1</td>
<td>5314794</td>
<td>2593</td>
<td>NC_005957</td>
</tr>
<tr>
<td>13</td>
<td><em>B. weihenstephanensis</em> KBAB4</td>
<td>3</td>
<td>5872743</td>
<td>2910</td>
<td>NC_010184</td>
</tr>
</tbody>
</table>

**Table 2.** Genomes of the *Bacillus cereus* group used in this study
Appendices

Appendix A: Details of the MCMC moves

Reversible-jump moves

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

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

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

where $R$ is the old value of the parameter and $R'$ 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 which is calculated using Equation 9. The second term is the ratio of priors which is calculated using Equation 7. This only leaves to calculate the third term, ie. the ratio of proposal distributions. For the “add” move we have (cf. Equation 7):

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

so that the acceptance ratio is:

$$\alpha = \min \left( 1, \frac{P(D|R')} {P(D|R)} \frac{\rho T}{2(R+1)} \right)$$  \hspace{1cm} \text{(A3)}$$

For the “remove” move, if we note the edge proposed to be removed with a * we have:

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

so that the acceptance ratio is:

$$\alpha = \min \left( 1, \frac{P(D|R')} {P(D|R)} \frac{2R}{\rho T} \right)$$  \hspace{1cm} \text{(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 $[0, D]$, ie:

$$Q(x \rightarrow x + d) = \frac{P(D|x+d)P(x+d)}{\sum_{j=0}^{D} P(D|x+j)P(x+j)}$$  \hspace{1cm} (A6)

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

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

$$\alpha = \min \left( 1, \frac{\sum_{j=0}^{D} P(D|x+j)P(x+j)}{\sum_{j'=-D}^{D} P(D|x+d-j')P(x+d-j')} \right) .$$  \hspace{1cm} (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 $\epsilon \sim \text{Normal}(0.1)$ to its age. If the age is decreased and a number $n$ of nodes of the clonal genealogy are crossed, then one of the $2^n$ daughter branches is chosen at random to place the point on. If the age is increased then let $-n$ denote the number of nodes of the clonal genealogy that are being crossed. The Metropolis-Hastings acceptance probability of this move is therefore:

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

where the ratio of 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 $\theta$, $\rho$ and $\delta$

Here we described the additional moves of the MCMC required when $\theta$, $\rho$ and $\delta$ are treated as additional parameters as is required in Step 2 of the analysis.

**Update of $\theta$**

We used an improper Uniform prior on $[0, \infty)$ for the mutation rate $\theta$ and updated its value by proposing the addition of a perturbation $\epsilon$ 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 which can be computed using Equation 9.

**Update of $\rho$**

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

**Update of $\delta$**

Given Equations 5 and 6, the likelihood of the mean length of imports $\delta$ is of the form:

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

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

29
References


Figure 1

A

\[
\begin{align*}
    x_1 &= 50 \\
    y_1 &= 200 \\
    a_1 &
\end{align*}
\]

\[
\begin{align*}
    x_2 &= 1 \\
    y_2 &= 100 \\
    a_2 &
\end{align*}
\]

B

From 1 to 50

From 51 to 100

From 101 to 200

From 201 to 300
Figure 2
Figure 3
Figure 4
Figure 6