Inference of recombination properties in bacteria from whole genomes

M. Azim Ansari
St John's College

Department of Statistics
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hillary Term 2014

To Becky for all her love and support. I could not have done it without you. For Jaanali, your smiles keep me going. For my mum and dad and all of my family who have supported me and set me on my path.

Acknowledgements

First of all I would like to thank my supervisor Xavier Didelot for teaching me a huge amount during the last few years. I will always be indebted to him. I would also like to thank Rory Bowden for his support and advice as my second supervisor. Many thanks to my office mates Jo Herman and Bethany Dearlove for many discussions we have had. I would also like to thank all of the people in the Modernising Medical Microbiology consortium who taught me, gave me advice and helped me when I needed it. Furthermore, I would like to thank my friends Philip Maybank, Sofia Piltz and Remi Bardenet for listening to me, many lunches that we had together and all the advice and help. Finally, I would like to thank my examiners Gil McVean and Daniel Falush for reading my thesis and their helpful discussion and constructive comments.

Inference of recombination properties

in bacteria from whole genomes

M. Azim Ansari, St John's College, Hillary 2014

Abstract

The concept of species in bacteria is a matter of contention. The current definition is based on DNA-DNA hybridisation and does not account for evolutionary forces that are important in demarcating species. In this thesis we investigate two evolutionary forces that are important in speciation in bacteria, propose novel statistical models for them and infer parameters of interest.

We present the first attempt at inferring the bias in the recombination process from whole bacterial genomes. Despite empirical evidence that recombination is biased and theoretical results that this bias is important in speciation, it is usually ignored. We propose a coalescent based model that accounts for the bias in the recombination process. We use approximate Bayesian computation for inference and describe an efficient method for simulating from the model. We show that our method performs well on simulated datasets and is robust to slight misspecification of the history of the samples. Application of our method to a *Bacillus cereus* dataset shows that it contain evidence that the recombination process depends on the evolutionary distance between donors and recipients. We demonstrate that the rate of bias in the recombination process for this dataset is far lower than what theoretical studies require for the spontaneous generation of populations that can be called species under neutral model.

Next we propose a model for occurrence of adaptive events on a phylogenetic tree. We use the model to infer the boundaries of clusters on a phylogenetic tree that correspond to ecologically distinct lineages. we characterise our method using simulated datasets and show that it is conservative in estimating the number of adaptive events. Finally we apply our method to two bacterial datasets of *Salmonella enterica* and *Vibrionaceae*. We show that there is decisive evidence that isolates in these datasets partition into numerous ecologically distinct lineages and use our method to delineate the boundaries of these lineages.

Contents

List of Figures

Chapter 1

Literature Review

1.1 Introduction

Bacteria are one of the three major domains of life along with Archaea and Eukarya. They are unicellular organisms that are generally able to carry out their life processes of reproduction and energy production independent of other cells. However, in nature they exist as populations that interact with other populations of cells and form communities. Due to their rapid growth (*Escherichia coli* for instance can double its population number every 20 minutes in ideal conditions, Martinko *et al.* 1997) they can quickly increase their number and affect the chemical and or physical properties of their environment.

Microbial interactions with animals and plants are widespread and in many cases it is symbiotic. For instance one of the most important bacteria plant relationship is that of Legumes (such as soybean, peas and beans) and nitrogen fixing bacteria. The plant produces the organic energy source that is needed by the bacteria and the bacteria provides the fixed nitrogen essential for the growth of the plant (Martinko *et al.*, 1997). On the other

hand there are many diseases in plants and animals that are caused by bacteria. Some of the bacterial species that cause diseases in humans (such as *Staphylococcus aureus*, *Pseudomonas Aeruginosa* and *Escherichia coli*) are gaining resistance to known antibiotics and are already causing diseases that are very difficult to treat (Arias and Murray, 2009).

Bacteria are an essential part of the life on Earth and understanding them will help to protect us, use them better in processes such as biotechnology and to understand and safeguard our environment.

One of the biological processes that bacteria engage in is recombination and a major part of this thesis is dedicated to investigating this process and to learn the properties of the recombination process from bacterial genomes.

This chapter is an introduction to the recombination process in bacteria and a review of the related models. The chapter begins with an introduction to the recombination process in bacteria. Subsequently, reproductive models that look forward in time, some of their properties and their drawbacks are presented. In the following section the coalescent model, which looks at the genealogy of a sample of a population backwards in time and its relationship to the forwards in time models, is established. Additions to the coalescent model such as mutation and recombination are presented in the subsequent section. This will then be followed by approximations to the coalescent model with recombination.

1.2 Bacteria and recombination

In higher organism which usually reproduce sexually, recombination is an integral part of the reproductive process and it introduces new genetic variants to the population. All recombination events in eukaryotes are homologous recombination i.e. the exchange of DNA is between

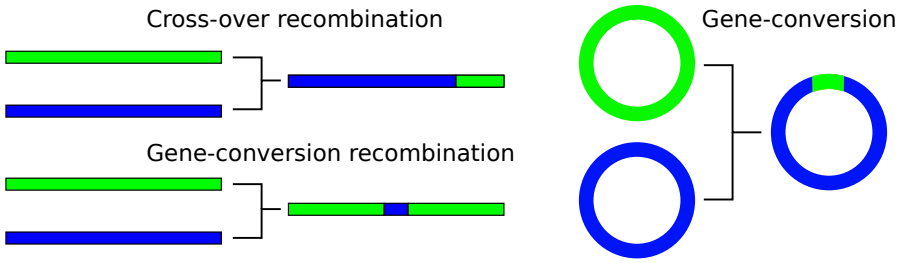
homologous segments of chromosomes. In eukaryotes there are two types of recombination:

Cross-over recombination usually occurs during meiosis between two paired chromosomes. Large segments of DNA from the recombination break point to the end of the chromosome are exchanged.

Gene-conversion recombination also usually occurs during meiosis between paired chromosomes. A small segment of one chromosome is exchanged with its homologue on another chromosome without its flanking regions.

Bacteria on the other hand reproduce by binary fission. Thus the two daughter cells are almost identical to the parent cell apart from perhaps a few mutations which occur during the DNA duplication process. Despite being clonal, recombination plays a crucial role in the evolution of bacteria (Feil and Spratt, 2001; Fraser *et al.*, 2007; Vos, 2009).

In bacteria recombination is defined as the process where a relatively small DNA segment from a donor cell is integrated into the recipient cell's genome. This process can lead to two outcomes, non-homologous and homologous recombination (Vos, 2009). Non-homologous recombination occurs when a novel segment of DNA from the donor cell is inserted into the genome of the recipient cell. On the other hand, homologous recombination happens when the DNA from the donor cell replaces its homologous counterpart in the genome of the recipient cell. In this study we are only concerned with the "core" genome of regions present in all sampled genomes (Medini *et al.*, 2008), and therefore only homologous recombination is relevant.



(a) Recombination in Eukarya (b) Recombination in Prokarya

Figure 1.1: Recombination in eukaryotes versus recombination in prokaryotes. (a) There are two types of recombination in eukaryotes: Cross-over and gene-conversion. (b) In prokaryotes all homologous recombination events are equivalent to gene-conversion recombinations in eukaryotes.

The process of homologous recombination in bacteria is asymmetric in terms of the genetic contributions made by donor and recipient cells, since typically a small segment of DNA from the donor in the order of a few hundreds or thousands of nucleotides in length is incorporated into the genome of the recipient which is on average millions of nucleotide long (Didelot and Maiden, 2010). This asymmetry contrasts with the well-studied mechanism of crossing-over in eukaryotic sexual reproduction where the two parents contribute equally. In addition bacteria have circular chromosomes, therefore homologous recombination in prokaryotes is equivalent to gene-conversion in eukaryotes (Wiuf and Hein, 2000; McVean *et al.*, 2002). Figure 1.1 shows recombination events in eukaryotes and prokaryotes.

1.2.1 Mechanisms of transfer of DNA between bacterial cells

In bacteria, foreign DNA can be taken up by the recipient cell through one of the three mechanisms: conjugation (transfer of DNA from one cell to another when they are in physical contact), transduction (bacteriophage mediated DNA transfer) or transformation (uptake of DNA from the

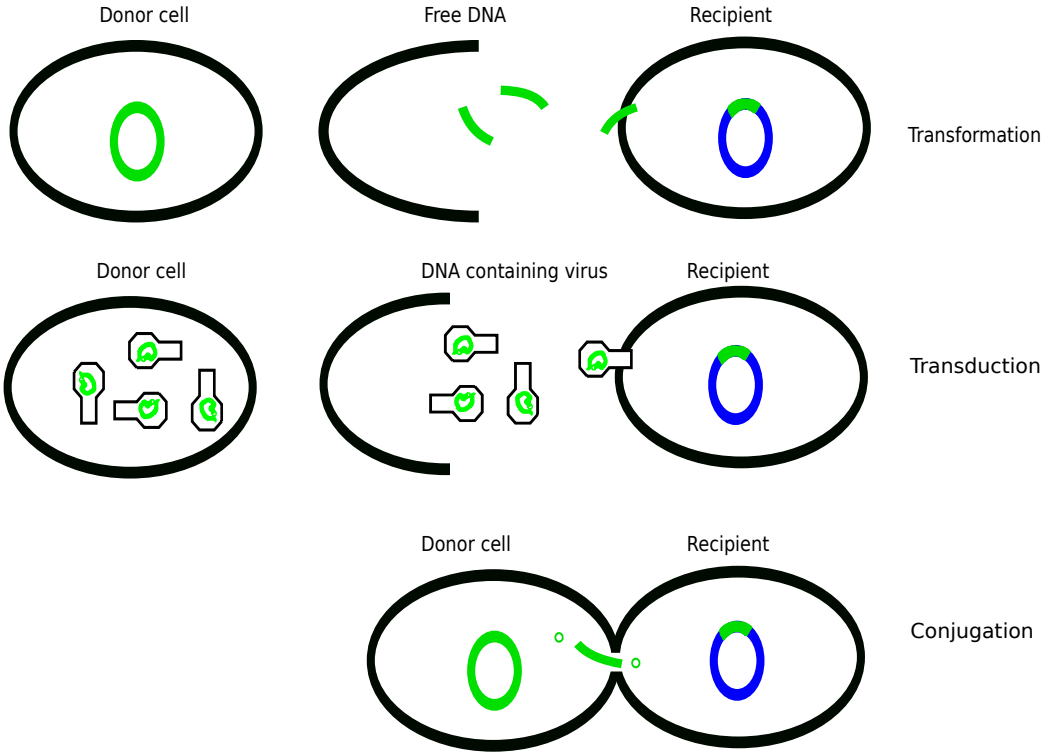


Figure 1.2: Illustration of the three mechanisms of recombination in bacteria. Transformation is shown at the top where naked DNA fragments are taken up by a recipient cell and used to replace the homologous counterparts on its genome. The second row illustrates transduction where a virus moves a fragment of DNA from a donor cell and injects it into the recipient cell. The foreign fragment then replaces the homologous section on the recipient cell's genome. The last row shows conjugation where two cells are in direct contact and exchange DNA fragments or plasmids or both.

environment by the recipient cell) (Thomas and Nielsen, 2005). In homologous recombination, the recipient cell then replaces the homologous section of its DNA with the foreign DNA segment. Figure 1.2 illustrates the three mechanisms of recombination in bacteria.

Mechanisms of recombination in bacteria are very different to that of higher organism (eukaryotes) where a lot of recombination modelling is focused and it is important to understand the differences and to take them

into account when modelling recombination in bacteria.

Transformation

Transformation is the uptake of exogenous DNA from the environment and its subsequent integration into the genome. For transformation to occur the recipient cell has to become competent. Competency happens through expression of specific proteins on the surface of the cell that binds exogenous DNA. Overall transformation involves around 20–50 proteins depending on the species (Thomas and Nielsen, 2005). Almost all transformable bacteria are thought to share common mechanisms for the uptake of the DNA and processing. These mechanisms use conserved genes that are all expressed on the onset of competence. The actual mechanisms of the transfer of DNA into the cell in most taxa are related to type IV pili and type II secretions systems (Dubnau, 1999; Chen *et al.*, 2005). A pilus on the recipient cell surface is bound to foreign double stranded DNA and one strand is degraded and the other strand is pulled into the cell through a trans-membrane channel (Chen *et al.*, 2005).

Although natural transformation has been observed in all the major groups of bacteria (around 80 species which are well represented among both gram-positive and gram-negative groups), not all species are naturally transformable (Lorenz and Wackernagel, 1994). Furthermore most species develop time-limited competency in response to specific environmental conditions such as cell density or starvation and the proportion of cells that develop competency can range from near zero to almost 100% (Johnston *et al.*, 2014). In some species such as *N. gonorrhoeae* and *H. influenzae* for transformation to be efficient the foreign DNA has to contain specific sequences that are overrepresented in

their own genomes (Goodman and Scocca, 1988), while species such as *B. subtilis* and *S. pneumoniae* display no preference for any specific sequence (Ochman *et al.*, 2000).

Transduction

Transduction occurs when foreign DNA is injected into the recipient cell by a bacteriophage. All transduction events are abnormal as the transfer of bacterial genetic material happens by machinery that is evolved primarily to transfer viral genetic material that is essential for its survival. When a cell is infected by a bacteriophage, the phage uses the cells machinery to replicate itself. Due to low fidelity of packaging, the phage capsid might contain bacterial DNA. When the cell dies and the phages are released, they can infect other cells and transfer the bacterial DNA to the recipient cell. Transduction is divided into two classes: generalised and specialised. In generalised transduction the bacteriophage has packaged random DNA fragments from the donor cell while in the specialised transduction DNA adjacent to the phage integration site in the donor cell is packaged into the capsid (Low and Porter, 1978).

Phages are thought to be extremely prevalent and to play a crucial role in the horizontal gene transfer process in bacteria (Weinbauer and Rassoulzadegan, 2003; Weinbauer, 2004). To survive, the bacteria have to be resistant to most type of phages that they encounter. This resistance mechanism can result in limited host range for some phages. As a result transduction can only lead to exchange of DNA between two strains if they are susceptible to the same phages. This has resulted in co-evolution of of phages and their host bacteria such that some species such as Salmonella can be identified by their susceptibility to a specific

phage (Le Minor, 1988).

Conjugation

Conjugation happens when two cells come into contact and plasmids or integrative conjugative elements (ICE) are exchanged. Conjugation is usually dependent on the presence of a plasmid in the donor cell that carries all the necessary genes to promote the transfer of DNA from the donor to recipient cell. In gram-negative bacteria such as *E. coli*, the donor cell needs to carry a filamentous tube on the cell surface. The structure of these tubes can be very different among different species. The pilus encoded by the F plasmid is long, thin and flexible while the pilus encoded by RP4 is short, thicker and rigid (Burrus *et al.*, 2002). The pili come into contact with the receptor proteins on the recipient cell's surface and form a mating pair. The pili then contracts and brings the two cells into direct contact. A pore is made between the two cell and then the transfer of DNA from donor to recipient occurs. In gram-positive bacteria such as *Enterococcus* and *Streptomyces* usually the number of genes required for the conjugation is much lower. This is partly due to the fact that gram-positive bacteria lack the outer membrane that the gram-negative bacteria have (Frost, 2009). In this specific species, the recipient releases a pheromone like protein into the environment that binds to a plasmid encoded receptor on the donor cell surface. Different types of plasmids encode for different types of receptors and therefore activated by different types of pheromones. The recipient cells can produce a range of pheromones and can mate with cells carrying different plasmids. The binding of pheromone to the receptor causes the donor cell to produce aggregating products that results in the aggregation of donor and recipient

cells (Frost, 2009).

Conjugation is common among all taxa of bacteria (Davison, 1999). For instance in *Vibrio* and *Pseudomonas* in gram-negative bacteria and *Streptomyces* and *Enterococcus* in gram-positive bacteria. Conjugation can happen between members of different species and even different kingdoms (Sawasaki *et al.*, 1996). Conjugation can only happen between cells which are physically close and usually requires long periods of contact between them and it tends to moves large portions of DNA between cells. Transfer of plasmid from the donor cell to recipient cell is the most common outcome of conjugation, however in some instances the transfer of chromosomal DNA also happens. Some plasmids can have a limited host range as they have to be able to replicate in the recipient cell. Conjugation is one of the main reasons for the transfer of antibiotic resistance genes (Courvalin, 1994) between cells.

1.2.2 Homologous recombination mechanism

With the exception of plasmids, once the foreign DNA is inside the cell, it has to be integrated into the recipient cells genome to persist for many generations. This is usually done through homologous recombination mechanism. Homologous recombination is a complex system with around 25 genes involved in the process. The necessary gene is *recA* which is present in all bacterial species studied. In homologous recombination the efficiency of integration depends on the degree of homology of the exogenous DNA and the host DNA (at least in parts) (Zawadzki *et al.*, 1995). Mismatch repair system (which their primary role is protection of genetic integrity) recognise and process mispaired and non-paired bases and can stop the recombination process if the foreign DNA is too divergent

(Matic *et al.*, 1996). Mutants that lack the mismatch repair system show a significant increase in their ability to recombine with more distant species (Rayssiguier *et al.*, 1989).

Restriction modification systems are another barrier to recombination. These system have evolved to protect bacteria from infection by exogenous DNA such as bacteriophages. However this system cannot respond to all threats. Many phages escape the effects of the host restriction system by absence of restriction sites on their DNA or by inactivating the restriction modification proteins (Matic *et al.*, 1996). In addition, small DNA segments are likely not to contain some of the restriction sites and therefore not to be affected by the restriction modification enzymes.

It seems the main determinant of the success of the recombination event to be the amount of sequence divergence between the exogenous DNA and the genomic DNA. Experimental work has shown that the probability of recombination has a negative log-linear relationship with the amount of sequence in diverse species such as *Bacillus* (Majewski and Cohan, 1999), *E. coli* (Vulić *et al.*, 1997) and *S. pneumoniae* (Majewski *et al.*, 2000).

1.2.3 Comparison of recombination rates in bacteria

The most commonly used measure of relevance of recombination in bacteria is r/m which is the ratio of probabilities that a given nucleotide is altered by recombination to mutation (Guttman and Dykhuizen, 1994). Therefore it estimates the relative effect of recombination to mutation in terms of how many substitution they cause.

There are different methods for estimating r/m in bacteria. However comparing these estimates from different methods is not meaningful due to different assumptions and properties of the methods. Instead we use the

result from a single study (Vos and Didelot, 2009) where they have looked at r/m for 48 different bacterial and archaeal species. r/m changes by 3 orders of magnitude from 0.2 in *Leptospira interrogans* to 63.6 in *Flavobacterium psychrophilum*. The authors observed that all marine and aquatic species had high or very high rates of homologous recombination. On the other hand, Terrestrial bacteria had low to intermediate rates of homologous recombination. This is to be expected as marine and aquatic species are less likely to be affected by the geographic population structure as the waves, tides and currents migration and movement of the isolates. The authors also observed pathogens vary widely in their rate of homologous recombination. This is true for both obligate pathogens and commensal and opportunists pathogens. Their results show that there is a large variation in the rate of homologous recombination in bacteria and it plays a signifcant role in the genetic diversity of a large number of bacterial species. For more information we refer the reader to Vos and Didelot (2009).

1.3 Reproductive models

1.3.1 The Wright-Fisher model

Wright (1931) and Fisher (1930) introduced a model of population reproduction which has become known as the Wright-Fisher (WF) model. This simple model describes how genes are transferred in an idealised population from one generation to the next. The notation used is slightly different for haploid and diploid organisms. As this work focuses on the bacterial evolution, we use the notation for haploid organisms. The WF model makes some important simplifying assumptions. They are as follows:

- Generations are non-overlapping.
- Population size is constant.
- All individuals are equally fit.
- The population has no structure.

Using the WF model assumptions, we can formulate some of the statistical properties of the model. Assuming a population of size N, generation $t+1$ is derived from generation t by uniformly sampling from it with replacement. Figure 1.3 shows an example of a WF population of size seven. It also shows the genealogy of a sample from the present population, going back in time. The **genealogy** is defined by the lines of ancestry going back in time until all the samples have found a common ancestor.

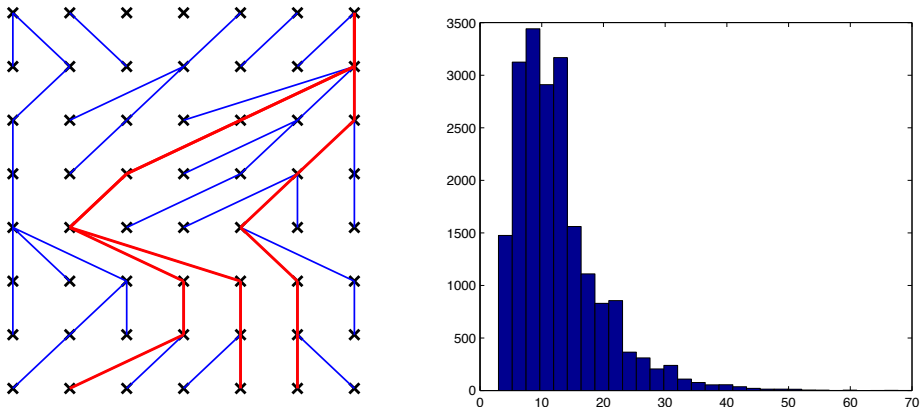
Probability of gene j in generation $t+1$ to have descended from gene i in generation t is $1/N$. Therefore the number of offspring of gene i in the next generation is binomially distributed. Number of descendants of gene i of generation t in the next generation C_i is distributed as:

$$\Pr\left(C_i = k\right) = \binom{N}{k}\left(\frac{1}{N}\right)^k\left(1-\frac{1}{N}\right)^{N-k} \qquad k = 0, 1, \ldots, N \tag{1.1}$$

For large values of N the binomial distribution can be approximated by the Poisson distribution with parameter $\lambda = N(1/N) = 1$:

$$\Pr\left(C_i = k\right) \approx \frac{1}{ek!} \tag{1.2}$$

Looking at genealogies an important property of the WF model is the distribution of number of generations to the first common ancestor event of a sample from the population going back in time. A **common ancestor event** is defined as when two lineages trace their ancestry to a single gene as shown in Figure 1.3. Let us assume we have a random sample of n genes



(a) Genealogy (b) Time to fixation

Figure 1.3: Illustration of a Wright-Fisher population and genetic drift. (a) An example of a WF population with seven genes. At each generation descendants are chosen by uniformly sampling with replacement from the current generation. The lines show the descendants of each gene. The red lines show the genealogy for a sample of size three from the present population going back in time. At some point every gene in the population will be descended from a single ancestor some generations ago. This event is called fixation. (b) Illustration of the randomness in a WF population. 20,000 simulations were done for a population of size seven where initially each gene has a different genotype and the time to fixation was recorded. The figure shows the histogram of the number of generations to fixation. There is a huge amount of randomness in the model which needs to be taken into account when drawing conclusions.

from a population of size N. The first gene is free to choose its parent, but the second gene has to choose its parent from $N-1$ remaining genes in the previous generation and so on. The probability that none of the n genes find a common ancestor in the previous generation is

$$\begin{aligned}\frac{N(N-1)\dots(N-n+1)}{N^n} &= (1-\frac{1}{N})\cdots(1-\frac{n-1}{N}) \\ &= \prod_{i=1}^{n-1}(1-\frac{i}{N}) = 1-\sum_{i=1}^{n-1}\frac{i}{N}+O(\frac{1}{N^2}) \\ &= 1-\binom{n}{2}\frac{1}{N}+O(\frac{1}{N^2}) \end{aligned} \tag{1.3}$$

In Equation 1.3, assuming that N is large, then $1/N^2$ and terms of higher order are negligible and can be ignored. As a result the probability of having no common ancestor event in the previous generation is approximately

$$1-\binom{n}{2}\frac{1}{N} = 1-\frac{n(n-1)}{2N} \tag{1.4}$$

And if $N >> n$, then the probability of having more than one common ancestor event happening in the same generation is negligible. Therefore the probability mass function of the number of generations to the first common ancestor event of n lineages T^n_{CA} is Geometrically distributed

$$\Pr(T^n_{\mathrm{CA}}=j) = \left(1-\binom{n}{2}\frac{1}{N}\right)^{j-1}\binom{n}{2}\frac{1}{N} \qquad j=1,\dots,\infty \tag{1.5}$$

It is important to note that the above approximation holds assuming that N is large and that $N >> n$. In addition it is important to note that despite the simplicity of the model, deriving exact probabilistic properties of this model are often difficult and require diffusion approximation (Kimura, 1985). Another way to study the model is Monte Carlo simulations i.e.

simulate the whole population for many generations, but these simulations tend to be computationally inefficient as population sizes are usually large and the process needs to run over many generations to reach equilibrium.

The Wright-Fisher model with mutation

Mutation is defined as a change of a base in a DNA sequence. Here we are only interested in the mutations that can be passed on to the next generation. As we are assuming neutrality, these mutations will not affect the reproductive patterns of the organism and thus can be seen as events that are independent from the reproductive process.

There are two main models of DNA sequence mutation. The first is the infinite site model (Kimura, 1969), which assumes mutations are rare events and that DNA sequences are very long. Hence multiple mutations on the same site are extremely rare so they can be ignored. Under the infinite site model all polymorphic sites are biallelic. The second is the finite site model, which assumes that mutations on all sites are equally likely and allows for several mutations on the same site. The finite site model has several variants, the simplest of which is the Jukes-Cantor model (Jukes and Cantor, 1969). This model assumes mutations to any of the three other nucleotides are equally probable. An improvement to the Jukes-Cantor model was introduced by Kimura (1980) which assumes transitions (A$\leftrightarrow$G and C$\leftrightarrow$T) are more likely than transversions (all other changes).

All of the above mutation models can be incorporated into the WF model. For simplicity we use the infinite site model. We assume that each gene passed on to the next generation is subject to mutation with probability u. This means following a single lineage backwards in time, the number of generations until the first mutation is observed T_{M} is

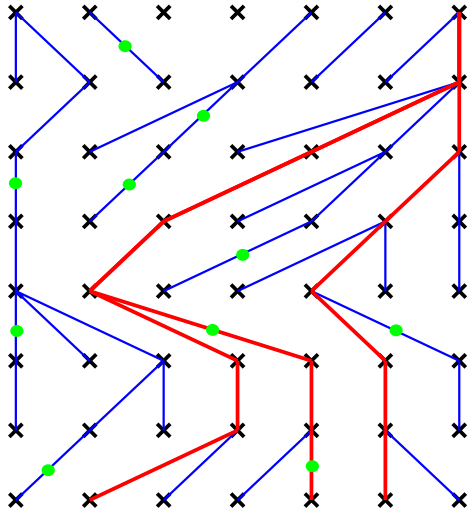


Figure 1.4: A Wright-Fisher population with mutation. The lines show the descendants of each gene and the green dots show the mutations that are passed on to the next generation. Under neutrality (mutations have no reproductive advantage), mutation is independent of the reproduction process and can be superimposed on top of it. The red lines show the genealogy of a sample of size three going back in time.

geometrically distributed

$$\Pr(T_{\mathrm{M}} = j) = u(1-u)^{j-1} \qquad j = 1, \ldots, \infty \tag{1.6}$$

and therefore the expected number of mutations in each generation is Nu.

The Wright-Fisher model with recombination

To add recombination to the WF model, let us assume that each gene passed to the next generation is subject to recombination with probability r. Hence following a single lineage back in time, the number of generations to the first recombination event T_{R} is geometrically distributed

$$P(T_{\mathrm{R}} = j) = r(1-r)^{j-1} \qquad j = 1, \ldots, \infty \tag{1.7}$$

Figure 1.5 shows a recombining WF population of size seven. At each recombination event the ancestry of the donor cell is shown in green dotted line. If the population is non-recombining then the entire length of the observed sequences evolve according to a single genealogy as shown in Figure 1.4. As recombination moves segments of one cell's DNA into another, in a recombining population different parts of the observed sequences will have different genealogies. As seen in Figure 1.11c, recombination in bacteria is asymmetric. The donor cell contributes a very small segment of DNA while the recipient cell provides the rest of the genome. Although there is no single genealogy that observed sequences evolve according to, we can define the concept of clonal genealogy (Didelot *et al.*, 2010). The **clonal genealogy** is the ancestry of a sample of isolates when at each recombination event the line of decent of the recipient cell is followed back in time. In a set of sampled isolates, given that the recombination rate is not extremely high, most parts of the observed genomes have evolved according to the clonal genealogy.

For a set of aligned sequences from a population of haploid organisms, all non-recombining sites have evolved according to the clonal genealogy, while the recombining sites have their own local genealogies. Parts of these local genealogies will be similar to the clonal genealogy as recombination affects a subset of the isolates. In Figure 1.5 the red line shows the clonal genealogy of a sample of size three. Most of the genome for the three sampled isolates has evolved according to this clonal genealogy, but some sections of the genomes will have different local genealogies for which the relevant green lines need to be traced back in time.

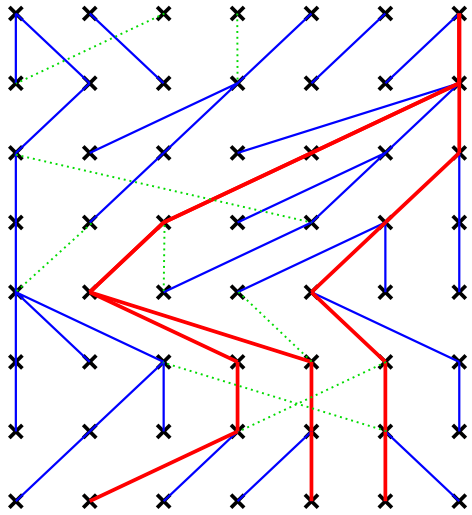


Figure 1.5: A Wright-Fisher population with gene-conversion recombination. Each gene passed on to the next generation has probability r of recombining. At each recombination event, the green dotted line shows the ancestry of the donor cell. The red lines show the clonal genealogy of a sample of size three taken from the population. Clonal genealogy is defined by following the ancestry of a sample back in time and at recombination events following the ancestry of the recipient cell. Clonal genealogy is an important concept in the bacterial evolution, as most of the sampled isolates' genomes have evolved according to it.

1.3.2 The Moran model

The Moran model (Moran, 1958) is an alternative reproduction model which is mathematically attractive. At each time step t an individual is chosen at random to reproduce and an individual is chosen at random to die. The same individual can be first chosen to reproduce and then die. Consequently, an individual in step t can have zero or one offspring in step $t+1$. The Moran model's simplifying assumptions are similar to that of WF model apart from the fact that as only one individual is allowed to reproduce in each time step, generations are allowed to overlap.

Let us assume we have a sample of size n from a population of size N. In each time step looking back in time, there are only two possibilities, two individuals finding a common ancestor or none finding a common ancestor. The probability that two of the individuals find a common ancestor in the previous time step is equal to the probability that the offspring did not replace its parent in the previous step $(N-1)/N$ times by the probability that both offspring and its parent are present in the sample $(n/N)(n-1/N-1)$:

$$\frac{N-1}{N}\frac{n}{N}\frac{n-1}{N-1} = \frac{n(n-1)}{N^2} = \binom{n}{2}\frac{2}{N^2} \tag{1.8}$$

Hence, following n lineages back in time, the number of time steps until the first common ancestor event T^n_{CA} is geometrically distributed

$$\Pr(T^n_{\mathrm{CA}} = j) = \left(1 - \binom{n}{2}\frac{2}{N^2}\right)^{j-1}\binom{n}{2}\frac{2}{N^2} \tag{1.9}$$

This distribution is exact, unlike the equivalent result of the WF model in Equation 1.5 where we used approximations. Figure 1.6 shows a population of size seven reproducing under the Moran model.

Calculating some of the properties of the Moran model are simpler than

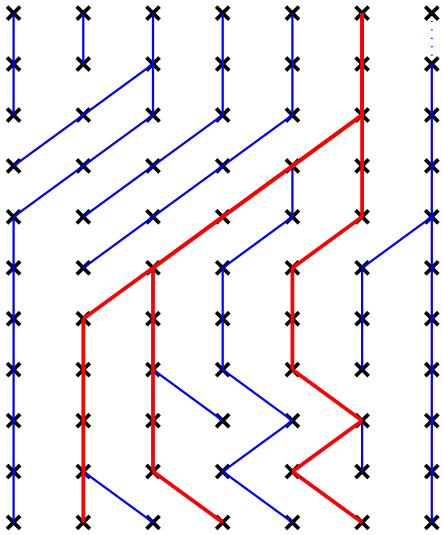


Figure 1.6: A population of size seven reproducing under the Moran model. In time step t an individual is chosen at random to reproduce and an individual is chosen at random to die. The same individual can be chosen to reproduce and then die. Such cases are shown with dotted lines. The red lines show the genealogy of a sample of size three going back in time.

the WF model, however this process can still be cumbersome and complex. Simulating populations under the Moran model is computationally inefficient due to large population sizes and the long number of time steps that simulation has to be carried over. Mutation and recombination can be added to the Moran model in a similar fashion as the WF model.

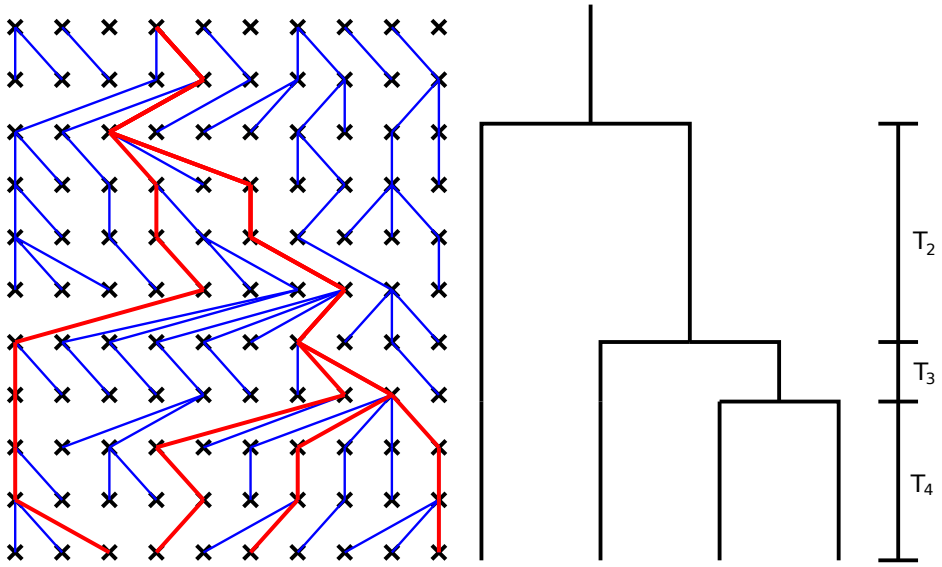
1.4 The coalescent model

KINGMAN (1982a,b) introduced the coalescent model. Until the early 80s most approaches were looking forwards in time (using models such as the WF and Moran) and were trying to calculate the expectation of the future events for a sample given present state of the population. In these models one usually needs to have information about the state of the whole population. Kingman took a different approach and introduced a model looking back in time. It looks at the genealogy of a sample backwards in time, ignoring genes in the population that are not relevant to the sample ancestry.

A coalescent event occurs when looking back in time two lineages find a common ancestor. Kingman expressed his model as coalescence of sampled lineages backward in time until all lineages find a common ancestor that is called the most recent common ancestor (MRCA). It can be proved that the coalescent model is the limiting case to the WF and Moran models (and many other population reproduction models) when the population size goes to infinity.

1.4.1 Deriving the coalescent model

Looking back in time at the ancestry of a sample of genes results in a tree that is called the **genealogy**. This tree has two parts, the branching order or the topology of the tree and the age of the nodes of the tree. Figure



(a) WF genealogy (b) Coalescent genealogy

Figure 1.7: The genealogy of a sample in a WF population and its equivalent coalescent tree. (a) The red lines show the genealogy of a sample of size four from a WF population going back in time until they meet their common ancestor. (b) The equivalent coalescent genealogy of the sample in the WF population is shown. It is important to note that different samples will have different genealogies and thus the randomness in the model has to be taken into account.

1.7 shows an example of such a tree. In the coalescent model the tree topology and the ages of the tree nodes are independent of each other. Any two branches are equally likely to find a common ancestor in the previous generation, therefore all branching orders are equally likely. The ages of the tree nodes T_n are defined as the time during which there are n lineages. In the following section we derive these times from the WF and Moran models.

Deriving the coalescent model from the WF model

Let us assume we have a sample of size n genes from a WF population of size N. From Equation 1.3, we know that following n lineages back in time the number of generations until the first coalescent (or common ancestor) event T^n_{CA} is geometrically distributed. Consequently, following n lineages back in time, the probability that the first coalescent event happens at generation $j+1$ or later is

$$\Pr(T^n_{\mathrm{CA}} > j) = \left(1 - \binom{n}{2}\frac{1}{N}\right)^j \tag{1.10}$$

For large values of N the probability of two simultaneous coalescent events becomes negligible and is ignored. Now if the time is measured in units of N generations and a change of variables is used where $t = j/N$ and $T_n = T^n_{\mathrm{CA}}/N$, then

$$\Pr(T_n > t) = \left(1 - \frac{\binom{n}{2}}{N}\right)^{Nt} \xrightarrow[N\to\infty]{} e^{-\binom{n}{2}t} \tag{1.11}$$

Hence

$$\lim_{N\to\infty} \Pr(T_n \leq t) = 1 - e^{-\binom{n}{2}t} \tag{1.12}$$

This change of variable allows us to map from a discrete unit of time measured in number of generations to a continuous unit of time where one unit of time is equivalent to N generations. Differentiating Equation 1.12

gives the probability density function of T_n. While there are k lineages, the time to the next coalescent event is exponentially distributed with parameter $\binom{k}{2}$

$$f_{T_k}(t) = \binom{k}{2} e^{-\binom{k}{2}t} = \frac{k(k-1)}{2} e^{-\frac{k(k-1)}{2}t} \tag{1.13}$$

This approximation is valid when population size N is large and the number of samples n is small relative to the population size N.

Based on Equation 1.13 the genealogy of the sample is defined by a series of coalescent events and the waiting times $T_n, T_{n-1}, \ldots, T_2$ between the coalescent events are exponentially distributed and are dependent on the current number of lineages. An important feature of Equation 1.13 is that it does not depend on the population size directly but through a scaling time. It also uses the fact that for a sample of size n from a population of size N much of the genealogy is irrelevant, which leads to very efficient simulation algorithms. To simulate the genealogy of a sample, we simulate the time to the next coalescent event T_k from Equation 1.13 and at each coalescent event we draw two lineages at random to coalesce.

Deriving the coalescent model from the Moran model

Let us assume that there exist a population of size N that is reproducing according to the Moran model. Based on Equation 1.9 we know that, following n lineages back in time, the time to the first coalescent event is geometrically distributed. Therefore, the probability that the first coalescent event happens at the latest by the time step j, is

$$\Pr(T^n_{\mathrm{CA}} \leq j) = 1 - \left(1 - \binom{n}{2}\frac{2}{N^2}\right)^j \tag{1.14}$$

Now if we scale the time, in units of $N^2/2$ time steps and introduce the change of variables $t = 2j/N^2$ and $T_n = 2T^n_{\text{CA}}/N^2$ then

$$\Pr(T_n \leq t) = 1 - \left(1 - \binom{n}{2}\frac{2}{N^2}\right)^{\frac{N^2 t}{2}} \xrightarrow[N\to\infty]{} 1 - e^{-\binom{n}{2}t} \tag{1.15}$$

Differentiating 1.15, gives the probability distribution function of T_n as $N \to \infty$. So, the time to the first coalescent event when there are k lineages is exponentially distributed with parameter $\binom{n}{2}$

$$f_{T_k}(t) = \binom{k}{2}e^{-\binom{k}{2}t} = \frac{k(k-1)}{2}e^{-\frac{k(k-1)}{2}t} \tag{1.16}$$

Both the Moran and the WF models give rise to the same distribution for the time to the next coalescent event. The difference is in the scaling of time. One unit of time in the coalescent model is equivalent to N generations in the WF model and equivalent to $N^2/2$ time steps in the Moran model.

1.4.2 Adding mutation to the coalescent model

Under the neutral theory mutations do not affect the number of offspring an individual will have in the next generation. This means there is a separation between the genealogy and the allelic states. Genealogy can be created without any input from the allelic states. Mutations can then be imposed on top of the genealogy.

Assume in a WF population mutations happen at a constant rate of u per generation per individual. Following a lineage back in time, Equation 1.6 shows that, the number of generations until the first mutation event is geometrically distributed. Therefore the probability that there are at least

one mutation on a single lineage in the first j generations is given by:

$$\Pr(T_{\mathrm{M}} \leq j) = 1 - (1-u)^j \tag{1.17}$$

Now if we use the WF scaled time for the coalescent model, and use the change of variables $t = j/N$, $T_{\mathrm{m}} = T_{\mathrm{M}}/N$ and $\theta = 2Nu$ (where θ is called the population mutation rate) we will have

$$\Pr(T_{\mathrm{m}} \leq t) = 1 - \left(1 - \frac{\theta}{2N}\right)^{tN} \xrightarrow[N\to\infty]{} 1 - e^{\frac{\theta t}{2}} \tag{1.18}$$

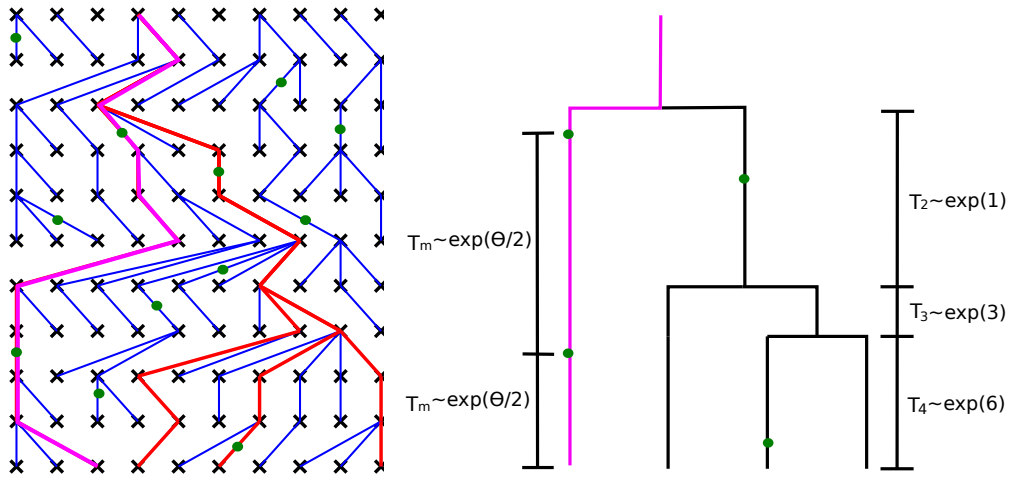
The derivative of Equation 1.18 is the probability density function of the time to the first mutation event following a lineage backwards in time which is exponentially distributed with parameter $\theta/2$:

$$f_{T_{\mathrm{m}}}(t) = \frac{\theta}{2} e^{\frac{\theta t}{2}} \tag{1.19}$$

Figure 1.8 shows the relationship between the WF model with mutations and the coalescent model with mutations. Using the relationship between the exponential distribution and the Poisson process, Equation 1.19 is equivalent to the mutations being dropped on the genealogy according to a Poisson process with rate $\frac{\theta}{2}$. Given that the length of a branch of genealogy is t, the number of mutations i on that branch is Poisson distributed:

$$\Pr(i|t) = \frac{(\frac{\theta t}{2})^i}{i!} e^{-\frac{\theta t}{2}} \tag{1.20}$$

Assuming an infinite site model, Equation 1.20 allows us to calculate the expected number of segregating sites S_n in the history of a sample of size n. Given that the total branch length of the genealogy is T_{total} and the time to



(a) WF with mutation (b) Coalescent with mutation

Figure 1.8: A WF population with mutations and the equivalent coalescent model with mutations. (a) Illustrates a WF population with mutation. Following a lineage back in time, the distribution of number of generations to the next mutation event is geometric. (b) In the coalescent model, following a lineage back in time, the distribution of time to the next event is exponential with parameter $\theta/2$. One unit of time in the coalescent model is equal to N generations in the WF model.

the next coalescent event when there are i lineages is given by T_i we have:

$$\begin{aligned}\mathbb{E}(S_n) = \mathbb{E}(\frac{\theta}{2}T_{\text{total}}) = \frac{\theta}{2}\mathbb{E}\big(\sum_{i=2}^{n}(iT_i)\big) = \frac{\theta}{2}\sum_{i=2}^{n} i\mathbb{E}(T_i) \\ = \frac{\theta}{2}\sum_{i=2}^{n}\frac{i}{\binom{i}{2}} = \frac{\theta}{2}\sum_{i=2}^{n}\frac{2i}{i(i-1)} = \theta\sum_{i=1}^{n-1}\frac{1}{i} \approx \theta\log(n-1) + 0.577\end{aligned} \tag{1.21}$$

A consequence of Equation 1.21 is that the expected number of mutations is linearly dependent on θ, but it is logarithmic on n. This means adding longer sequences to the data will increase the expected number of polymorphic sites faster than adding more sequences to the sample. In addition the expected number of segregating sites between two sequences is θ.

A commonly reported estimator of the population mutation rate is the Watterson's estimator (Watterson, 1975):

$$\hat{\theta}_W = \frac{S_n}{\sum_{i=1}^{n-1}\frac{1}{i}} \tag{1.22}$$

which uses Equation 1.21 and replaces the expected number of polymorphic sites with the observed number. Another estimator that is commonly used is Tajima's estimator (Tajima, 1989):

$$\hat{\theta}_T = \frac{2}{n(n-1)}\sum_{i<j}\pi_{ij} \tag{1.23}$$

where π_{ij} is the number of segregating sites between sequences i and j in the sample. As the number of polymorphic sites between two sequences is expected to be θ, Tajima's estimator simply replaces the expected value by the observed value, and estimates θ by averaging π_{ij} for all possible sequence pairs in the sample. Mutations that occur further up the genealogy affect

more sequences while mutations that occur on the terminal branches of the genealogy only affect a single branch. The Watterson's estimator puts equal weights on all of these mutations whereas the Tajima's estimator puts more weight on mutations further up in the genealogy as they are counted several times. TAJIMA (1989) proposed the statistic,

$$D = \frac{\hat{\theta}_T - \hat{\theta}_W}{std(\hat{\theta}_T - \hat{\theta}_W)} \tag{1.24}$$

known as the Tajima's D test which tests if sequence data fit a basic neutral (coalescent) model.

1.4.3 Adding recombination to the coalescent model

Despite the fact that bacterial recombination is not of the cross-over form, it will be introduced here as it can be extended to include gene-conversion. In addition a substantial proportion of the literature is focused on the coalescent with cross-over recombination. Following this section, we will present the coalescent with gene-conversion which is relevant to bacteria.

Cross-over recombination

HUDSON (1983) added cross-over recombination to the coalescent model. When a recombination event happens, a point on the chromosome is randomly chosen and the DNA to the right of the point comes from one parent and the DNA to the left of the point comes from the other parent. Looking back in time, when a recombination event happens, there are two ancestors for a single lineage. Consequently the entire length of sampled sequences cannot be represented by a single tree, since different sections of the sampled sequences will have different genealogies (trees). Adding

recombination to the coalescent turns the tree structure of the genealogy to a graph structure and that is where the name ancestral recombination graph (ARG) originates (Griffiths and Marjoram, 1997). This graph represents the history of the entire length of the sampled sequences, but different segments of the sequences will have different local trees which are embedded in the ARG. Figure 1.9 shows a WF population with recombination and the equivalent ARG. It also shows the local trees for different sections of the observed sequences.

For a WF population of size N, let r be the probability of recombination per generation per individual. Equation 1.9 states that following a lineage back in time, the time to the first recombination event is geometrically distributed. As a result, following a lineage back in time, the probability that there are at least one recombination in the first j generations is

$$\Pr(T_{\mathrm{R}} \leq j) = 1 - (1 - r)^j \tag{1.25}$$

Now if we use the coalescent unit of time for a WF population and do a change of variables $t = j/N$, $T_{\mathrm{r}} = T_{\mathrm{R}}/N$, and $\rho = 2Nr$, where ρ is the population recombination rate, we will have

$$\Pr(T_{\mathrm{r}} \leq t) = 1 - (1 - \frac{\rho}{2N})^{Nt} \xrightarrow[N\to\infty]{} 1 - e^{-\frac{\rho}{2}t} \tag{1.26}$$

Thus distribution of time (in coalescent unit) to the next recombination event on a single lineage converges to an exponential distribution with parameter $\rho/2$ as $N \to \infty$:

$$f_{T_{\mathrm{r}}}(t) = \frac{\rho}{2} e^{-\frac{\rho}{2}t} \tag{1.27}$$

As the lineages recombine independently, given there are k lineages

present, the time to the first recombination event is exponentially distributed with the total rate of $k\rho/2$. On the other hand the total coalescent rate when there are k lineages is $k(k-1)/2$. Hence looking back in time, the coalescent with recombination is a birth and death process where the rate of lineage birth is linearly dependent on k, and the rate of lineage death is quadratically dependent on k. This guarantees that the number of lineages will stay finite (Griffiths and Marjoram, 1997).

Both the coalescent and recombination events have exponential distributions. Using standard properties of exponential distributions one can show that while there are k lineages the time before the first event (coalescent or recombination) is exponentially distributed as:

$$f_{T_k}(t) = \left(\binom{k}{2} + \frac{k\rho}{2} \right) e^{-(\binom{k}{2} + \frac{k\rho}{2})t} \tag{1.28}$$

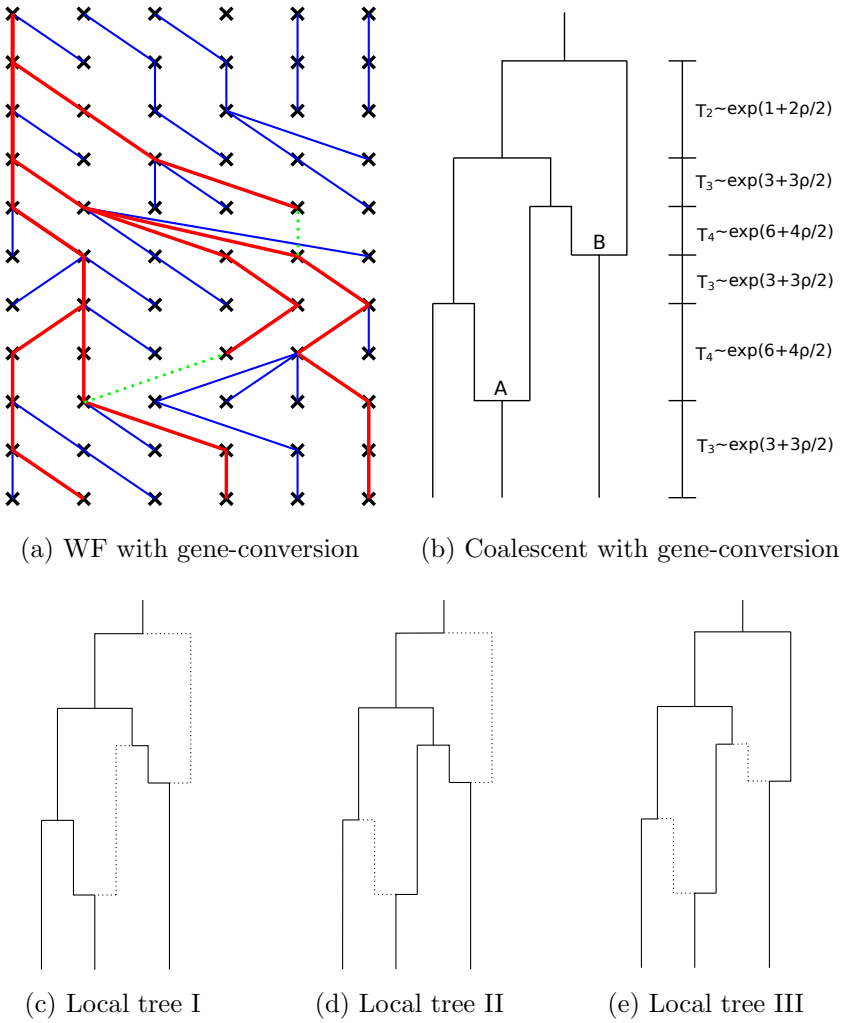
and

$$\Pr(\text{Event is recombination}) = \frac{\frac{k\rho}{2}}{\binom{k}{2} + \frac{k\rho}{2}} = \frac{\rho}{k-1+\rho} \tag{1.29}$$

$$\Pr(\text{Event is coalescent}) = \frac{\binom{k}{2}}{\binom{k}{2} + \frac{k\rho}{2}} = \frac{k-1}{k-1+\rho} \tag{1.30}$$

Gene-conversion recombination

Wiuf and Hein (2000) added gene-conversion recombination to the coalescent model. All recombinations in bacteria are of the gene-conversion form. The model is similar to the cross-over recombination introduced by Hudson (1983). The main difference is that at each cross-over recombination there is a single break point on the affected sequence, therefore the only parameter of interest is the recombination rate. However, at each gene-conversion recombination there are two break points on the affected sequence and therefore there are two parameters of interest,



(a) WF with gene-conversion (b) Coalescent with gene-conversion

(c) Local tree I (d) Local tree II (e) Local tree III

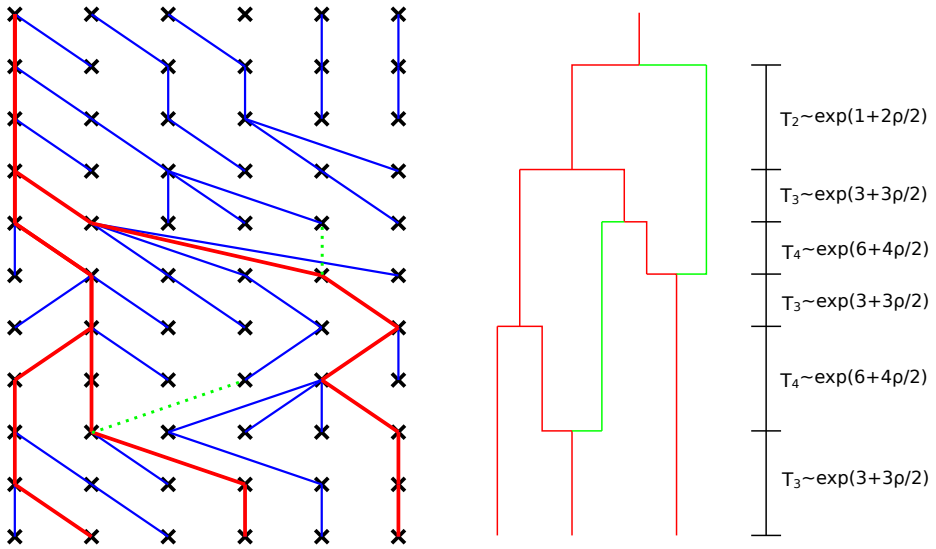
Figure 1.9: The coalescent model with cross-over recombination for a sample of size three and observed sequences of 1000 bp. (a) A WF population with recombination. At each recombination event the green dotted line shows one of the parents. The red and green lines show the genealogy of a sample of size three looking back in time. (b) The equivalent ARG and the distribution of the time to the next event. The ARG represents the history of the entire length of observed sequences. Different parts of the observed sequences will have different local trees that are embedded in the ARG. Assuming that recombination A happens at the 100th bp and recombination B happens at 700th bp, there will be three local trees that represent the history of the region. (c, d, e) The local trees for the observed sequences at 1–99, 100-699 and 700–1000 bp respectively. The solid lines show the history of the segment and the dotted lines show the rest of ARG.

the recombination rate and the mean length of recombination events. The DNA segment between the two break points (usually hundreds to thousands of nucleotides long) comes from the donor cell and the rest of the genome (usually millions of nucleotides long) comes from the recipient cell. This asymmetry allows us to define the concept of **clonal genealogy** (Guttman, 1997) for a given set of isolates by following the ancestry of the isolates back in time and following the ancestry of the recipient cells at recombination events. Parts of the core genomes that are not affected by recombination have evolved according to the clonal genealogy and parts that have been affected by recombination will have their own local trees. Figure 1.10 shows a WF population with gene-conversion recombination and its equivalent coalescent tree with the clonal genealogy highlighted in red.

Mathematically the coalescent with gene-conversion recombination is an extension to the coalescent with cross-over recombination and the equations derived in the cross-over recombination section are valid. There is one added complexity, the distribution of the recombination tract length. Recombination tract length is assumed to be geometrically distributed partly due to mathematical simplicity and partly due to empirical evidence (Falush *et al.*, 2001). If the mean recombination tract length is assumed to be δ then the length of recombination event d is assumed to be distributed as:

$$\Pr(d = i) = \delta^{-1}(1 - \delta^{-1})^{i-1} \qquad i = 1, 2, 3, \cdots \tag{1.31}$$

For a recombination event the probability that it starts at any nucleotide in the sequence except for the first one is identical and is denoted by r.



(a) WF with gene-conversion (b) Coalescent with gene-conversion

Figure 1.10: The WF and coalescent models with gene-conversion recombination. Bacterial recombinations are of gene-conversion form. A donor cell contributes a small fragment of DNA, while the recipient cell contributes the rest of the genome as shown in Figure 1.11c. If at each recombination event the line of the recipient cell is followed back in time then the clonal genealogy which is a binary tree is acquired. (a) Illustration of a WF population with gene-conversion recombination and the genealogy of a sample of size three. At each recombination event the ancestry of the donor cell is shown with a green dotted line. The red lines highlight the clonal genealogy. (b) The equivalent ancestral recombination graph. The red lines show the clonal genealogy and at each recombination event the green lines show the ancestry of the donor cell.

The probability that the event starts at the first nucleotide of an observed sequence is equal to the probability that it started d base pairs before the first observed nucleotide multiplied by the probability that it it is longer than d base pairs, and summed over all possible values of d

$$\begin{aligned} r' &= r \sum_{i=0}^{\infty} \Pr(d > i) \\ &= r \sum_{i=0}^{\infty} (1 - \delta^{-1})^i = r\delta \end{aligned} \tag{1.32}$$

Assuming that the observed sequences are L base pairs long, summing over all possible sites where recombination could occur, we have

$$r(L-1) + r' = 1 \tag{1.33}$$

so that:

$$r = \frac{1}{L - 1 + \delta} \tag{1.34}$$

$$r' = \frac{\delta}{L - 1 + \delta} \tag{1.35}$$

1.4.4 Coalescent with cross-over recombination as a point process along the genomes

Wiuf and Hein (1999) introduced an alternative algorithm for simulating genealogies under coalescent with cross-over recombination. This algorithm instead of looking at the genealogy of the sample back in time, moves along the sequences and modifies the genealogy as the recombination breakpoints are encountered. They show that the distribution of the sequence length to the next recombination event is exponentially distributed with the parameter dependent on the size of the local graph. They also show that the recombination point on the local

graph is uniformly distributed and that recombinant edges coalesce back with the local graph according to the normal coalescent probabilities.

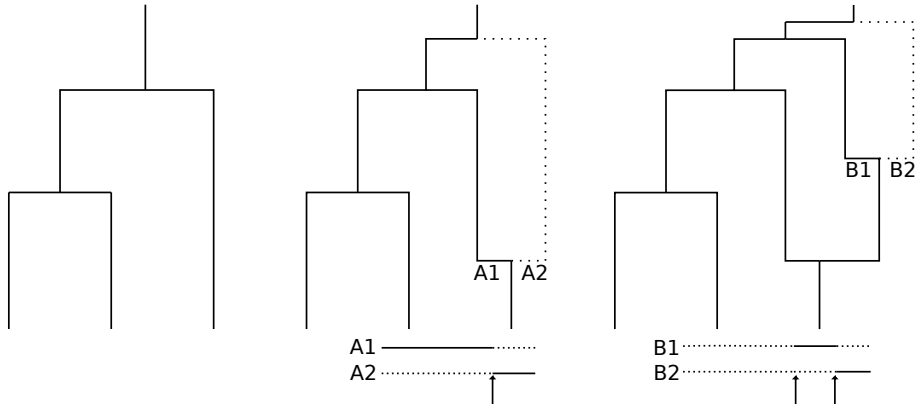
It is easy to simulate a graph from Wiuf's algorithm that has the local genealogies for each site embedded in it. It is important to note that this process is not Markovian along the sequences. With each new recombination break point encountered along the sequences, a new recombinant edge is added to the local graph (which has all the local trees up to that pint embedded in it). This local graph (rather than the local genealogy) is needed to simulate the next break point along the sequences and the new recombinant edge is added to this local graph. This means the size of the local graph is expanding as one moves along the sequences and it needs to be stored in memory which can be very memory expensive for large values of recombination rate. Figure 1.11 illustrates this algorithm.

1.5 Approximations to the coalescent

Inference under the coalescent with recombination model is difficult (Stumpf and McVean, 2003; McVean and Cardin, 2005). There are three reasons for this difficulty. Firstly, there is no way of calculating the likelihood of the data without knowing the ARG. Thus the data likelihood can only be calculated by conditioning on ARGs. Given the model parameters θ, we have:

$$\Pr(D|\theta) = \int \Pr(D|\theta, G) \Pr(G|\theta) dG \tag{1.36}$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \Pr(D|\theta, G_i) \qquad \text{and} \quad G_i \sim \Pr(G|\theta) \tag{1.37}$$



(a) Step 1 (b) Step 2 (c) Step 3

Figure 1.11: Illustration of coalescent as a point process along sequences. The length of sequences is mapped to interval [0,1). (a) Simulate the genealogy for the first position in the sequences. This is a coalescent tree. (b) Given the total length of the graph (tree), simulate the distance to the next recombination point on the sequences (The distance is exponentially distributed with parameter dependent on total length of the graph). Choose a point uniformly on the graph and add a recombination event to it. The left branch will follow the path of the existing line at that point and the right branch will coalesce at some point higher up on the graph based on the coalescent probabilities. (c) Continue until the next recombination point falls outside the length of the sequences.

However the state-space of possible ARGs is huge. Simulating sufficient number of samples from the state-space of ARGs to get a reasonable estimate for the data likelihood is not possible, except for the smallest and simplest of datasets.

Secondly, mutations provide information about the local genealogies, but recombination events break up a sequence into small regions each with its own genealogy, thus reducing the power to identify the local genealogies. This means if the mutation rate is low relative to the recombination rate, there generally will be very little information in the data about the local genealogies of the sample.

Finally, the ARG is more than the sum of the local genealogies, yet the data is only informative about the local genealogies. Given all the local trees, they can be embedded in infinitely many ARGs that are compatible with them and these ARGs can have very different likelihoods.

Attempts have been made to do inference under the coalescent with recombination model using Monte Carlo methods. Nielsen (2000) and Kuhner *et al.* (2000) developed MCMC methods to sample local genealogies and Fearnhead and Donnelly (2001) introduced an importance sampling method such that local genealogies that have higher likelihoods are sampled more often. These Monte Carlo methods work for small datasets, but are not suitable for any dataset of reasonable size.

A standard technique to solve large problems is to divide and conquer. Although the likelihood of the full data under the coalescent with recombination is difficult to compute, the data can be divided into small subsets (pairs of sites) for which likelihood can be calculated for. One can approximate the likelihood of the full data by multiplying the likelihood of the subsets of the data (Hudson, 2001; McVean *et al.*, 2002;

Fearnhead and Donnelly, 2002; Wall, 2004). These methods are called composite likelihood and are the most widely used methods in practice. As these methods ignore the dependence structure in the data, the estimated likelihood surface tends to be more peaked about their maximum than they should be. As a result obtaining reliable measures of uncertainty for the estimates is difficult.

An alternative to inference under the coalescent with recombination is to use approximations to the model, which are biologically valid and make the inference of recombination properties easier. In the next section we first introduce some of the approximations to the coalescent with cross-over recombination and then the approximations to the coalescent with gene-conversion recombination that are relevant for prokaryotes.

1.5.1 Coalescent with cross-over recombination

Product of approximate conditional (PAC) likelihoods

Li and Stephens (2003) developed a model which approximates the process of recombination while remaining computationally tractable for large genomic regions. Let $h_1, \cdots, h_n$ denote n sampled haplotypes from a population and ρ denote the population recombination rate. The model relates the distribution of the sampled haplotypes to the recombination rate, by using the following identity:

$$\Pr(h_1, \cdots, h_n|\rho) = \Pr(h_1|\rho)\Pr(h_2|h_1, \rho)\cdots\Pr(h_n|h_1, \cdots, h_{n-1}, \rho) \quad (1.38)$$

There are no exact methods to calculate the conditional likelihoods given in Equation 1.38. However, these conditional likelihoods can be approximated using models that generate a new haplotype given a set of

known haplotypes. Specifically, PAC uses a hidden Markov model in which haplotype k, is an imperfect mosaic of the previous $k-1$ haplotypes. The transition probabilities are a function of the recombination rate and the distance between the markers and the emission probabilities are a function of mutation rate and the sample size.

Informally the PAC model creates a new haplotype by copying from the existing haplotypes allowing for mutations such that segments derived from the same ancestor may differ. Under the PAC model the likelihood of the data can be calculated as a sum over all possible histories that could give rise to the data. This summation can be done quickly using dynamic programming algorithms (forward algorithm).

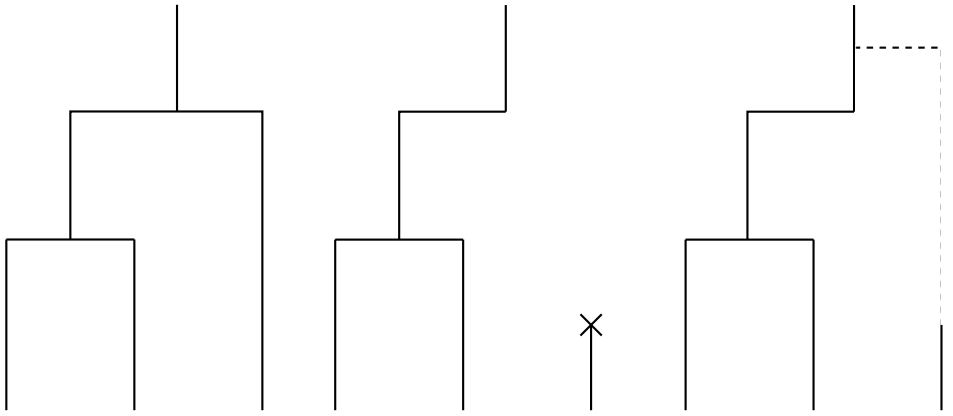
The PAC model captures some aspects of the coalescent with recombination (Li and Stephens, 2003) although it is not a genealogical model. These include how haplotypes are related, the likelihood of seeing new haplotypes decreases as the sample size increases and as the mutation rate increases the likelihood of seeing new haplotypes increases. However there are issues with the PAC model. The main issue is that the approximation introduced in calculating the conditional likelihoods depends on the order in which the haplotypes are used. Thus different orders lead to different likelihood estimates. To solve this problem the authors suggest averaging over a number of runs with different orders. Despite this, the parameter estimates are biased and the bias is difficult to model. It is important to remember that the advantage of the PAC model is its computational tractability, so that it can be used for large datasets. For instance the PAC model has been applied to thousands of human genomes to produce a coancestry matrix which is then clustered to detect population structures in humans (Lawson *et al.*, 2012).

Sequentially Markov coalescent (SMC)

McVean and Cardin (2005) introduced the sequentially Markov coalescent model. It makes a simplification to the ARG model where two ancestral lineages that have no ancestral material in common are not allowed to coalesce. This change makes the state-space of ARGs smaller, leads to fewer recombination events in the history of the sample and results in a Markovian process when sequentially generating genealogies along the sequences.

This modification is best understood when comparing it to the Wiuf's spatial algorithm (Wiuf and Hein, 1999) illustrated in Figure 1.11. In spatial algorithm when sampling genealogies along the sequences, with each recombination event, a new recombinant edge is added to the local graph. This local graph is needed for simulating the next recombination event. However under the SMC model with each new recombination event, before adding the new edge corresponding to the recombination event, the existing ancestry of the branch on which the recombination is occurring (above the recombination point) is deleted. This results in a floating lineage which coalesces back with the rest of the tree according to the coalescent probabilities. This process leads to new sampled genealogy which is a tree rather than a graph. Only this new local genealogy (tree) is needed for simulating the next recombination event along the sequences.

The authors show that there is little difference in the LD patterns when data is simulated under the SMC and the ARG models. Figure 1.12 illustrates how a recombination event affects the previous local tree in the SMC model. The SMC model is used in simulation software (Marjoram and Wall, 2006; Chen *et al.*, 2009) although it has not been used in inference procedure. Next we introduce approximations to the coalescent



(a) Step 1 (b) Step 2 (c) Step 3

Figure 1.12: Illustration of the SMC algorithm. (a) Simulate the genealogy for the first position in the sequences according to the coalescent probabilities (b) Given the total length of the previous tree, simulate the distance to the next recombination point on the sequences. Choose a point uniformly on the tree for the recombination event. Remove the branch above that point (part of the line which is between the point and where it coalesces with the tree). (c) Let the hanging branch to coalesce with the tree at some higher point according to the coalescent probabilities.

with gene-conversion that is appropriate for bacteria.

1.5.2 Coalescent with gene-conversion recombination

The ClonalFrame model

Didelot and Falush (2007) developed the ClonalFrame model and software. The ClonalFrame model is a very rough approximation of the coalescent with gene-conversion recombination. An important concept in bacterial recombination is the concept of **clonal genealogy**, see section 1.3.1. The model assumes that given the clonal genealogy, recombination events are independent of each other. In addition the ancestry of the recombination event on the ARG is not modelled. Instead it is assumed that recombination events introduce novel substitutions to the sequences

at a constant rate.

Figure 1.13 illustrates the ClonalFrame model. Recombination events happen as a Poisson process on the branches of the clonal genealogy and introduce substitutions at a uniform rate to the affected sequences. The length of the recombination events is modelled by a geometric distribution. Inference under the ClonalFrame model is done in a Bayesian context and MCMC is used to get a posterior distribution for the model parameters including the clonal genealogy. Although the ancestry of the recombination events is not modelled, it is possible to post process the output of the ClonalFrame and use BLAST to detect where the recombinant segments are coming from and to identify putative donors (Didelot *et al.*, 2009a).

The ClonalFrame model can be seen as an approximation to the coalescent with recombination where recombination is coming from outside of the population. Thus it tends to underestimate the recombination rate relative to mutation rate as it will miss recombination events that do not introduce many polymorphisms. Another issue with ClonalFrame is that it does not use all the information available in the data, namely it does not take into account homoplasy. A site is called homoplasic when given a tree, it could not have arisen without either recombination or repeat mutation (Maynard Smith and Smith, 1998; Maynard Smith, 1999). ClonalFrame was developed for multilocus sequence typing (MLST) data (Maiden *et al.*, 1998) which are usually around 3000-4000 nucleotides long. Therefore the method does not scale well for large datasets of whole bacterial genomes. For instance for 30 bacterial whole genomes with average diversity of around 5% it can take weeks to get results from the software.

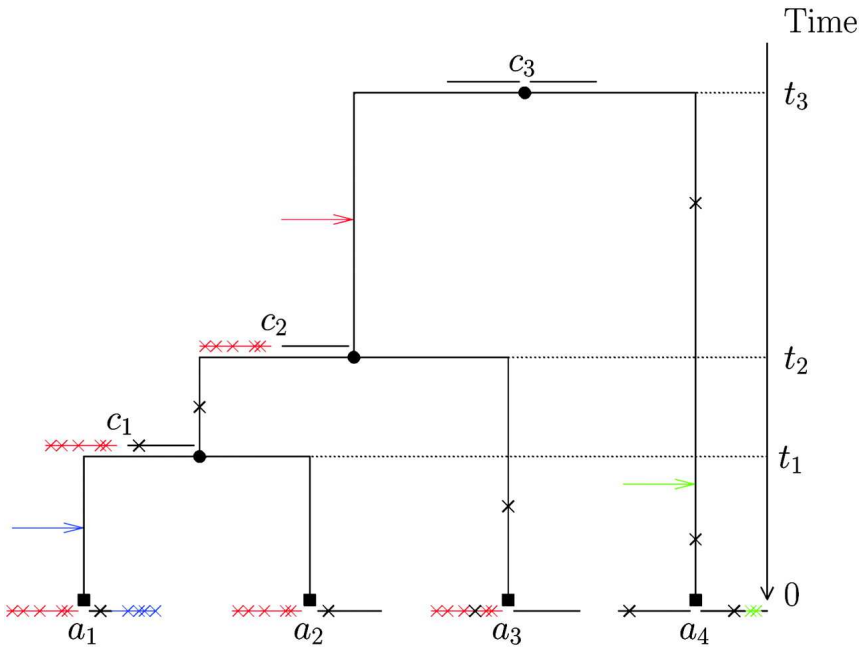


Figure 1.13: Illustration of the ClonalFrame model. a_1, a_2, a_3 and a_4 show the four sampled sequences. c_1, c_2 and c_3 show the ancestral sequences. Recombination events are shown as arrows and mutation events are shown as black crosses. ClonalFrame model does not model the source of recombination events. Each recombination event affects a segment of the observed sequences and introduces a number of substitutions to the segment with a constant rate. Mutations affect a single site. In this instance there are three recombination events where the sources are not known and four mutation events. Figure reproduced with permission from (Didelot and Falush, 2007).

The ClonalOrigin model

Didelot *et al.* (2010) introduced the ClonalOrigin model and software. In comparison to ClonalFrame it is a better approximation of the coalescent with gene-conversion recombination. Unlike ClonalFrame, ClonalOrigin models the origin of recombination events as points on the clonal genealogy.

ClonalOrigin model assume that given the clonal genealogy, recombination events are independent of each other. This has two effects on the ARG. Firstly, the recombinant edges are not allowed to coalesce with each other, they are only allowed to coalesce with the clonal genealogy. Secondly, the recombinant edges are not allowed to recombine.

The assumption of independence of recombination events given the clonal genealogy can be examined by its two effects on the ARG. The first simplification follows the same logic as the SMC algorithm, which is an approximation of the coalescent with cross-over recombination (McVean and Cardin, 2005). Lines that are not part of the clonal genealogy (recombinant edges) carry little ancestral material and if two such lines coalesce, they are unlikely to carry any overlapping ancestral material. This change makes the model Markovian along the sequence and thus simplifies the inference. McVean and Cardin (2005) showed that ignoring these cases has little effect on the patterns of LD relative to the full ARG model. The second simplification is justified, as the lines that are not part of the clonal genealogy do not carry much ancestral material and therefore it is unlikely for those small ancestral material to be affected by recombinations.

Therefore recombination events can be seen as edges that start and end on the clonal genealogy. Figure 1.14 illustrates the model. The clonal

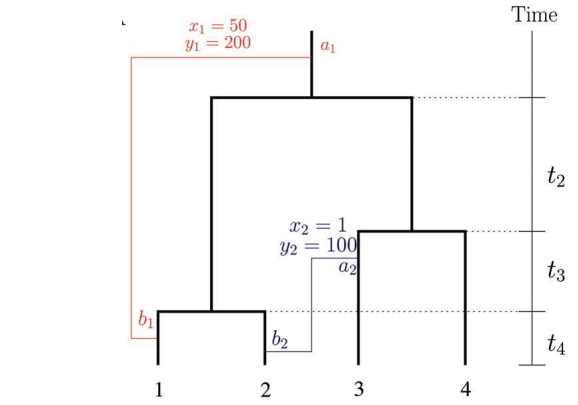
genealogy is shown in solid black lines and the two recombinant edges are shown using blue and red lines. At each recombination event the ancestry of the donor cell is followed back in time to its common ancestor on the clonal genealogy. The local tree for any site is given by following back the correct lines of ancestry (on the clonal genealogy and the recombinant edges) for that site.

As our work is based on the ClonalOrigin model, the details of the model are presented. Let $\mathcal{T}$ denote the clonal genealogy which is a coalescent tree and T denote the total branch length of the clonal genealogy. Each recombination event is an edge superimposed on this tree. Recombination events occur independently on the branches of the clonal genealogy at a constant rate of $\rho/2$ where ρ is the population recombination rate. The total number of recombination events R on the clonal genealogy is Poisson distributed as:

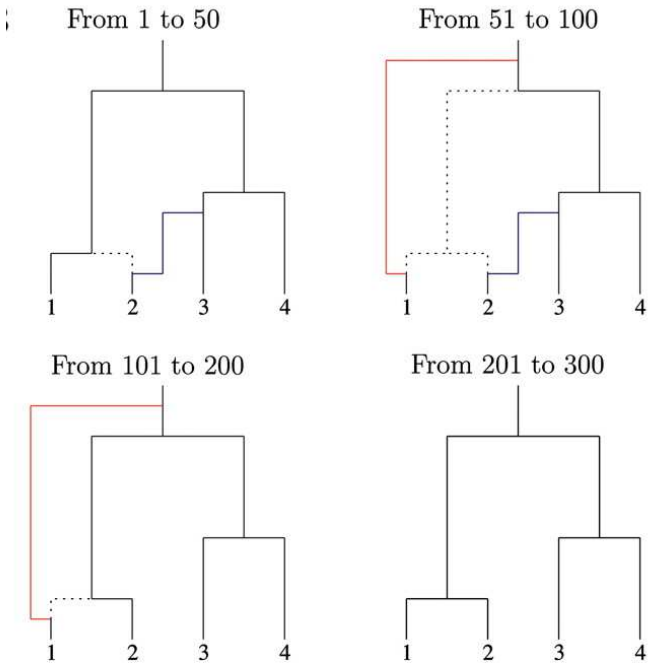
$$\Pr(R = r|T, \rho) = \frac{\left(\frac{\rho T}{2}\right)^r e^{-(\frac{\rho T}{2})}}{r!} \tag{1.39}$$

Each recombination event i where $i = 1, \ldots, R$ is identified by four variables. An arrival point on the clonal genealogy b_i where the recombination occurs, a departure point on the clonal genealogy a_i where the ancestry of the donor cell meets the clonal genealogy, the site on the chromosome where recombination starts x_i and the site on the chromosome where recombination event ends y_i.

It is important to note that to identify a_i and b_i we need to specify the time they occur and the branch on the clonal genealogy they occur on. The arrival points are uniformly distributed on the branches of the clonal genealogy as recombination happens at a constant rate on the clonal



(a) Ancestral graph



(b) Local trees

Figure 1.14: Illustration of the ClonalOrigin model. (a) Ancestral graph. The black lines show the clonal genealogy and colour lines show the ancestry of the donor cell for recombination events. The red recombination event affects the ancestor of sample 1 from base pair 50 to 200 and the donor shares ancestry with the clonal genealogy at point a_1. The blue recombination event affects the ancestor of sequence 2 from base pair 1 to 100 and it shares a common ancestor with the clonal genealogy at point a_2. (b) The two recombination events have created 4 regions where the local trees are different. The figure shows the local trees for each of the four regions. Figure reproduced with permission from (Didelot *et al.*, 2010).

genealogy:

$$f(b_i|\mathcal{T}) = \frac{1}{T} \qquad b_i \in [0, T] \tag{1.40}$$

Given b_i, the recombinant edge reconnects with the clonal genealogy according to the coalescent probabilities i.e. the time for the recombinant edge to coalesce with a single lineage is exponentially distributed with parameter 1. This means the departure point a_i given b_i is exponentially distributed as:

$$f(a_i|b_i, \mathcal{T}) = e^{-L(a_i,b_i)} \tag{1.41}$$

where $L(a_i, b_i)$ is the sum of the branch lengths of the clonal genealogy between a_i and b_i.

Furthermore, the model assumes that recombination events are uniformly distributed along the sequences and their length is geometrically distributed with mean δ. Sequence data is made of B independent blocks with total length L. A recombination event may start before a block and be long enough to affect the beginning of the block, so that the probability of observing the recombination start at the beginning of the block is δ times greater than within the block and the normalising factor would be $\delta B + L - B$. If s denotes the location of the nucleotide, then the

distribution of x_i and y_i are:

$$\Pr(x_i = s|\delta) = \begin{cases} \frac{\delta}{L-B+\delta B} & \text{if } s \text{ is at the start of the sequence} \\ \frac{1}{L-B+\delta B} & \text{otherwise} \end{cases} \tag{1.42}$$

$$\Pr(y_i = s|x_i, \delta) = \begin{cases} \delta^{-1}(1-\delta^{-1})^{s-x_i} & \text{if } s \text{ is before the end of the sequence} \\ (1-\delta^{-1})^{s-x_i+1} & \text{if } s \text{ is the end of the sequence} \end{cases} \tag{1.43}$$

Mutations are assumed to occur according to a Poisson process with rate $\theta/2$ on the branches of the clonal genealogy and the recombinant edges. Let M be the total number of mutations on a local tree with sum of branch lengths T_i. M is distributed as:

$$\Pr(M = m|T_i, \theta) = \frac{\left(\frac{\theta T_i}{2}\right)^m e^{-\frac{\theta T_i}{2}}}{m!} \tag{1.44}$$

It is important to note that the ClonalOrigin model is a genealogical approximation to the coalescent with gene-conversion recombination which aims to make inference simpler. Despite the simplified model, full Bayesian inference remains difficult and inferring all the parameters at the same time impractical (Didelot *et al.*, 2010). The model parameters are inferred in three steps. First, the clonal genealogy is inferred using ClonalFrame. Second, ρ, θ and δ are inferred for blocks of the core genome given the clonal genealogy. Third, given the clonal genealogy and the mean of ρ, θ and δ from the previous steps, the recombination events are inferred independently for each block in parallel. The inference for 13 *Bacillus cereus* genomes took several weeks using several hundred processors.

Chapter 2

Inference

2.1 Bayesian inference

The Bayesian paradigm models unknown parameters as random variables. This provides a cohesive and intuitive framework for combining complex models and external knowledge. In Bayesian settings we specify the likelihood $p(D|\theta)$ for the observed data D given the unknown parameter θ and assume that θ is a random variable sampled from a known prior distribution $p(\theta)$. The product of these two distributions defines the joint distribution, $p(D, \theta)$. Given the joint distribution, one can use Bayes' theorem to get the distribution of $\theta|D$ which is called the posterior distribution of the model parameters. Inference concerning θ is based on its posterior distribution:

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)} \\ &= \frac{p(D|\theta)p(\theta)}{\int_\theta p(D|\theta)p(\theta)d\theta} \qquad (2.1) \\ &\propto p(D|\theta)p(\theta) \qquad (2.2) \end{aligned}$$

In many realistic problems Equation 2.1 does not have a closed form solution. There are two underlying reasons for this problem. The first is that the integral in the denominator of equation 2.1 which is called the marginal likelihood may not be tractable. The second is that the likelihood function $p(D|\theta)$ may be intractable.

To solve the first problem we can use Monte Carlo methods to sample from the posterior distribution. These methods do not require the marginal likelihood to be known. A sample from the posterior distribution is usually sufficient for reliable inference. In addition an estimate can be made arbitrarily accurate just by increasing the sample size. These methods include rejection sampling, importance sampling and MCMC sampling.

However solving the second problem is harder. The last decade has seen the rise of approximate Bayesian computation (ABC) for addressing problems where the model likelihood $p(D|\theta)$ is intractable. The main reason for the rise of ABC is the ubiquity of cheap and powerful computers, as the essence of ABC is simulation. In ABC inference, one simulates from the model and then compares the simulated data to observed data. The parameter values that give rise to simulated data which are similar to observed data are said to come from the posterior distribution. For real problems we have to make several approximations that will be discussed in the following sections.

The methods described in this chapter will be used in the following chapters. We first introduce Monte Carlo methods for sampling from a target distribution and then introduce ABC methods for sampling from the posterior distribution where the model likelihood is intractable.

2.2 Monte Carlo methods

Although there are earlier versions of Monte Carlo method such as Buffon's needle problem, the term Monte Carlo (MC) methods was introduced by Metropolis and Ulam (1949). Monte Carlo methods generally refers to estimating quantities by repeated random sampling, although more specifically it can be seen as a framework to sample from a distribution. The sample approximates the distribution and can be used for estimation. For instance, to compute the expectation of the posterior distribution, we can use the MC approximation:

$$\begin{aligned}\mathbb{E}(\theta|D) &= \int_{\theta} \theta p(\theta|D) d\theta \\ &\approx \frac{1}{n}\sum_{i=1}^{n} \theta_i \qquad \text{and} \quad \theta_i \sim p(\theta|D) \end{aligned} \tag{2.3}$$

The law of large numbers guarantees the convergence of $1/n \sum_{i=1}^{n} \theta_i \rightarrow \mathbb{E}(\theta|D)$ as $n \rightarrow \infty$. The central limit theorem gives the rate of convergence for the above MC estimation:

$$\sqrt{n}(\frac{1/n \sum_{i=1}^{n} \theta_i - \mathbb{E}(\theta|D)}{\sigma}) \rightarrow \mathrm{N}(0,1) \qquad \text{as} \qquad n \rightarrow \infty. \tag{2.4}$$

where σ^2 is the variance of the posterior distribution $p(\theta|D)$ which can be estimated using sample variance. In other words the standard deviation of $1/n \sum_{i=1}^{n} \theta_i$ is $\sigma/\sqrt{n}$. Therefore to get an answer that is twice more accurate, we have to simulate four times more and the variance of the estimate goes to zero as $n \rightarrow \infty$.

Thus the question is how do we sample from a distribution where the normalising constant is not known as is the case for the posterior distribution in most Bayesian inference settings. In the following sections, we introduce

several methods of sampling from a target distribution when the normalising constant is not known. We start with rejection sampling and then introduce importance sampling. Both of these methods can be seen as producing independently and identically distributed (IID) samples. At the end of the section we introduce MCMC sampling which is one of the most widely used algorithm in science and produces correlated samples from a distribution.

2.2.1 Rejection sampling

In the literature this method is referred to by different names including, acceptance-rejection and accept-reject sampling. In this method, given that sampling from distribution g (target distribution) is difficult, we sample from another distribution f (proposal distribution) and reject some of these samples such that the introduced bias leads to values sampled from g.

Given $g(x)$ is the density of the target distribution and $f(x)$ is the density of the proposal distribution they have to satisfy three conditions for rejection sampling to work. Firstly, we should be able to sample from distribution f. Secondly, we can calculate the function $g(x)/f(x)$ and finally $g(x) \leq cf(x)$ for all x in the support of distribution g, where c is some finite constant. Rejection sampling is performed using Algorithm 1.

To prove that the Algorithm 1 produces samples distributed according

Algorithm 1 Rejection sampling

1: $t = 1$.
2: **while** $t \leq n$ **do**
3: Sample a candidate point x' from proposal distribution f.
4: Generate a random number $U \sim \text{Unif}[0, 1]$.
5: **if** $U \leq g(x')/(cf(x'))$ **then**
6: $x_t = x'$ and $t = t + 1$.
7: **end if**
8: **end while**

to density g, we show that the CDF of random variable X conditional on X being accepted is G. Let X and U be independent random variables where X has density $f(x)$ and $U \sim \text{Unif}(0,1)$.

$$\begin{aligned}
\Pr(X \leq y | U \leq \frac{g(X)}{cf(X)}) &= \frac{\Pr(X \leq y, U \leq \frac{g(X)}{cf(X)})}{\Pr(U \leq \frac{g(X)}{cf(X)})} \\
&= \frac{\int_{-\infty}^{y} \int_{0}^{g(x)/cf(x)} f(x) du dx}{\int_{-\infty}^{\infty} \int_{0}^{g(x)/cf(x)} f(x) du dx} \\
&= \frac{\frac{1}{c}\int_{-\infty}^{y} g(x) dx}{\frac{1}{c}\int_{-\infty}^{\infty} g(x) dx} \\
&= \int_{-\infty}^{y} g(x) dx \\
&= G(y) \qquad (2.5)
\end{aligned}$$

Although the above proof used properly normalised densities, it can be shown that the algorithm works when the target density $g(x)$ is known up to a constant of proportionality. On a more intuitive level rejection sampling can be justified as follows: the proposed samples have density $f(x)$ and they are accepted with probability $M(x)$ thus the accepted ones are distributed with density $f(x)M(x)$. Now if we set $M(x) = g(x)/f(x)$, we have $f(x)M(x) \propto g(x)$ and therefore the accepted samples are distributed with density $g(x)$.

There are two main issues with this method. First, it can be difficult to find a distribution with density $f(x)$ which is easy to sample from and such that $g(x) \leq cf(x)$ for all x. Second, the probability of a sample being accepted is $1/c$ and for large values of c (common in high dimensional problems) the algorithm becomes inefficient.

Algorithm 2 Importance sampling

1: **for** $i = 1, \ldots, n$ **do**
2: Sample a point x_i from distribution f.
3: Set $w(x_i) = g(x_i)/(f(x_i))$.
4: **end for**
5: Return $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} w(x_i)h(x_i)$

2.2.2 Importance sampling algorithm

An alternative to the rejection sampling algorithm is the importance sampling algorithm. Instead of accepting candidates from the proposal distribution with a given probability, all the candidates are accepted, but are weighted to correct for the discrepancy between the proposal and the target distributions. In contrast to the rejection sampling algorithm, the importance sampling algorithm does not return a sample from the target distribution, but a weighted sample that can be used for calculating expectations and probabilities.

Suppose we would like to estimate $\mu = \mathbb{E}(h(X))$ where X has density $g(x)$. Let $f(x)$ be another distribution defined on the same support for X. Then

$$\begin{aligned} \mu = \mathbb{E}(h(X)) &= \int_{-\infty}^{\infty} h(x)g(x)dx = \int_{-\infty}^{\infty} \frac{h(x)g(x)}{f(x)} f(x)dx \\ &\approx \frac{1}{n}\sum_{i=1}^{n} \frac{h(x_i)g(x_i)}{f(x_i)} \qquad x_i \sim f \end{aligned} \tag{2.6}$$

Where g is the target distribution and f is the proposal distribution. The importance sampling is performed using Algorithm 2.

The estimator of Equation 2.6 is unbiased and law of large numbers guarantees that it converges to $\mathbb{E}(h(X))$ as $n \to \infty$. The estimator of Equation 2.6 has finite variance if the importance weights have finite variance which is an important condition to check. However in practice the

above estimator usually cannot be used as the normalising constant of $g(x)$ and $f(x)$ need to be known. If the normalising constant of g or f or both are not known, we can use a normalised version of Equation 2.6 which is biased, but asymptotically consistent:

$$\mathbb{E}(h(X)) \approx \frac{\sum_{i=1}^{n} \frac{g(x_i)}{f(x_i)} h(x_i)}{\sum_{i=1}^{n} \frac{g(x_i)}{f(x_i)}} \qquad x_i \sim f \tag{2.7}$$

Simulation studies show that estimator of Equation 2.7 is more stable than estimator of Equation 2.6 (Ripley, 2009). Therefore it might be better to use Equation 2.6 even when the normalising constants can be calculated.

Importance sampling has two drawbacks. First, for the importance sampling estimator to have a finite variance, the variance of the weights has to be bounded which may be difficult to guarantee. In addition Importance sampling works best when the shape of the proposal distribution is close to target distribution. For high dimensional problems finding a suitable proposal distribution $f(x)$ is difficult and thus the algorithm becomes very inefficient.

2.2.3 Markov chain Monte Carlo (MCMC) sampling

Rejection sampling and importance sampling produce IID samples for which the central limit theorem and the law of large numbers are valid and can be used to estimate distribution of the sample mean. In many cases the target distribution is too complex to generate IID samples from it. In such cases MCMC algorithm allows us to simulate correlated samples from the distribution.

A Markov chain is a time discrete stochastic process such that the future is conditionally independent of the past given the present. More formally the

Markov property for a time discrete and finite state homogeneous Markov chain is defined as follows:

$$\Pr(X_{t+1} = j | X_t = i, \ldots, X_0 = k) = \Pr(X_{t+1} = j | X_t = i) = p_{ij} \tag{2.8}$$

And we can define the transition matrix P as the matrix where element (i, j) is the probability (p_{ij}) of going from state i to state j. It can be shown that for an irreducible (possible to get from any state to any other state with positive probability) and aperiodic chain there is a unique stationary distribution. A distribution π is a stationary distribution if:

$$\pi = \pi P \tag{2.9}$$

which indicates that if the chain is distributed according to π then distribution of the chain does not change with time. A Markov chain with transition matrix P is reversible if there is a probability vector S such that:

$$s_i p_{ij} = s_j p_{ji} \qquad \text{for all states } i, j \tag{2.10}$$

Intuitively reversibility means that if the chain starts at some distribution and then it evolves, if then it is played back and shown to someone else they would not know the sequence has been reversed. Equation 2.10 also is known as the detailed balance equation due to its origin in statistical physics literature.

It can be proven that if a chain is reversible with respect to a probability vector S then S is the stationary distribution of that Markov chain. Reversibility is important because when we construct an MCMC, it is much easier to show that the Markov chain is reversible with respect to

the target distribution. As a result the target distribution is the stationary distribution of the Markov chain. It can also be proven that any irreducible and aperiodic Markov chain converges to its stationary distribution as $n \to \infty$.

The main idea behind the MCMC algorithm is to simulate a Markov Chain in the state space S such that the stationary distribution of the chain is the target distribution. This is in contrast to the usual Markov chain analysis, where the transition matrix is known and one is interested in estimating the stationary distribution. In MCMC simulation, we know the stationary distribution and would like to find an efficient transition matrix, so that the Markov chain simulation reaches its stationary distribution quickly.

Metropolis *et al.* (1953) introduced the first algorithm for Markov Chain Monte Carlo which is extremely simple and powerful. This algorithm later known as Metropolis algorithm and its many variations introduced since, theoretically can be used to generate correlated random samples from almost any target distribution, no matter how complex.

Metropolis sampling from target distribution $g(x)$ is performed using Algorithm 3.

It is easily shown that Algorithm 3 leads to a Markov chain which is reversible with respect to g and therefore g is its stationary distribution. As a result its distribution will converge to g as $n \to \infty$. Note that the algorithm does not need the normalising constant of the target distribution $g(x)$, as it would cancel out in step 5 Equation.

Hastings (1970) extended the metropolis algorithm so that any proposal function can be used rather than symmetric ones. The Metropolis-Hastings sampling from target distribution $g(x)$ is performed

Algorithm 3 Markov chain Monte Carlo (Metropolis)

1: Choose a starting point x_0.
2: **for** $i = 1, \ldots, n$ **do**
3: Propose a random unbiased perturbation to the current state x_t, to generate a new state x' using a symmetric proposal density q i.e. $q(x'|x_t) = q(x_t|x')$.
4: Generate a random number $U \sim \text{Unif}[0, 1]$.
5: **if** $U \leq \frac{g(x')}{g(x_t)}$ **then**
6: $x_{t+1} = x'$
7: **else**
8: $x_{t+1} = x_t$
9: **end if**
10: **end for**

Algorithm 4 Markov chain Monte Carlo (Metropolis-Hastings)

1: Choose a starting point x_0.
2: **for** $i = 1, \ldots, n$ **do**
3: Propose a random perturbation to the current state x_t, to generate a new state x' using proposal density q.
4: Generate a random number $U \sim \text{Unif}[0, 1]$.
5: **if** $U \leq \frac{g(x')q(x_t|x')}{g(x_t)(q(x'|x_t)}$ **then**
6: $x_{t+1} = x'$
7: **else**
8: $x_{t+1} = x_t$
9: **end if**
10: **end for**

using Algorithm 4.

It can be shown that Algorithm 4 produces a Markov chain which is reversible with respect to g and therefore g is its stationary distribution. It is important to note that the simulated samples are correlated and if the proposal is not chosen well, the chain becomes sticky (the samples will be highly correlated) and therefore to have a representative sample from the target distribution, the chain has to run for longer. The ergodic theorem (law of large numbers for Markov chains) states that if the Markov chain $x_1, x_2, \cdots$ simulated by M-H algorithm is irreducible and aperiodic, then

for any bounded function f:

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \rightarrow \int_{\text{All } x} f(x)g(x)dx \tag{2.11}$$

with probability 1, as $n \rightarrow \infty$.

One of the main drawbacks of the MCMC algorithm is that despite much theory about the convergence of the Markov chain, in practice it can be difficult to know if the chain has reached its stationary distribution. Another issue is that if the distribution is multi-modal, the chain may get stuck in one of the modes without exploring the whole state space.

2.3 Approximate Bayesian computation

Monte Carlo methods are a general framework for sampling from a distribution. Most of these methods can deal with unknown normalising constants which is the case for the posterior distribution in most realistic Bayesian inference settings. Another problem that very often arises in complex settings is when simulation from the model is possible, but the likelihood function does not have an analytic form. This frequently happens in population genetics.

ABC methods can sample from the posterior distribution when simulation from the model is easy, but no analytical form is present for the likelihood function. ABC methods are rapidly gaining popularity as they allow us to use Bayesian inference for problems that otherwise would be intractable. In the following sections we first introduce rejection ABC. Next we present ABC-MCMC which improves on the acceptance ratio of the rejection ABC. In the final section sequential ABC is introduced which is a form of sequential importance sampler.

Algorithm 5 Rejection ABC 1

1: $t = 1$.
2: **while** $t \leq n$ **do**
3: Generate θ' from the prior distribution $p(\theta)$.
4: Simulate x' from the model $p(x|\theta')$.
5: **if** $x' = D$ **then**
6: $x_t = x'$ and $t = t + 1$.
7: **end if**
8: **end while**

2.3.1 Rejection ABC

The simulation procedure used within ABC can be seen as a data augmentation procedure. In this context we augment the posterior distribution $p(\theta|D) \propto p(D|\theta)p(\theta)$ such that:

$$p(\theta, x|D) \propto p(D|x, \theta)p(x|\theta)p(\theta) \tag{2.12}$$

where D is the observed data and the auxiliary parameter x is simulated data given θ. As $p(D|x, \theta)$ is not known, approximations of it can be used. These approximations are usually chosen such that the regions in the posterior where x and D are similar, have higher density. To get the marginal posterior density which we are interested in we integrate out x:

$$p_{\text{ABC}}(\theta|D) \propto p(\theta) \int_x p(D|x, \theta)p(x|\theta)dx \tag{2.13}$$

The simplest case is to set $p(D|x, \theta)$ to one when $D = x$ and zero otherwise. One of the earliest versions of ABC was introduced by Tavaré *et al.* (1997) given in Algorithm 5. Given that $p(D|x, \theta)$ is a point mass at $x = D$ and zero elsewhere, Equation 2.13 gives exactly the posterior distribution. Therefore the accepted observations from Algorithm 5 are independent draws from the posterior distribution. The only

approximation present in algorithm 5 is the Monte Carlo approximation due to using a finite number of samples. The issue is that for anything apart from the simplest cases, the probability of simulating $x = D$ is zero.

Pritchard *et al.* (1999) introduced the first version of ABC in a population genetics setting which can be applied to real problems. In practice two modifications are made to $p(D|x, \theta)$ to increase the probability of acceptance of x that are close to D. These modifications mean that $p_{\text{ABC}}(\theta|D)$ in Equation 2.13 is an approximation to the true posterior distribution. Firstly, we allow $p(D|x, \theta)$ to be an indicator function on set A where $A = \{x|\ |x - D| \leq \epsilon\}$:

$$p(D|x, \theta) = \mathbb{1}_{A_\epsilon}(x) \tag{2.14}$$

This Kernel puts the weights on the regions of the posterior where $x \approx D$. In many real situations the data are high dimensional such that comparing two datasets becomes meaningless. To address this problem we add the second modification to $p(D|x, \theta)$. This concession allows the comparison of the datasets to be performed through a low dimensional vector of summary statistics $S(x)$ where $\dim(S(x)) \geq \dim(\theta)$. Therefore $p(D|x, \theta)$ is approximated by:

$$p(D|x, \theta) = \mathbb{1}_{B_{\epsilon,K}}(x) \tag{2.15}$$

where $B = \{x|K(S(x) - S(D)) \leq \epsilon\}$. This kernel puts larger weight on regions in the posterior when $S(x) \approx S(D)$. If summary statistic S is sufficient for θ then this modification will not add any further approximation in estimating $p(\theta|D)$. However for most realistic problems there are no sufficient summary statistics and therefore we have an added approximation due to the loss of information by using summary statistics.

Algorithm 6 Rejection ABC 2

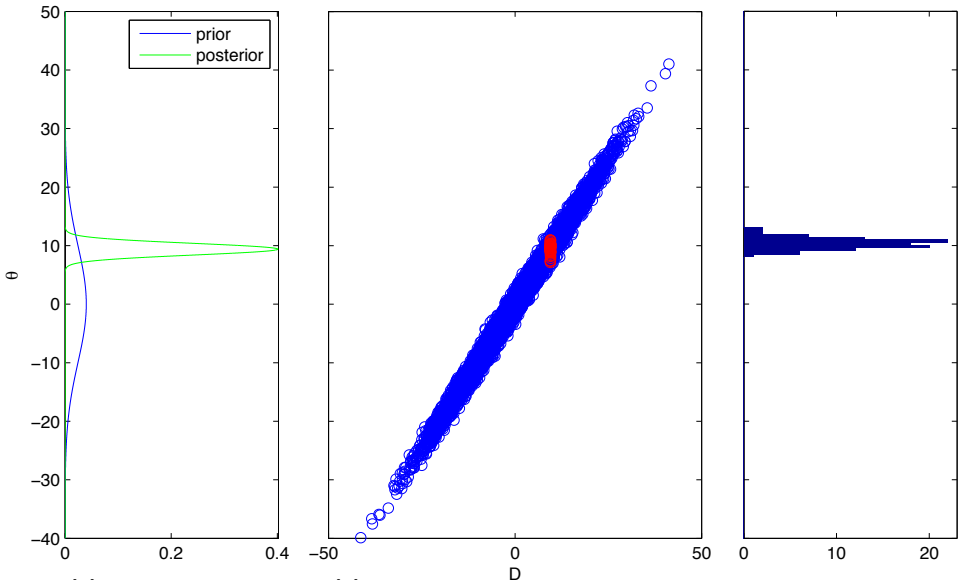
1: $t = 1$.
2: **while** $t \leq n$ **do**
3: Generate θ' from the prior distribution $p(\theta)$.
4: Simulate x' from the model $p(x|\theta')$.
5: **if** $K(S(x) - S(D)) \leq \epsilon$ **then**
6: $x_t = x'$ and $t = t + 1$.
7: **end if**
8: **end while**

Therefore the new method with both modifications is given by Algorithm 6. The most widely used distance metric K is Euclidean distance. The choice of ϵ is related to the trade-off between computability and accuracy. Reducing the bandwidth ϵ will decrease the approximation introduced by it, but at the expense of reducing the acceptance ratio. Figure 2.1 shows an implementation of Algorithm 6 for a toy example. Figure 2.2 shows for the same toy example how as ϵ becomes smaller the ABC estimate of the posterior becomes more accurate, at the expense of more simulation.

The advantages of the rejection ABC algorithm are that it is easy to code and can use parallel processing as samples are independent of each other. In addition, ABC usually leads to reasonable results where no other method is able to provide answers. The disadvantages are that it can be hard to know how the summary statistics affect the ABC posterior, if they are not sufficient. Furthermore the choice of the distance metric K and the bandwidth ϵ will also affect the approximation.

2.3.2 ABC-MCMC

The rejection ABC uses independent samples from the prior which can lead to an unacceptable rejection ratio if the posterior is highly peaked relative to the prior. Marjoram *et al.* (2003) introduced an MCMC version of the



(a) Posterior (b) Simulation from model (c) ABC posterior

Figure 2.1: ABC toy example. We have a single data point $D = 10$ that comes from a normal distribution with unknown mean θ and a known variance of 1 so that $D \sim \mathcal{N}(\theta, 1)$. If we use a prior distribution on θ that is normal $p(\theta) = \mathcal{N}(0, 10)$ then the posterior will be a normal distribution $p(\theta|D) \sim \mathcal{N}(9.9, 0.99)$. (a) The prior and the posterior distributions. (b) Histogram of the samples from the ABC estimate of the posterior distribution with $\epsilon = 0.1$. (c) The ABC algorithm. Y-axis is the parameter θ and the x-axis is the simulated data x. Each circle shows the values of θ_i generated from the prior $p(\theta)$ and the simulated data x_i for it. The red circles show all the θ_is that are accepted with $\epsilon = 0.1$ i.e. their simulated data is less than 0.1 unit away from the observed data.

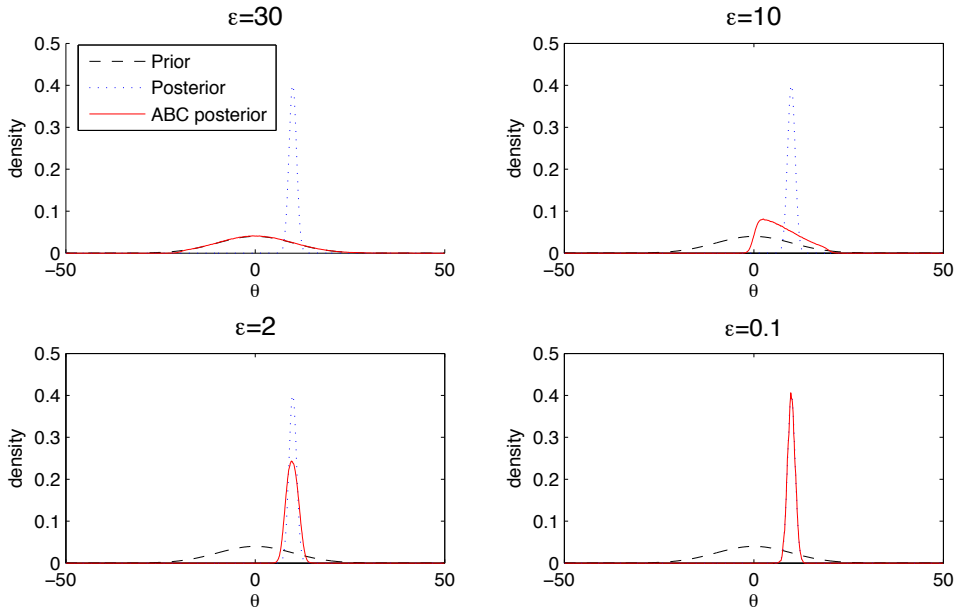


Figure 2.2: Effect of choice of bandwidth ϵ on the estimation of the posterior distribution for the toy example of Figure 2.1. In this example there are two sources of error, the Monte Carlo error and the error due to use of a bandwidth ϵ. Monte Carlo error can be reduced by simulating more samples and the error due to use of bandwidth can be reduced by decreasing ϵ. For large values of ϵ, the ABC estimate of the posterior distribution is approximately the prior distribution. As $\epsilon \to 0$ the ABC estimate of the posterior distribution gets closer to the true posterior distribution as that is the main source of error. Although as $\epsilon \to 0$, the acceptance ratio reduces. Therefore choice of bandwidth ϵ is a trade off between accuracy and computational cost.

ABC algorithm which targets the augmented posterior $p_{\text{ABC}}(\theta, x|D)$. ABC-MCMC is an auxiliary variable MCMC sampler on the state $\theta \times X$ such that proposal can be factorised as:

$$q((\theta, x) \to (\theta', x')) = q(\theta'|\theta)p(x'|\theta') \tag{2.16}$$

with the augmented target density of

$$p_{\text{ABC}}(\theta, x|D) \propto p(D|x, \theta)p(x|\theta)p(\theta) \tag{2.17}$$

where $p(D|x, \theta) = K_\epsilon(S(x) - S(D))$ is the kernel that measures the similarity between $S(x)$ and $S(D)$ and includes a bandwidth ϵ. Then the Metropolis-Hastings ratio is given by:

$$\begin{aligned}\alpha((\theta, x), (\theta', x')) &= \min(1, \frac{K_\epsilon(S(x') - S(D))p(x'|\theta')p(\theta')}{K_\epsilon(S(x) - S(D))p(x|\theta)p(\theta)} \frac{q(\theta|\theta')p(x|\theta)}{q(\theta'|\theta)p(x'|\theta')}) \\ &= \min(1, \frac{K_\epsilon(S(x') - S(D))p(\theta')}{K_\epsilon(S(x) - S(D))p(\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)})\end{aligned} \tag{2.18}$$

As it can be seen the Metropolis-Hastings ratio does not require the model likelihood to be calculated. The ABC-MCMC sampling is given by Algorithm 7. It can be shown that Algorithm 7 produces a Markov chain with stationary distribution $K_\epsilon(S(x) - S(D))p(x|\theta)p(\theta)$ by showing that it is reversible with respect to it (Marjoram *et al.*, 2003).

One advantage of ABC-MCMC over rejection ABC is an improved acceptance ratio. On the other hand, possible drawbacks of ABC-MCMC are that if the chain is initiated at the tail of the posterior distribution when the proposal distribution and the kernel density K_ϵ are badly designed, the chain may get stuck and it would take a very long time for it to converge to the correct distribution. In addition, for multi-modal

Algorithm 7 ABC-MCMC

Choose a starting point (θ_0, x_0).
for $i = 1, \ldots, n$ **do**
    Propose a random perturbation to the current state for θ_t, to generate a new state θ' using proposal density q.
    Simulate x' given θ' from the model.
    Generate a random number $U \sim \text{Unif}[0, 1]$.
    if $U \leq \frac{K_\epsilon(S(x')-S(D))p(\theta')q(\theta_t|\theta')}{K_\epsilon(S(x_t)-S(D))p(\theta_t)q(\theta'|\theta_t)}$ **then**
        $(\theta_{t+1}, x_{t+1}) = (\theta', x')$
    else
        $(\theta_{t+1}, x_{t+1}) = (\theta_t, x_t)$
    end if
end for

posterior distributions the chain can get stuck in one of the modes and it will not explore the whole of state space fully in limited amount of time.

2.3.3 Sequential ABC

Sisson *et al.* (2007) introduced partial rejection control ABC (ABC-PRC) which uses sequential importance sampling (SIS) methodology. In this algorithm a population of samples from the parameter space are propagated from an initial distribution through intermediary distributions until they represent a sample from the target distribution. There are several versions of the algorithm with slight modifications (Sisson *et al.*, 2007; Beaumont *et al.*, 2009; Toni *et al.*, 2009). Here we use the ABC-SMC algorithm introduced by Toni *et al.* (2009).

We would like to have N samples from the ABC posterior $p_{\text{ABC}}(\theta, x|D) \propto p(D|x, \theta)p(x|\theta)p(\theta)$. To reduce error, we would like ϵ to be as small as possible, but setting the bandwidth ϵ to very small values will lead to poor acceptance ratio. To overcome this problem ABC-SMC uses a natural sequence of intermediate distributions that are constructed by modifying the bandwidth ϵ. To get to a final bandwidth of ϵ_t the

Algorithm 8 ABC-SMC

Initialise $\epsilon_1, \cdots, \epsilon_T$.
for $t = 1, \ldots, T$ **do**
  for $i = 1, \ldots, n$ **do**
    if $t = 1$ **then**
      Sample θ'' from the prior $p(\theta)$.
    else
      Sample θ' from the previous population $\{\theta_{t-1}\}$ according to their weights w_{t-1}.
      Propose a perturbation from θ' to θ'' according to $q(\theta''|\theta')$.
    end if
    Generate x'' given θ'' from the model.
    if $K(S(x'') - S(D)) < \epsilon_t$ **then**
      $\theta_t^{(i)} = \theta''$.
      if $t = 1$ **then**
        $w_t^{(i)} = 1$.
      else
        $w_t^{(i)} = \frac{p(\theta_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} q(\theta_t^{(i)}|\theta_{t-1}^{(j)})}$
      end if
    else
      go to 4
    end if
    $i = i + 1$.
  end for
  Normalise weights.
end for

algorithm moves through several ϵ_i such that $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_t$.

In ABC-SMC, a number of sampled parameter values (particles), $\{\theta^{(1)}, \cdots, \theta^{(N)}\}$, from the prior distribution $p(\theta)$ are propagated through a series of intermediate distributions until they represent a sample from the target distribution $\mathbb{1}_{B_{\epsilon_T,K}}(x)p(x|\theta)p(\theta)$. The tolerances ϵ_i are chosen such that $\epsilon_1 > \cdots > \epsilon_T$. The ABC-SMC is given by Algorithm 8.

ABC is usually used to infer model parameters, however all of the ABC algorithms can be used to compare models by calculating Bayes factor (Kass and Raftery, 1995) as discussed by Grelaud *et al.* (2009); Didelot *et al.*

(2011b); Robert *et al.* (2011).

Chapter 3

Biased Recombination

3.1 Introduction

Bacteria are organisms that reproduce clonally, but they occasionally exchange fragments of DNA with one another. This process can lead to two outcomes, non-homologous and homologous recombination (Vos, 2009). Non-homologous recombination occurs when a novel segment of DNA from the donor cell is inserted into the genome of the recipient cell. On the other hand, homologous recombination happens when the DNA from the donor cell replaces its homologous counterpart in the genome of the recipient cell. In this study we are only concerned with the "core" genome of regions present in all sampled genomes (Medini *et al.*, 2008), and therefore only homologous recombination is relevant. Foreign DNA can be taken up by the recipient cell through one of the three mechanisms: conjugation (transfer of DNA from one cell to another when they are in physical contact), transduction (bacteriophage mediated DNA transfer) or transformation (uptake of DNA from the environment by the recipient cell) (Thomas and Nielsen, 2005). In homologous recombination, the

recipient cell then replaces the homologous section of its DNA with the foreign DNA segment.

A first concept that has helped appreciate the role of recombination in bacteria is linkage disequilibrium (LD), or the non-random association of alleles at different loci (Maynard Smith *et al.*, 1993). LD between a pair of sites is expected to decrease as more and more recombination events affect exclusively one or other site, so that LD is a function of the distance between pairs of sites. In bacteria on average LD decreases down to a plateau level when pairs of sites are considered that are further and further away from each other on the genome, and this is often represented graphically (e.g. Namouchi *et al.* 2012; Takuno *et al.* 2012). Another important concept is tree homoplasy which is said to occur when given a known tree, a site could not have arisen without either recombination or repeat mutation (Maynard Smith and Smith, 1998; Maynard Smith, 1999). The probability of a site being homoplasic increases with the number of recombination events affecting the site. For this reason, tree homoplasy is commonly used as an indicator of the prevalence of recombination (e.g. Nübel *et al.* 2008; Harris *et al.* 2010). A related notion is incompatibility between pairs of sites (also known as the four-gamete test or G4), which occurs when two sites cannot be explained by a shared phylogenetic tree without either recombination or repeat mutation (Hudson and Kaplan, 1985; Maynard Smith, 1999). Incompatibility between pairs of sites is often used to identify recombination events (e.g. Takuno *et al.* 2012; Yahara *et al.* 2012).

Recombination plays a key role in shaping the patterns of all these summary statistics, but they are also crucially affected by other factors, which makes them difficult to interpret. This includes the population structure underlying the relationships between the individuals under study

(McVean *et al.*, 2002; Wakeley and Lessard, 2003), and this effect is likely to be especially important in bacteria because of their clonal mode of reproduction. Another factor likely to be important in bacterial population genetics is biased recombination, which we define in contrast to free recombination where all individuals in the population are equally likely to recombine. There are many factors contributing to recombination being biased rather than free. Laboratory experiments have shown that the recombination process is homology dependent so that it tends to happen more often between individuals that are less diverged (Roberts and Cohan, 1993; Zawadzki *et al.*, 1995; Majewski *et al.*, 2000; Majewski, 2001). Furthermore, the geographical and ecological structures observed in many bacterial populations implies a greater opportunity of recombination for pairs of cells that are closely related (Feil and Spratt, 2001; Majewski, 2001; Cohan, 2002; Didelot and Maiden, 2010). Purifying selection may also effectively prevent recombination between distantly related bacteria. All these effects would clearly be hard to disentangle, and here we group them all under the single concept of biased recombination. The strength of this bias is an important factor to take into account in order to understand recombination in bacteria.

Biased recombination determines how often recombination happens within clades in a population than between clades. If recombination is biased then it happens more often between isolates that are closely related than between isolates that are distantly related. The signature of biased recombination in the sequence data therefore will be distinct amount of recombination within clades than between clades in a population. To measure the amount of recombination that moves between clades, we define clade homoplasy for biallelic sites as follows: divide the clonal

genealogy into two clades. If both alleles are present in both clades then the site is clade homoplasic. This measure should give us an indication of how much recombination is moving between the two clades.

Furthermore biased recombination determines how often recombination happens within the diversity of the population under study rather than from other sources. Such recombination events from external sources would strongly affect LD, but have little or no effect on homoplasy and G4 since they introduce what is in effect new polymorphism from the viewpoint of the studied population.

Here we introduce a new statistical framework for inferring the recombination parameters, including the rate of bias in recombination, from a sample of bacterial genomes. Our starting point is an evolutionary model of free recombination which describes the ancestral recombination process given the clonal relationships in the sample. We show how this can easily be extended to allow recombination to be biased. We describe how data can be efficiently simulated under the model, which is crucial to allow the use of approximate Bayesian computation techniques (ABC; Pritchard *et al.* 1999; Beaumont *et al.* 2002) to estimate parameters. We use informative summary statistics about the recombination process such as LD, homoplasy and G4 to infer parameters. Applications are presented on simulated datasets as well as on a real dataset of *Bacillus* genomes.

3.2 Model and methods

3.2.1 Free recombination model

The process of homologous recombination in bacteria is asymmetric in terms of the genetic contributions made by donor and recipient cells, since typically a small segment of DNA from the donor in the order of a few hundreds or thousands of nucleotides in length is incorporated into the genome of the recipient which is much longer (Didelot and Maiden, 2010). This asymmetry contrasts with the well-studied mechanism of crossing-over in eukaryotic sexual reproduction where the two parents contribute equally. Consequently, it is possible to consider the (potentially empty) set of genomic sites that have not been affected by recombination since a sample of isolates evolved from a common ancestor, and the ancestral relationships between the isolates at these sites is called the clonal genealogy (Guttman, 1997). Alternatively, the clonal genealogy of a set of isolates can be defined as the ancestral tree obtained by tracing the ancestry of the isolates back in time and following the ancestral line of the recipient cell (rather than the donor cell) whenever a recombination event occurred.

The coalescent model with gene-conversion describes the ancestry of a bacterial sample subject to homologous recombination (Wiuf, 2000; Wiuf and Hein, 2000; McVean *et al.*, 2002; Didelot *et al.*, 2009b). A useful approximation of this process is the ClonalOrigin model (Didelot *et al.*, 2010), where given the clonal genealogy the recombinant lines of ancestry are assumed to be independent of each other. This means that given the clonal genealogy the recombinant lines of ancestry are not allowed to recombine and are only allowed to coalesce with the clonal genealogy.

Symbol	Description
	Symbols used for the data
A	Aligned sequence data
L	Total length of the alignment
B	Number of blocks in the alignment
W_i	i^{th} summary statistic of data
	Symbols used for the clonal genealogy
$\mathcal{T}$	Clonal genealogy
T	Sum of branch lengths of the clonal genealogy
	Symbols used for the recombination events
R_1	Number of recombination events affecting the first site
$R_2^\star$	Number of recombination events affected the first site that survive to the second site
R_2^I	Number of recombination events that start between the two sites and affect the second site
a_i	Where the ancestry of the donor meets the clonal genealogy for event i (departure point)
b_i	Where the transfer of DNA fragment from donor to recipient occurs for event i (arrival point)
$L(a_i, b_i)$	Sum of branch lengths on the clonal genealogy between the departure and arrival of event i
$D(a_i, b_i)$	Distance in coalescent unit of time between donor and recipient cells of event i
	Symbols used for the global 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
λ	Rate of bias in the recombination process
δ	Mean of the geometric distribution modelling the length of recombinant segments

Table 3.1: Table of symbols

Consequently, the clonal and recombination processes can be separated. Here however we exploit another property of this model, namely the fact that it has a simple Markovian structure along the genome, similar to that of the Sequentially Markov Coalescent in approximating the crossing-over ancestral recombination graph (McVean and Cardin, 2005; Marjoram and Wall, 2006). Given the clonal genealogy this allows for the simulation of pairs of sites at a given physical distance from each other on the genome. As both LD and G4 are defined for pairs of sites, we use this Markovian property of the model to simulate these summary statistics in a computationally efficient manner. A formal description of this model follows, and the mathematical symbols used are summarized in Table 3.1.

In the ClonalOrigin model (Didelot *et al.*, 2010), recombination events are independent of one another given the clonal genealogy and the total number of recombination events R given the total branch length of the clonal

genealogy T and the population recombination rate ρ is Poisson distributed:

$$\mathbf{P}(R = r|\rho, T) = \frac{(\frac{\rho T}{2})^r e^{-\frac{\rho T}{2}}}{r!} \tag{3.1}$$

Each recombination event i has four properties: the departure point on the clonal genealogy a_i where the ancestry of the donor cell meets the clonal genealogy, the arrival point on the clonal genealogy b_i where the recombination occurs, the site on the chromosome where recombination starts x_i and the site on the chromosome where recombination ends y_i. Figure 3.1 shows three recombinations with their arrival and departure points on the clonal genealogy. The three event have the same arrival points, but different departure points on the clonal genealogy.

The arrival points b_i are uniformly distributed on the clonal genealogy as recombination happens at a constant rate on the branches of the clonal genealogy. A recombinant edge reconnects with the clonal genealogy at a rate equal to the number of ancestors in the clonal genealogy as in the standard coalescent model (KINGMAN, 1982b). Thus a_i conditional on b_i is distributed as:

$$\mathbf{P}(a_i|b_i, \mathcal{T}) = e^{-L(a_i,b_i)} \tag{3.2}$$

where $\mathcal{T}$ is the clonal genealogy and $L(a_i, b_i)$ is the sum of branch lengths on the clonal genealogy between the time of a_i and b_i. In addition we assume that the recombination events are uniformly distributed along the observed sequences and that their length is geometrically distributed with mean δ.

3.2.2 Biased recombination model

We extend the ClonalOrigin model to incorporate the bias in recombination and modify Equation 3.2 such that a recombinant edge

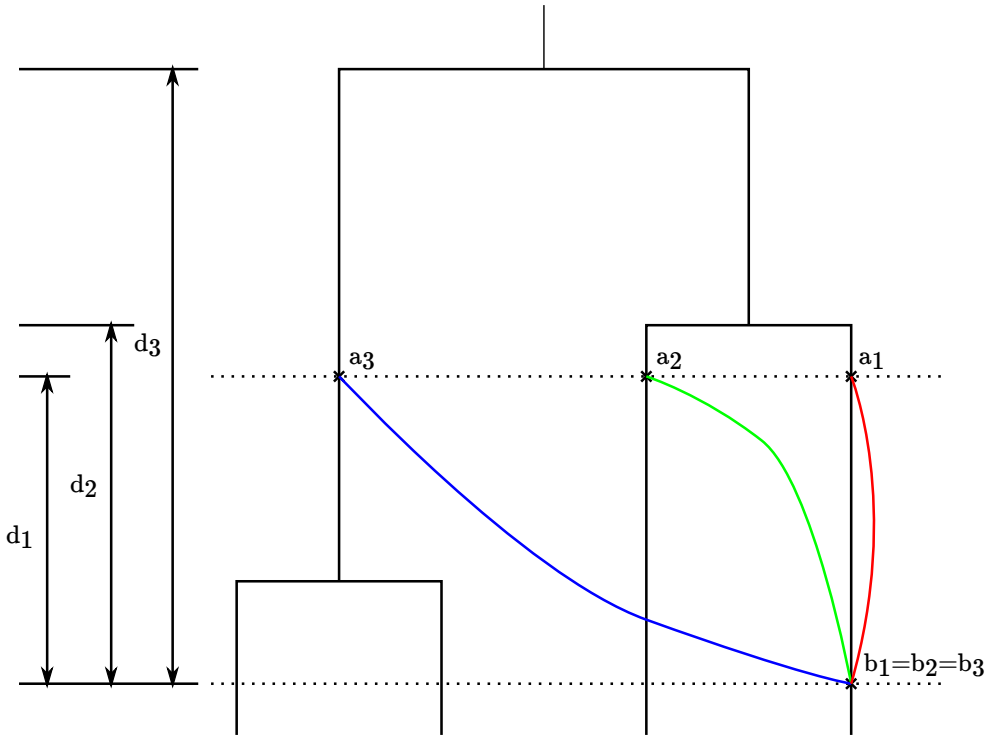


Figure 3.1: Illustration of the recombination model. Consider three recombination events arriving at points $b_1 = b_2 = b_3$ and departing from points a_1, a_2 or a_3 on the clonal genealogy. In the ClonalOrigin model (free recombination, Equation 3.2) these three departure points are equally likely because the sums of branch lengths between the times of each b_i and a_i are equal: $L(a_1, b_1) = L(a_2, b_2) = L(a_3, b_3)$. The amount of evolutionary distance between the donor and recipient cells for the three recombination events is given by $D(a_1, b_1) = 2d_1, D(a_2, b_2) = 2d_2$ and $D(a_3, b_3) = 2d_3$. In the biased recombination model (Equation 3.3), the probability of departing from a_1 is higher than from a_2 which is higher than from a_3, because the amount of evolutionary distance between the donor and recipient cells is increasing: $D(a_1, b_1) < D(a_2, b_2) < D(a_3, b_3)$.

coalesces with the clonal genealogy at a rate that depends on both the sum of number of ancestors in the clonal genealogy and the amount of evolutionary distance between donor and recipient cells. Therefore we propose the following distribution for a_i:

$$\mathbf{P}(a_i|b_i, \mathcal{T}) \propto e^{-L(a_i,b_i)} \times e^{-\lambda D(a_i,b_i)} \tag{3.3}$$

where $D(a_i, b_i)$ is the evolutionary distance in coalescent unit of time between the donor and recipient cells for recombination i and λ is the strength of the recombination bias. Experimental data have shown that the probability of success of a recombination event is inversely proportional to the exponential of the level of sequence divergence in a variety of bacterial species such as *Bacillus* (Majewski and Cohan, 1999), *Escherichia coli* (Vulić *et al.*, 1997) and *Streptococcus pneumoniae* (Majewski *et al.*, 2000). Instead of modelling the rate of recombination as inversely proportional to the exponential of the amount of sequence divergence, we model it as inversely proportional to the exponential of amount of evolutionary distance between the donor and recipient cells. Firstly, the amount of evolutionary distance is a good approximation for the amount of sequence divergence between the donor and recipient cells. Secondly, computationally this is much easier to handle. As our inference procedure is based on simulation, if sequence divergence is used in our model we then have to simulate sequences and use those to decide if a recombination is accepted or not. However with our proposed model we can simply simulate the recombination events by looking at the clonal genealogy without simulating sequences. Our model is an approximation, but computationally it is much more attractive and makes the inference

procedure possible as shown in the following section.

Free recombination is nested in this model, as setting $\lambda = 0$ results in Equation 3.2. For values of λ greater than zero, we have that the probability of recombination decreases with the evolutionary distance between donor and recipient. Figure 3.1 shows the relationship between $D(a_i, b_i)$ and $L(a_i, b_i)$ for three recombination events with the same arrival points, but three different departure points on the clonal genealogy. Under a free recombination model, the three recombination events would have the same probability because the sum of branch lengths of clonal genealogy between the arrival and departure points on the clonal genealogy are the same. However the amount of evolutionary distance between the donor and recipient cells increases from recombination events 1 to 2 to 3. Thus in the model of biased recombination described by Equation 3.3 with $\lambda > 0$, the probability of event 1 is more than that of event 2 which is more than that of event 3.

3.2.3 Simulating pairs of sites

The sequentially Markovian property of our model allows us to simulate pairs of sites at a given physical distance from each other given the clonal genealogy. The simulation is done in three steps. First we simulate recombination events affecting the first site and their properties. In the second step, we simulate recombination events affecting the second site. This include some of the recombination events from the first site that are long enough to affect the second site and some new recombinations initiated between the two sites. In the third step the local trees for the two sites are computed and mutations are added.

The sequence data is made of B independent blocks with total length

L, and subject to mutation and recombination at population rates θ and ρ respectively. A recombination event may start before a block and be long enough to affect the beginning of a block, so that the probability of observing the recombination start at the beginning of a block is δ times greater than within a block (DIDELOT and FALUSH, 2007). There are B sites at the beginning of blocks and $L - B$ sites within blocks, thus the recombination rate per site is defined as $\rho_s = \rho/(\delta B + L - B)$, and since mutation affects any site with equal probability, the mutation rate per site is $\theta_s = \theta/L$ (DIDELOT *et al.*, 2010).

Given the clonal genealogy $\mathcal{T}$, recombination rate per site ρ_s, mean length of recombination tract δ, the rate of bias in recombination λ, the physical distance between the two sites on the chromosome k and the mutation rate per site θ_s, a pair of sites is simulated as follows:

Step 1 Simulate recombination events for the first site.

(a) We assume that recombinations start between nucleotides and that they are at least one nucleotide long. As the length of recombination events are geometrically distributed with mean δ, the rate at which a site k nucleotides before the first site initiates a recombination that survives to the first site is:

$$\frac{\rho_s}{2} \times (1 - \delta^{-1})^k$$

Summing over all sites before the first site, we get the expected rate of recombination affecting the first site:

$$\sum_{i=0}^{\infty} \frac{\rho_s}{2}(1 - \delta^{-1})^i = \frac{\rho_s}{2} \sum_{i=0}^{\infty}(1 - \delta^{-1})^i = \frac{\rho_s}{2}\delta$$

Therefore the number R_1 of recombination events affecting the first site is Poisson distributed:

$$R_1|T, \rho_s, \delta \sim \text{Poisson}\left(\frac{\rho_s \delta T}{2}\right) \tag{3.4}$$

(b) For each recombination event i, the arrival point on the clonal genealogy b_i is uniformly distributed and the departure point a_i is drawn from Equation 3.3. To simulate from Equation 3.3, we use rejection sampling where the proposal distribution is Equation 3.2 and the simulated a_i is accepted with probability $e^{-\lambda D(a_i, b_i)}$.

Step 2 Simulate recombination events for the second site. Two types of recombination events can affect the second site. Some events affecting the first site may have survived to the second site and new recombinations could have started between the two sites and have survived to the second site.

(a) As the length of recombination events is geometrically distributed, the probability of a recombination that is affecting the first site to have survived to the second site is:

$$\mathbf{P}(\text{Survival}) = (1 - \delta^{-1})^k$$

Thus the number of recombination events $R_2^\star$ from the first that survive to the second site is Binomially distributed:

$$R_2^\star \sim \text{Binomial}(R_1, (1 - \delta^{-1})^k) \tag{3.5}$$

(b) The number of recombination events R_2' that start between the

two sites and that affect the second site is distributed as:

$$R_2'|T, \rho_s, \delta' \sim \text{Poisson}\left(\frac{\rho_s \delta' T}{2}\right) \text{ where } \delta' = \sum_{i=0}^{k-1}(1-\delta^{-1})^i \quad (3.6)$$

This is because there are only k positions between the two sites where recombination could have started.

(c) For each of the R_2' recombination events affecting the second but not the first site, the departure and arrival points on the clonal genealogy are simulated as detailed in Step 1.

Step 3 For both sites, extract the local trees backwards in time (from tips to root), given the clonal genealogy and the recombination events. Mutations are then simulated on these local trees as follows.

(a) The number M_j of mutations affecting the local tree at site j is distributed as:

$$M_j|T_j, \theta_s \sim \text{Poisson}\left(\frac{\theta_s T_j}{2}\right) \quad (3.7)$$

where T_j is the total branch length of the local tree at site j.

(b) We are only interested in simulating polymorphic sites and Equation 3.7 for plausible values of θ_s and T_j leads to many non-polymorphic sites. To remedy this problem, we use an importance sampling strategy (Fearnhead, 2007). A local tree is in the target distribution if Equation 3.7 leads to at least one mutation on that local tree. The proposal distribution is made of all local trees simulated by Steps 1 and 2. Therefore the

importance sampling weight is:

$$w_j = \frac{\mathbf{P}(\text{Local tree } j \text{ is in the target distribution})}{\mathbf{P}(\text{Local tree } j \text{ is in the proposal distribution})} = \frac{\mathbf{P}(M_j > 0)}{1} = 1 - \mathbf{P}(M_j = 0) = 1 - e^{-\frac{\theta_s T_j}{2}} \quad (3.8)$$

The simulated local trees are importance sampled using the weights from Equation 3.8, and the number of mutations on the local tree is simulated from the truncated Poisson with one or more mutations.

(c) Mutations are uniformly distributed on the local trees. For simplicity we use the Jukes-Cantor model where all mutations are equally likely, but any mutation model could be used (Whelan *et al.*, 2001).

3.2.4 Inference using whole genomes

Whole genomes can be compared using Mauve that detects and aligns the conserved genomic regions in the presence of rearrangements (Darling *et al.*, 2004, 2010). Given a core alignment A of whole bacterial genomes and the clonal genealogy $\mathcal{T}$ (estimated for example using ClonalFrame; Didelot and Falush 2007), we want to infer the posterior density of the model parameters $\mathbf{P}(\rho_s, \delta, \lambda, \theta_s | A, \mathcal{T})$. However due to the complexity of the model, the likelihood function is intractable and therefore we cannot use standard approaches such as a Markov chain Monte Carlo (MCMC). One solution would be to use data augmentation techniques as in Didelot *et al.* (2010). Instead here we use Approximate Bayesian Computation (ABC; Pritchard *et al.* 1999; Beaumont *et al.* 2002) where the likelihood does not have to be computed, but simulation from the model

has to be efficient. In effect, the likelihood is approximated through a distance metric on a set of informative summary statistics between the simulated and observed data. There are several implementations of the ABC algorithm (reviewed in Beaumont 2010; Csilléry *et al.* 2010) and we have implemented and tested both ABC-MCMC (Marjoram *et al.*, 2003) and ABC-SMC (Beaumont *et al.*, 2009) approaches. The results presented used a parallel ABC-MCMC implementation where given the current chain state θ_j, n states $\theta'_1, \ldots, \theta'_n$ are proposed independently and for each one data $x'_1, \ldots, x'_n$ are simulated in parallel (where n is the number of cores available). The proposed states and their simulated data are examined sequentially in the ABC-MCMC algorithm. For each rejected proposed state, the MCMC stays at θ_j. If a proposed state θ'_i is accepted then remaining proposed states $\theta'_{i+1}, \ldots, \theta'_n$ are discarded. If all proposed states are rejected, then the MCMC has stayed at θ_j for n states and the process is repeated with proposal of n new states. This parallelisation scheme is similar to that of pre-fetching which was developed for MCMC with known likelihood (Brockwell, 2006).

3.2.5 Summary statistics and distance metric

Since one of the parameters we need to infer is the mutation rate θ_s, we included in the summary statistics the proportion of segregating sites S which is highly informative about this parameter (Watterson, 1975). To calculate S from the simulated data, Equation 3.8 was used which gives the probability that a simulated site is polymorphic. The most widely used summary statistics that are informative about the recombination process are LD, homoplasy and incompatibility between pairs of sites (meaning for bi-allelic sites, all four possible haplotypes are present, G4). To measure

LD, r^2 was used which quantifies the amount of association between a pair of bi-allelic sites (HILL and ROBERTSON, 1968). As r^2 and G4 are distance dependent, for empirical datasets, we plot the mean of r^2 and G4 against distance between the sites. Figure 3.2 shows an example for a sample of 13 *Bacillus* whole genomes (DIDELOT *et al.*, 2010). As summary statistics we choose three points on the LD and G4 plots that capture the decay and the constant part of the plots. These points are shown with blue circles in Figure 3.2 and here correspond to pairs of sites at distances of 50, 200 and 2000 nucleotides from each other. These distances need to be chosen according to the r^2 and G4 plots of the given empirical dataset. Background LD can be affected by other factors than recombination such as genetic drift (FALUSH *et al.*, 2003), although these would not affect the variation in LD at different distances. To account for this, and since we are here interested in recombination, we used the differences in LD as summary statistics i.e. $\mathrm{LD}_{100} - \mathrm{LD}_{2000}$ and $\mathrm{LD}_{100} - \mathrm{LD}_{200}$. We also included as summary statistic the proportion of homoplasic sites relative to the clonal genealogy and a new variable which we called clade homoplasy and which is calculated as follows: Given a clonal genealogy, it is divided into its two largest clades and for biallelic sites if both alleles are present in both clades, we say that the site is clade homoplasic. We are introducing this new summary statistic as an indicator of the amount of recombination between the clades which will be informative about the rate of bias in the recombination process.

In total, we therefore use eight summary statistics: the proportion of segregating sites, two distance-based differences in LD, three distance-based values of G4, one value for homoplasy and one for clade homoplasy. These summary statistics are compared between the observed and simulated datasets using a metric equal to the sum of the squared

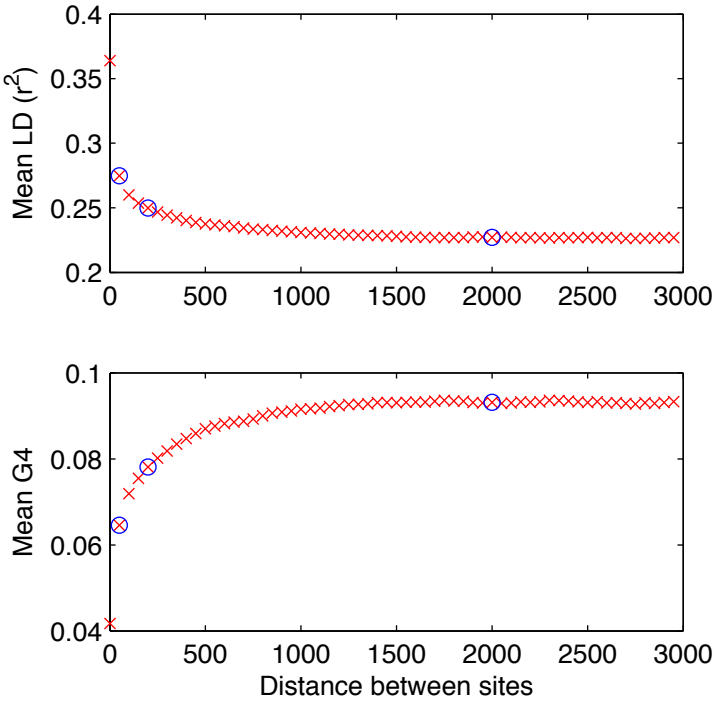


Figure 3.2: LD and G4 plots for 13 *Bacillus* whole genomes, as a function of the distance between pairs of sites. LD decreases and G4 increases until they both plateau at around 1000 bp. The slopes in the plots indicate that recombination is present in this dataset. The blue circles show the three values of LD and G4 that were used as summary statistics in the inference procedure.

normalized distances:

$$\text{dist}(x', x) = \sum_i \left(\frac{W_i(x') - W_i(x)}{W_i(x)} \right)^2 \tag{3.9}$$

where x' is the simulated data, x is the observed data and W_i is the i-th summary statistic of the data.

3.2.6 Monte Carlo estimation of r/m

An important quantity in bacterial population genetics is the ratio r/m of rates at which nucleotides are substituted due to recombination and mutation (Guttman and Dykhuizen, 1994; Vos and Didelot, 2009). In our model this is equal to:

$$\begin{aligned} r/m &= \frac{(\text{Recombination rate per site}) \times \mathbf{P}(\text{substitution}|\text{recombination})}{(\text{Mutation rate per site})} \\ &= \frac{\rho_s \delta \times \mathbf{P}(\text{substitution}|\text{recombination})}{\theta_s} \end{aligned} \tag{3.10}$$

Given a recombination on the clonal genealogy, the probability of a substitution being introduced due to the recombination event at the site is given by

$$\mathbf{P}(\text{substitution}|\text{recombination}) \approx \frac{\theta_s}{2} \mathbb{E}(D) \tag{3.11}$$

Where $\mathbb{E}(D)$ is the expected distance between the donor and recipient cells in coalescent unit of time given a recombination event. Therefore for a given set of parameters, the probability of substitution given a recombination event is estimated using Equation 3.11 by simulating many recombination events on the clonal genealogy and computing the average distance between donors and recipients. Equation 3.10 is then used to estimate r/m. This Monte-

Carlo procedure is applied for each value of the parameters in the posterior sample in order to obtain a sample from the posterior distribution of r/m.

3.2.7 Simulation of future events on the clonal genealogy

Population genetics models have shown that recombination with high rates of homology dependency can be on one hand a strong cohesive force between highly homologous bacteria and on the other hand very rare between diverged bacteria, thus resulting in clusters of diversity which could be considered to represent separate species (Falush *et al.*, 2006; Hanage *et al.*, 2006; Fraser *et al.*, 2007; Achtman and Wagner, 2008; Fraser *et al.*, 2009). To test this hypothesis further for a given dataset, we can consider the next evolutionary events likely to happen to any one of the isolates using our model.

We would like to know the rate at which any two isolates on the clonal genealogy diverge or converge relative to each other due to a future event. We first consider the effect of mutation on isolates i and j. In our model the rate at which an isolate is affected by mutation at a given site is $\theta_s/2$ and we are assuming Jukes-Cantor mutation model (Jukes and Cantor, 1969). In addition we assume the proportion of segregating sites between isolates i and j is π_{ij}.

Let us assume the two isolates are polymorphic relative to each other at a given site. Therefore if a mutation affects isolate i at that site, with probability 2/3 the two isolates stay polymorphic (no change) and with probability 1/3 they become non-polymorphic (convergence) relative to each other at that site. If isolates i and j are not polymorphic relative to each other at the given site, then a mutation on isolate i will lead to the two isolates becoming polymorphic (divergence) relative to each other with

probability 1. Thus the rate of divergence and convergence between isolates i and j due to a mutation affecting isolate i is given by:

$$\text{rate of convergence between } i \text{ and } j \mid i \text{ mutating} = \frac{\theta_s}{2}\pi_{ij}\frac{1}{3} \tag{3.12}$$

$$\text{rate of divergence between } i \text{ and } j \mid i \text{ mutating} = \frac{\theta_s}{2}(1-\pi_{ij}) \tag{3.13}$$

The same rates can be calculated for a mutation affecting isolate j. The total rates of divergence and convergence can be calculated by adding the results for the two case. The difference between the total rate of divergence and convergence between the two isolates will tell us if the overall affect of mutation on the two isolates is divergence or convergence.

Next we look at the effect of recombination on divergence and convergence between isolates i and j. The rate at which an isolate is affected by recombination at a given site is $\rho_s/2$. Let us call the probability of isolates i and j being polymorphic relative to each other, before recombination affecting isolate i, P_{before}, and the probability of the two isolates being polymorphic after a recombination affecting isolate i, P_{after}. We can estimate the rate of divergence between the two isolates, due to a recombination affecting isolate i by:

$$\text{rate of div/conv between } i \text{ and } j \mid i \text{ recombining} = (P_{\text{after}} - P_{\text{before}})\frac{\rho_s}{2} \tag{3.14}$$

P_{after} and P_{before} cannot be calculated analytically. For this reason we use Monte Carlo simulation to estimate these probabilities. This is done in two steps. We first simulate a "normal" site and identify if isolates i and j are polymorphic relative to each other. Next we simulate a recombination affecting isolate i and identify if the two isolates are polymorphic relative

to each other. We repeat this procedure many time to estimate $P_{i\text{before}}$ and P_{after}. The details of the Monte Carlo procedure are as follows:

1. Simulate a "normal" site.

   (a) Given the clonal genealogy, population recombination rate ρ_s, mean recombination tract length δ and the rate of bias in recombination λ, simulate recombinations for the site.

   (b) extract the local tree for the site.

   (c) Given the population mutation rate θ_s, simulate mutations on the local tree. Identify if isolates i and j are polymorphic relative to each other.

2. Simulate an additional recombination affecting isolate i.

   (a) Simulate an additional recombination event affecting isolate i at the tip of the clonal genealogy. Given the previous recombination events affecting the site and the additional one, extract the new local tree.

   (b) Given the population mutation rate θ_s, simulate mutations on the new local tree. Identify if isolates i and j are polymorphic relative to each other or not.

3. repeat the above procedure many times.

In Equation 3.14 negative rates indicate convergence and positive rates indicate divergence. The same procedure can be applied to isolate j and thus, we can calculate the rate of divergence and or convergence for all pairs of isolates due to recombination. Rates of divergence and convergence due to mutation and recombination can be compared to estimate the overall rate of divergence or convergence for all pairs of isolates.

3.3 Results

3.3.1 Parameters and summary statistics

We used simulated data to investigate the relationship between the model parameters and the summary statistics. A clonal genealogy with fifteen taxa was simulated under the coalescent model (Figure 3.3) and the following parameters were used $\rho_s = 0.02, \delta = 300, \lambda = 1.2$ and $\theta_s = 0.05$ which represents reasonable values for a real bacterial population (Fraser *et al.*, 2007; Didelot *et al.*, 2010). We then changed one parameter at a time in the intervals $\rho_s \in [0, 0.4]$, $\delta \in [0, 4000]$, $\lambda \in [0, 10]$, $\theta_s \in [0, 0.3]$ and simulated the summary statistics in order to see how they varied with the parameters. For each parameter value, we simulated 2000 pairs of sites distant from each other by 50, 200 and 2000 bp.

Figure 3.4 shows how the summary statistics change with the model parameters. ρ_s, δ and λ have large influence on r^2, G4 and homoplasies and relatively small effect on the proportion of segregating sites S. On the other hand θ_s has little impact on r^2, G4 and homoplasies, but it has a large influence on S. In the absence of recombination ($\rho_s = 0$) the differences in mean r^2 is zero which indicates r^2 is independent of distance between pairs of sites. As ρ_s increases the differences in mean r^2 increase to a maximum, beyond which as ρ_s increases the differences in mean r^2 decreases and for very high values of ρ_s, the differences approach zero which indicate r^2 again becomes independent of distances between pairs of sites. Increasing ρ_s increases homoplasies and G4 up to a maximum beyond which the mean G4 and homoplasy slightly decrease. δ has a similar but non-identical effect on r^2, G4 and homoplasies. However λ has the opposite effect on r^2, G4 and homoplasies. This is because as λ

increases, the effect of the recombination decreases as the donor cells tend to have a smaller evolutionary distance relative to the recipient cells and therefore local trees become more and more similar to the clonal genealogy. For extremely high values of λ this results in no differences in mean r^2, no homoplasies and zero incompatible pairs of sites (G4) which is similar to those observed in the absence of recombination. θ_s has the largest influence on the proportion of segregating sites S, but ρ_s, δ and λ also slightly affect it. This is because as the number of recombination events increases, the probability that a recombination edge reattaches itself higher up the clonal genealogy increases and that would increase the total branch length of local trees relative to the clonal genealogy.

It is important to note that the clonal genealogy has a large impact on the observed patterns of LD and homoplasy. To illustrate this, we performed the same sensitivity analysis as above but using a different clonal genealogy (Figure 3.5). The resulting relationships between model parameters and summary statistics are shown in Figure 3.6. These relations are quantitatively the same as we described above based on Figure 3.4, but the exact values differ significantly. It is therefore essential to account for the clonal genealogy as we do here in order to correctly interpret the values of the summary statistics. Having done this, there are strong relationships between model parameters and the summary statistics (Figure 3.4) which means that inference via ABC on the basis of these statistics should provide good statistical power to infer parameter values.

3.3.2 Application to simulated datasets

We first applied our inference methodology to a dataset simulated under our model. Fifteen genomes of length one million bp were simulated, based

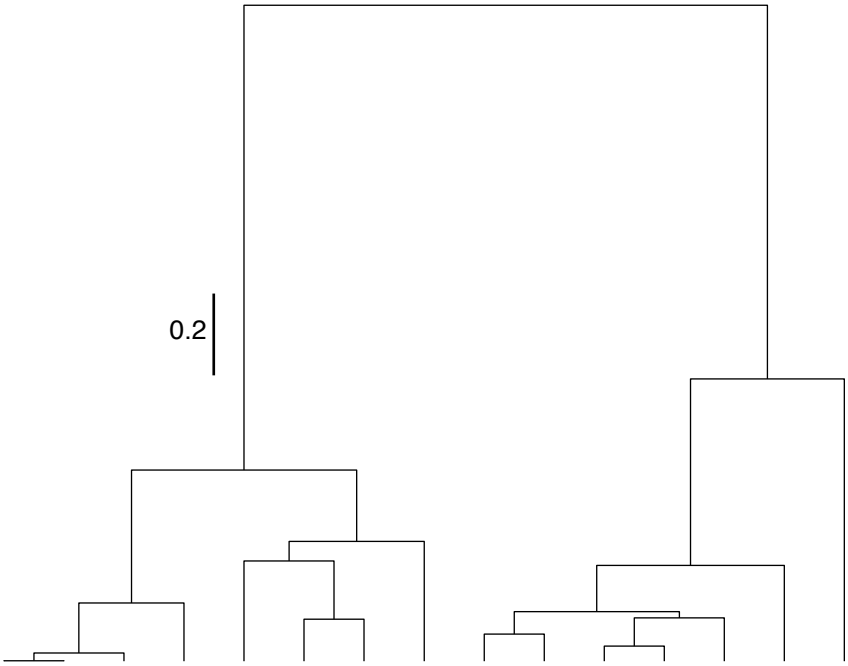


Figure 3.3: Clonal genealogy with fifteen taxa simulated under the coalescent and used to simulate data to investigate the relationship between summary statistics and model parameters. The result are shown in Figure 3.4.

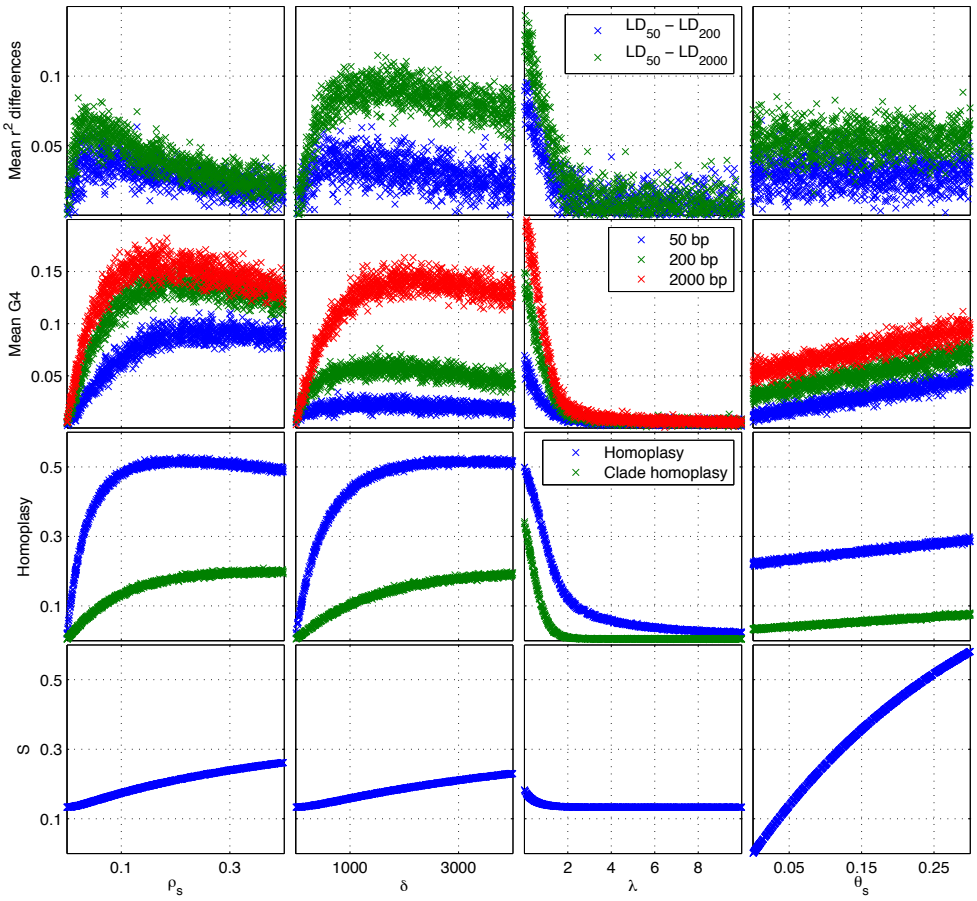


Figure 3.4: Relationship between model parameters and the summary statistics. For a given clonal genealogy (shown in Figure 3.3), the four model parameters were changed one at a time and the summary statistics were simulated. When unchanged, the parameters were equal to $\rho_s = 0.02$, $\delta = 300$, $\lambda = 1.2$ and $\theta_s = 0.05$. θ_s has large influences only on the proportion of segregating sites S while λ, δ and ρ_s have large influences on homoplasy, G4 and LD, but little influence on S.

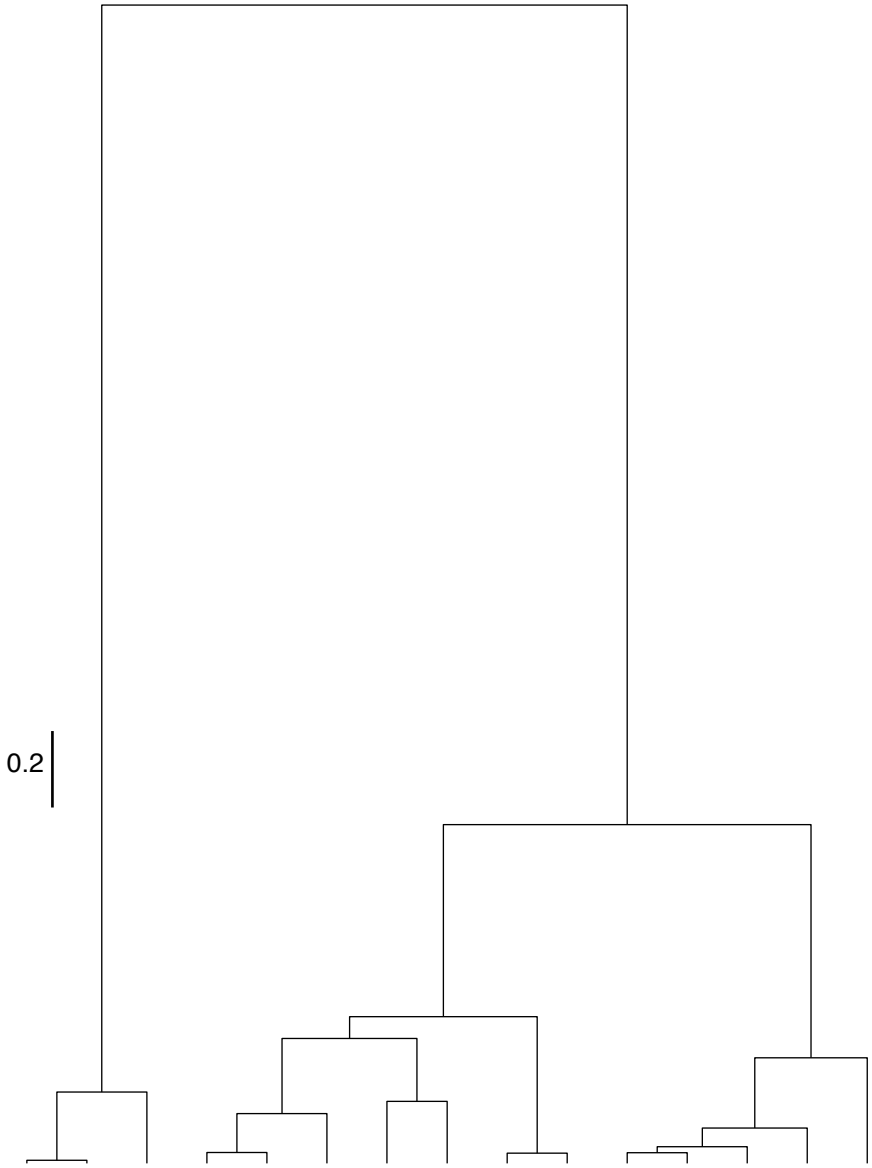


Figure 3.5: Clonal genealogy with 15 taxa simulated under the coalescent and used to simulate data to investigate the relationship between summary statistics and model parameters. The result are shown in Figure 3.6. Note the differences between this tree and tree of Figure 3.3.

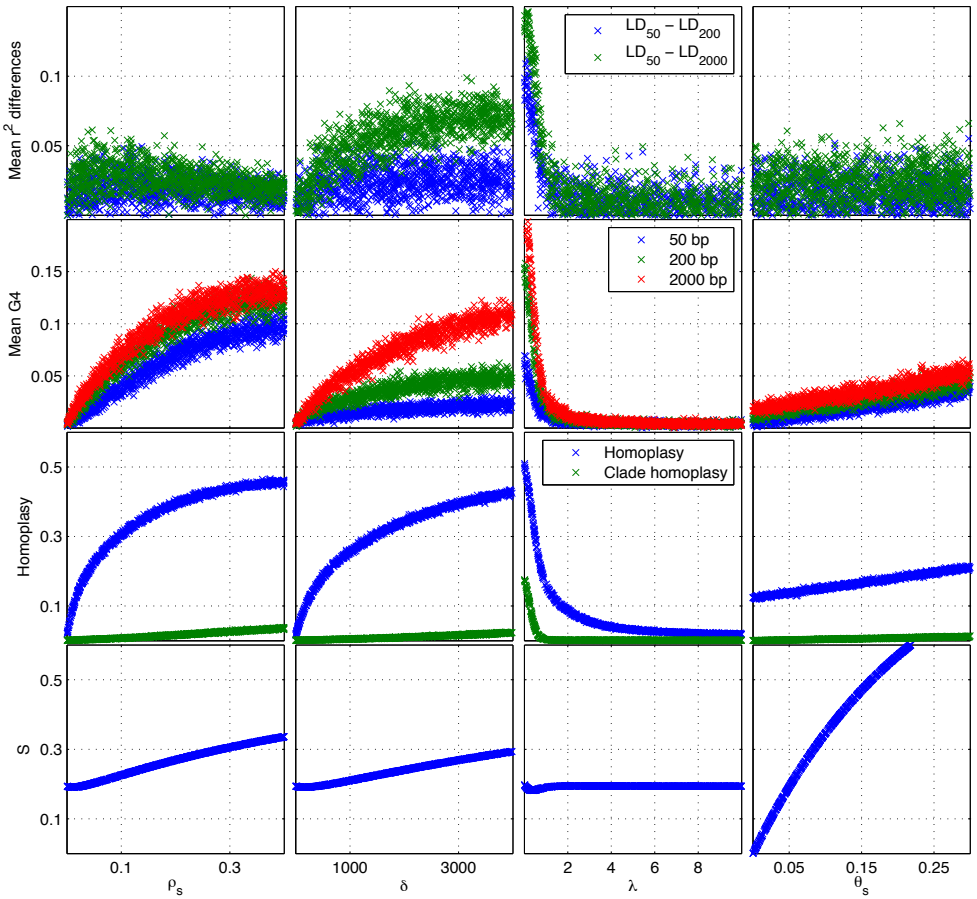
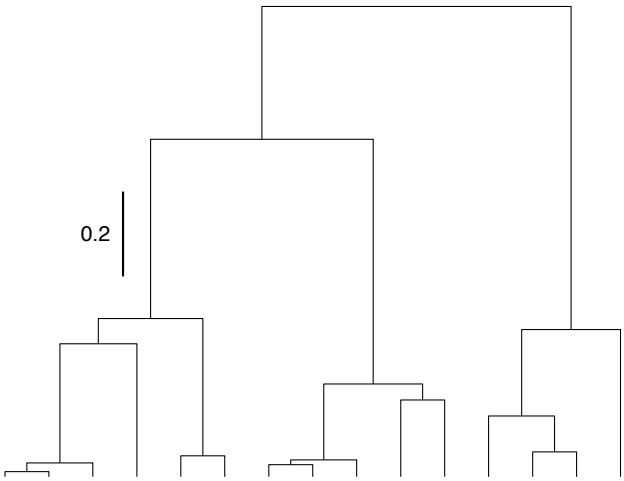


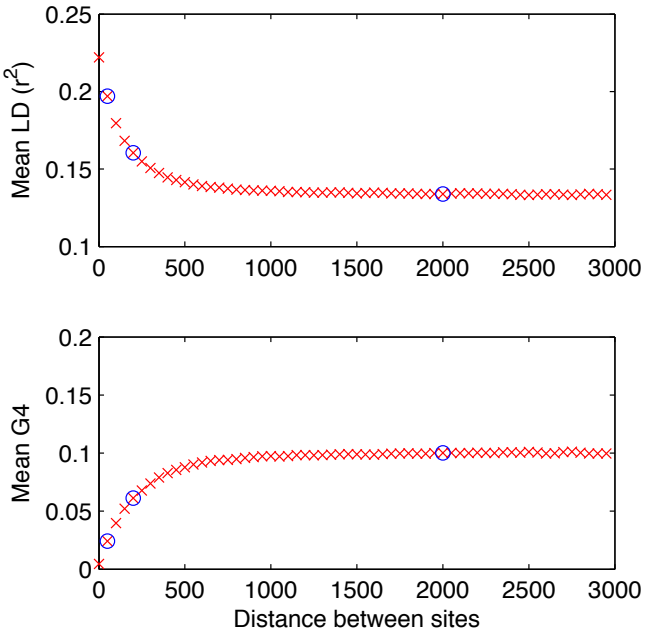
Figure 3.6: Relationship between the summary statistics and model parameters when simulating using the clonal genealogy in Figure 3.5. Note the difference with Figure 3.4. The clonal genealogy (or population structure) has a large impact on the chosen summary statistics. Although the overall relationships between the summary statistics and the model parameters are the same in this figure and Figure 3.4, the exact values differ significantly. Thus it is essential to account for the clonal genealogy to interpret the summary statistics correctly.

on the clonal genealogy shown in Figure 3.7a, and using the following parameters: $\rho_s = 0.02, \delta = 300, \lambda = 1.2$ and $\theta_s = 0.05$. Figure 3.7b shows the LD and G4 plots for this dataset. The LD measure r^2 for this simulated data were equal to (0.1970, 0.1605, 0.1339) for pairs of sites distant by (50, 200, 2000) bp, respectively. The proportions of G4 were equal to (0.0242, 0.0613, 0.1002) for pairs of sites distant by the same respective amounts. The proportion of homoplasic and clade homoplasic sites were respectively (0.3067, 0.0931). Finally the proportion of segregating sites was equal to $S = 0.1235$.

We chose uniform priors for all model parameters on the following ranges: $\rho_s \in [0, 0.2], \delta \in [0, 2000], \lambda \in [0, 10]$ and $\theta_s \in [0, 0.2]$. We ran a parallel ABC-MCMC chain of 300,000 iterations with the ABC threshold $\epsilon = 0.015$ and the proposal density tuned so that acceptance rate was 0.4%. The histograms in Figure 3.8 show the marginal distribution of posterior samples for each of the four parameters. The posterior distribution of the recombination rate per site ρ_s had a mean of 0.020 with a 95% credibility interval CI=0.012-0.028. The posterior of the mean recombination tract length δ had a mean of 309 with CI=226-449. The posterior of the rate of bias of recombination λ had a mean of 1.18 with CI=0.81-1.47. The posterior of the mutation rate θ_s had a mean of 0.050 with CI=0.046-0.054. For each of the four parameters, the true value that was used for simulation was well within the 95% credibility interval and in each case close to the mean of the posterior distribution. Furthermore, Figure 3.8 shows that the posterior distributions are much tighter than the prior distributions for each of the four parameters. This means that the summary statistics upon which inference is based carry significant information about the underlying values of the parameters, as had



(a) Simulated Clonal genealogy.



(b) LD and G4 plots of the simulated data.

Figure 3.7: Clonal genealogy, LD and G4 plots for simulated dataset. Fifteen genomes of length 1 million bp were simulated under the clonal genealogy shown in (a) using the following parameters: $\rho_s = 0.02$, $\delta = 300$, $\lambda = 1.2$ and $\theta_s = 0.05$. (b) LD and G4 plots of the simulated dataset. The blue circles indicate the points used as summary statistics in the inference procedure.

previously been suggested by the correlations between parameters and summary statistics in simulated datasets (Figure 3.4).

To assess the effect of inferring the clonal genealogy incorrectly, we performed two additional simulations. Given the clonal genealogy of Figure 3.7a, the distance matrix $l_{i,j}$ was computed between all pairs of leaves. A modified distance matrix was then computed by replacing each $l_{i,j}$ with a uniform draw from the interval $[0.75l_{i,j}, 1.25l_{i,j}]$, and a modified tree was computed using UPGMA on the modified distance matrix. The two resulting genealogies are shown in Figures 3.9a and 3.9b, and these differ from the true clonal genealogy of Figure 3.7a in both tree topology and branch lengths. These two incorrect genealogies were then used to infer the model parameters. The posterior marginal densities are shown in Figure 3.10, indicating that that our model and inference procedure are robust to slight misspecification of the clonal genealogy. In both cases the true parameters used for simulation of data are well within the 95% credible interval of the posterior densities.

Running this inference procedure on a cluster of 12 Intel 3.33 GHz cores took about 70 hours. The computing time of the inference procedure depends on the range of parameters being inferred as higher values of ρ_s and δ lead to slower simulations. As the inference procedure is time consuming, testing our model on hundreds of simulated datasets is not possible and we tested our algorithm on 11 additional simulated datasets with a range of parameters. We limited our parameter ranges to biologically meaningful values. The parameter ranges used are as follows: $\rho_s = [0, 0.07], \delta = [0, 1000], \lambda = [0, 2]$ and $\theta_s = [0.02, 0.08]$. We used the clonal genealogy of Figure 3.7a and used different parameter values to simulate 11 datasets each made of 15 whole genomes of 1 million bp. We

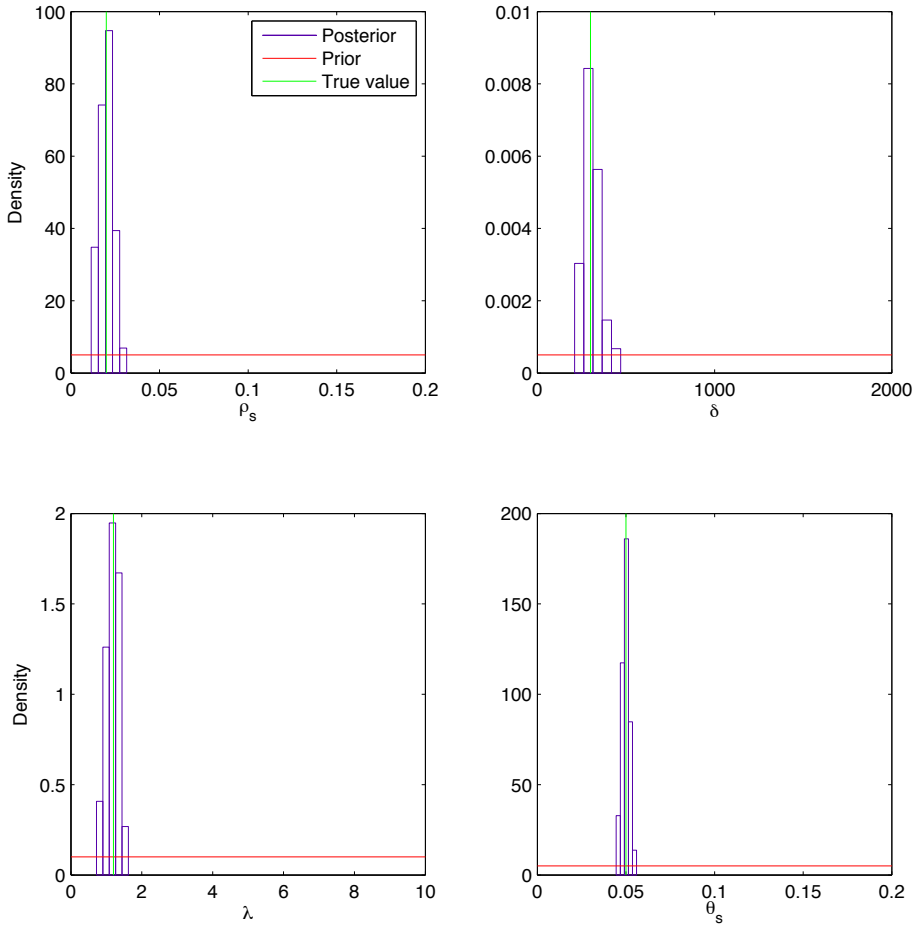
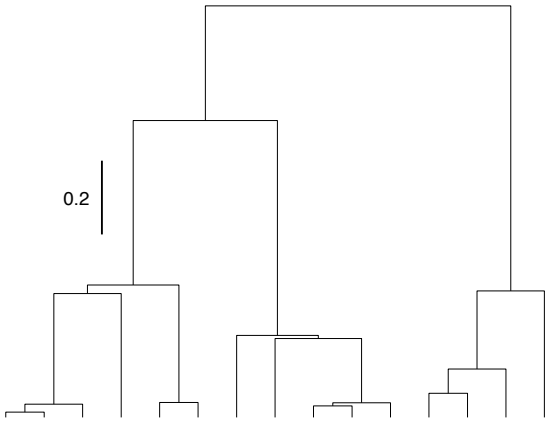
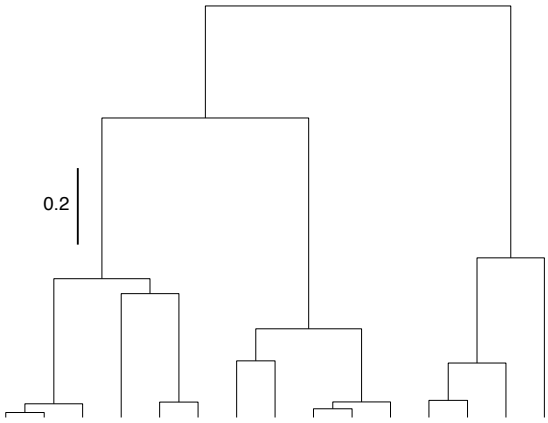


Figure 3.8: Estimated marginal posterior densities of the parameters for the simulated dataset. The values used in simulation are shown in green and equal to $\rho_s = 0.02, \delta = 300, \lambda = 1.2$ and $\theta_s = 0.05$. The red lines show the uniform prior densities used for the model parameters and the blue histograms show the marginal posterior densities estimated using ABC-MCMC.



(a) Clonal genealogy I.



(b) Clonal genealogy II.

Figure 3.9: Incorrect clonal genealogies used to test robustness to inaccuracies in the clonal genealogy. The distance matrix between all pairs of leaves $[l_{ij}]$ of clonal genealogy of Figure 3.7a was modified by replacing each l_{ij} with a uniform draw from the interval $[0.75l_{ij}, 1.25l_{ij}]$. The modified distance matrix was used to construct a UPGMA tree. The two resulting genealogies were used in inference procedure to test the effect of the inaccuracies in the clonal genealogy.

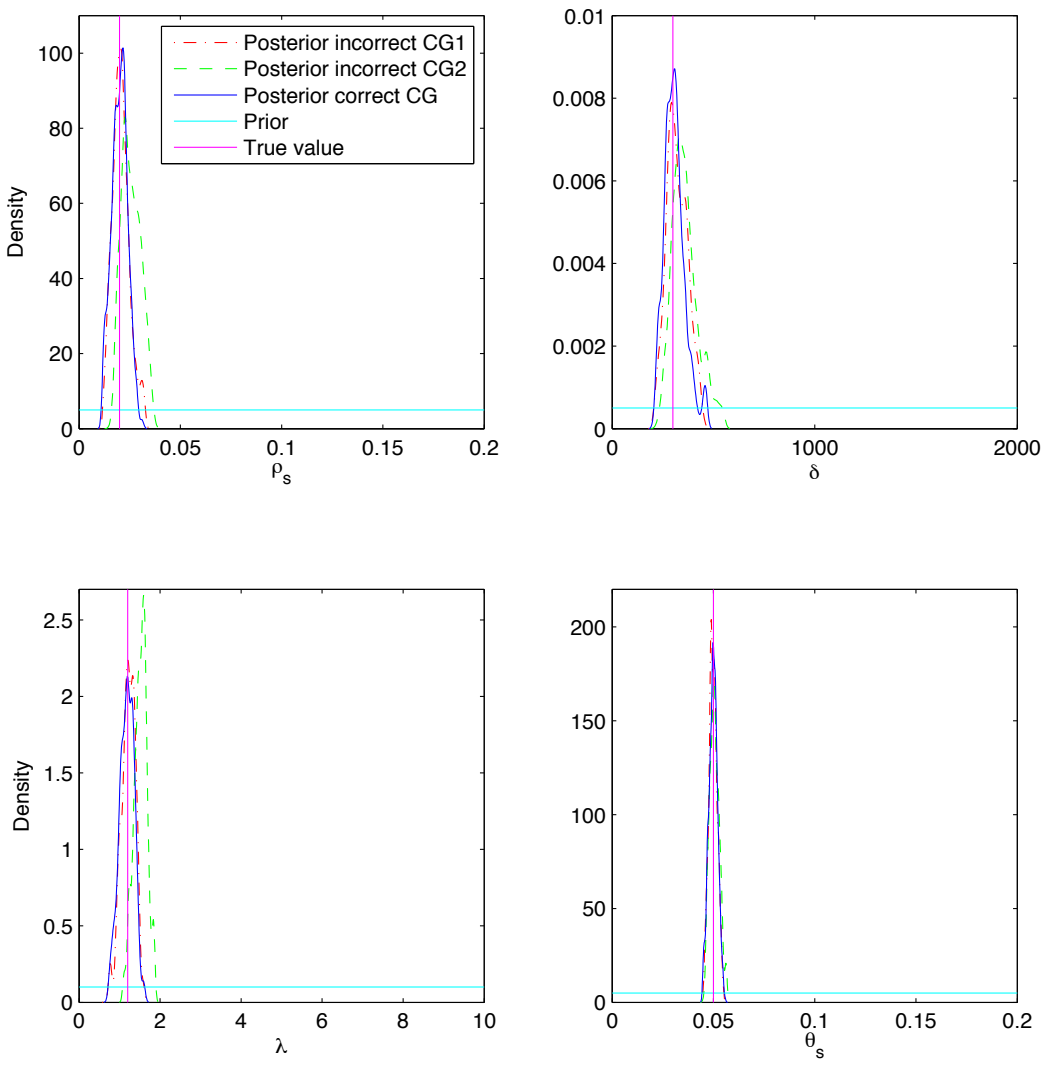


Figure 3.10: Estimated marginal posterior densities of the parameters for the simulated dataset using the two incorrect clonal gnealogies of Figure 3.9b. In both cases the true parameters used for simulation are well within the 95% credible intervals of the posterior densities.

then used our method to infer the parameter values for each of the 11 datasets. Figures 3.11 and 3.12 show the marginal posterior density for each of the 11 datasets. For values of ρ_s or δ equal to zero, there are no recombination events. In such cases either ρ_s and δ can be close to zero while the other parameter and λ can change freely. In addition extremely high values of λ lead to patterns similar to that of no recombination case. Such instances are easily recognised as LD and G4 plots are straight lines and therefore could be excluded from further analysis. For all other reasonable values of ρ_s, δ, λ and θ_s, posterior range is much tighter than the prior range. This shows that our inference method works as expected and that inference is possible for a wide range of parameter values.

3.3.3 Application to *Bacillus* dataset

We applied our method to estimate recombination properties based on a core alignment of 13 whole genomes of *Bacillus cereus*, *B. anthracis*, *B. thuringiensis* and *B. weihenstephanensis* which are all closely related including the first genome of this clade to be fully sequenced (Ivanova *et al.*, 2003). This is the same data as previously analysed by Didelot *et al.* (2010), thus allowing comparisons between the two analyses to be drawn. This previous analysis was performed using the ClonalOrigin model, which does not account for the bias in recombination. Nevertheless, the posterior distribution of recombination events contained a clear excess of recombination between close relatives (cf. Figure 5 of Didelot *et al.* 2010). Table 3.2 shows the year and country that these isolates were collected. Most of the samples are clinical isolates which come from North America and Europe. There are three isolates from China, Namibia and Iraq which are not clinical isolates. The amount of sequence divergence

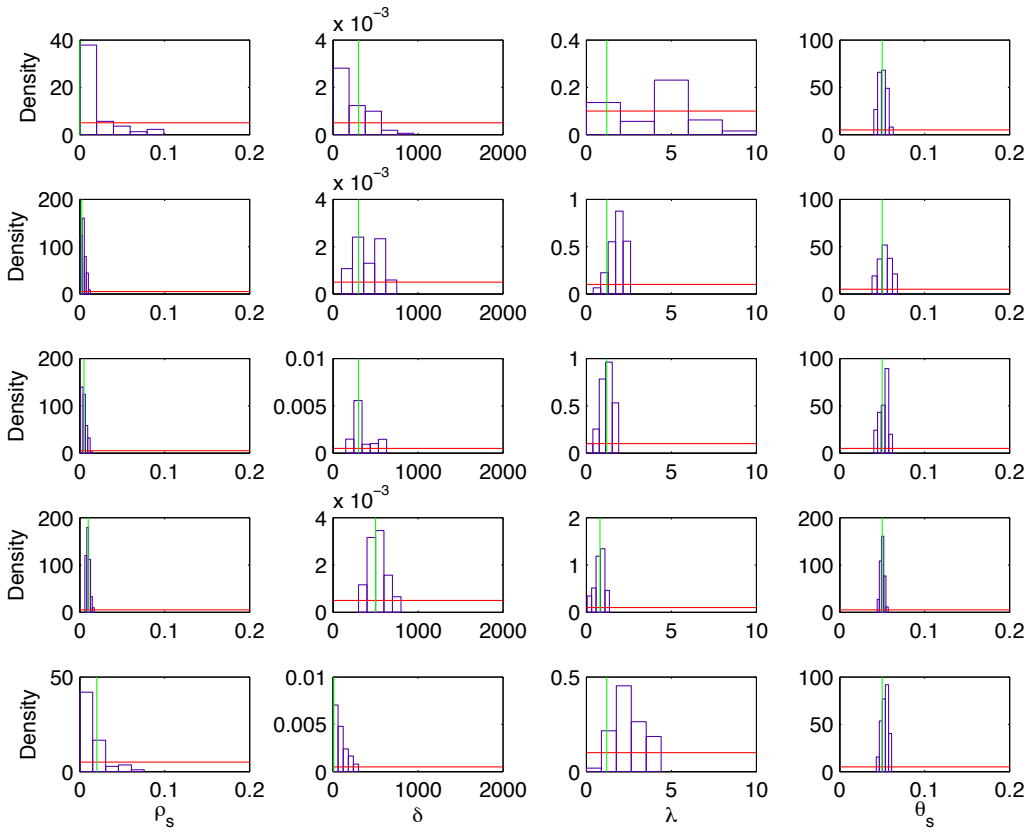


Figure 3.11: Marginal densities for five of the eleven simulated datasets. The clonal genealogy of Figure 3.7a with a range of parameters were used to simulate additional datasets to test our model. The first and last row highlight the fact that for values of ρ_s and or δ equal to zero (no recombination), either parameter can be close to zero while the other parameter and λ can change freely. In addition extremely high values of λ leads to patterns similar to that of no recombination case.

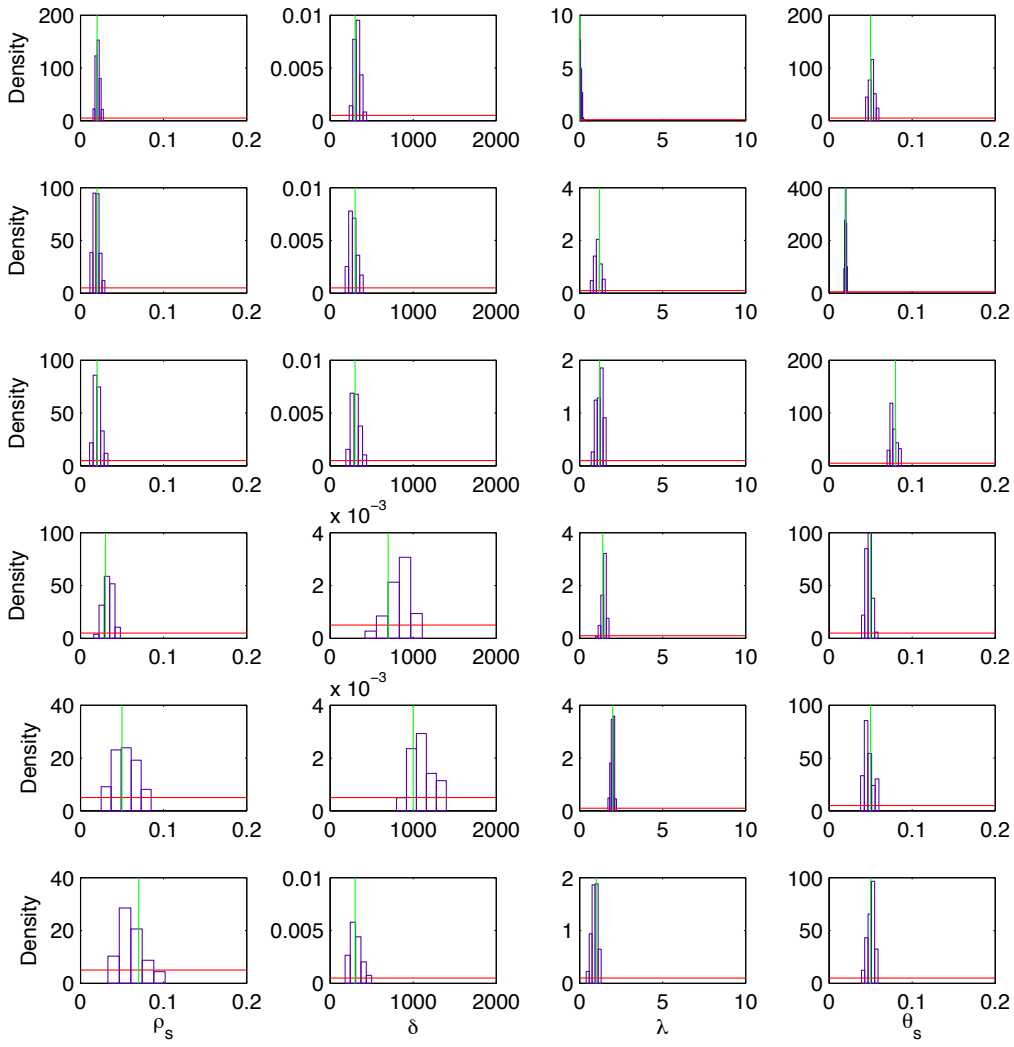


Figure 3.12: Marginal densities for six of the eleven simulated datasets. For all of the parameter values used in simulation the 95% credible interval includes the true value. The variance of the posterior density of the parameters is much smaller than that of the prior densities on the parameters.

between closest isolates is around 1% or in the region of 30,000 SNPs in the core genome of around 3,600,000 bases. This means the most recent common ancestor for the isolates is tens of thousands of years ago and the tree represents long evolutionary history. These isolates are very unlikely to have been in close geographical proximity and they represent the whole diversity of the species.

Figure 3.13 shows the clonal genealogy of the data that was estimated by DIDELOT *et al.* (2010) using ClonalFrame (DIDELOT and FALUSH, 2007). A unique tree topology with little uncertainty in the branch lengths was reconstructed. Figure 3.2 shows the LD and G4 plots for this dataset. Three points on the plots were selected to be used as summary statistics, with distance between the pairs of sites at 50, 200 and 2000 bp. The mean r^2 and G4 for pairs of sites at these distances were respectively (0.2738, 0.2493, 0.2270) and (0.0679, 0.0808, 0.0932), the proportion of segregating sites in this dataset was $S = 0.174$ and the proportion of homoplasic and clade homoplasic sites were respectively 0.29 and 0.15.

We chose uniform priors for all model parameters on the following ranges: $\rho_s \in [0, 0.2], \delta \in [0, 2000], \lambda \in [0, 4]$ and $\theta_s \in [0, 0.2]$. Several independent ABC-MCMC chains were run with similar results. The histograms on Figure 3.14 show the marginal posterior densities for the estimated parameters. The posterior mean for the recombination rate ρ_s is 0.077 with CI=0.036-0.127. The posterior mean of the recombination tract length δ was 152 bp with CI=74-279. The posterior mean of the rate of bias in recombination λ was estimated to be 1.32 with CI=0.812-1.788. The posterior mean of the mutation rate θ_s was 0.0528 with CI=0.0437-0.0640. The estimates of θ_s and δ were in agreement with previous estimates (median of $\theta_s = 0.0438$ and $\delta = 236$; DIDELOT *et al.* 2010). However, this previous analysis had

	Genome	Clade	Length	GenBank	Country	Year	Comment
1	*B. anthracis* Ames 0581	1	5503926	NC_007530	USA	1981	Isolated from a dead cow
2	*B. cereus* 03BB102	1	5449308	NC_012472	USA	2003	Fatal Pneumonia 39 year old male
3	*B. cereus* AH187	1	5599857	NC_011658	UK	1972	Outbreak of food poisoning
4	*B. cereus* AH820	1	5588834	NC_011773	Norway	1995	Periodontal pocket of 76 year old female
5	*B. cereus* ATCC 10987	2	5432652	NC_003909	Canada	1930	isolated from cheese spoilage
6	*B. cereus* ATCC14579	2	5427083	NC_004722			
7	*B. cereus* B4264	2	5419036	NC_011725	USA	1969	Fatal Pneumonia male patient
8	*B. cereus* G9842	1	5736823	NC_011772	USA		Isolated from stool of food poisoning patient
9	*B. cereus* Q1	1	5506207	NC_011969	China		Isolated from subsurface oil reservoir in North-eastern China
10	*B. cereus* ZK	1	5843235	NC_006274	Namibia		Isolated from a swab of a zebra carcass
11	*B. thuringiensis* Al Hakam	1	5313030	NC_008600	Iraq		Collected by UN special mission at a suspected bio-weapon facility
12	*B. thuringiensis* konkukian	1	5314794	NC_005957			Isolated from a necrotic human wound
13	*B. weihenstephanensis* KBAB4	3	5872743	NC_010184	France		Isolated from forest soil

Table 3.2: Genomes of the *Bacillus* group used in this study

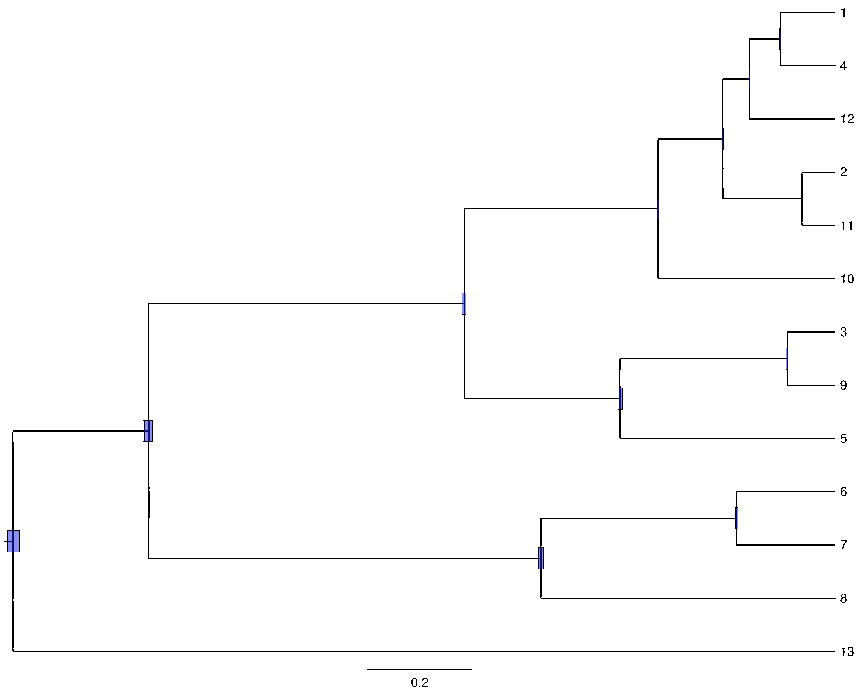


Figure 3.13: Clonal genealogy of the *Bacillus* dataset inferred by ClonalFrame. The blue bars represent the uncertainty on the age of the internal nodes.

estimated that recombination was significantly less frequent ($\rho_s = 0.017$; DIDELOT *et al.* 2010). In this previous study, the recombination rate had probably been underestimated as a result of not accounting for the bias in recombination. In a model with biased recombination, a larger fraction of recombination events are between close relatives and therefore have little effect, and would tend to go undetected by methods that do not account for it. The relative impact of recombination to mutation r/m (GUTTMAN and DYKHUIZEN, 1994; VOS and DIDELOT, 2009) was estimated using Equations 3.10 and 3.11. r/m had a mean of 3.4 with CI of 1.6-6.7 (Figure 3.15). This is slightly higher than the previous estimate from ClonalFrame (mean of $r/m = 2.41$; DIDELOT *et al.* 2010).

The posterior distributions of the four model parameters were significantly correlated as shown by the scatter plots in Figure 3.14. θ_s had moderate levels of negative correlation with ρ_s (Pearson's linear correlation coefficient, $r = -0.26$, $p = 1.3 \times 10^{-16}$), δ ($r = -0.12$, $p = 2.5 \times 10^{-4}$) and λ ($r = -0.16$, $p = 3.5 \times 10^{-7}$). The strongest associations however were the positive correlation of ρ_s with λ ($r = 0.83$, $p = 2.0 \times 10^{-253}$) and δ with λ ($r = 0.75$, $p = 1.3 \times 10^{-178}$). ρ_s and δ where also slightly correlated ($r = 0.34$, $p = 3.2 \times 10^{-29}$). Since higher values of λ translate into a higher bias in recombination (where recombination occurs between more similar isolates) and therefore a smaller effect of recombination, it is logical that there is to some extent a trade-off between smaller ρ_s and λ on one hand (meaning less recombination with more effect per recombination) and higher ρ_s and λ on the other hand (meaning more recombination with less effect per recombination). Likewise, higher values of δ indicate larger recombination events and therefore a higher effect per event, which explains the trade-off between λ and δ.

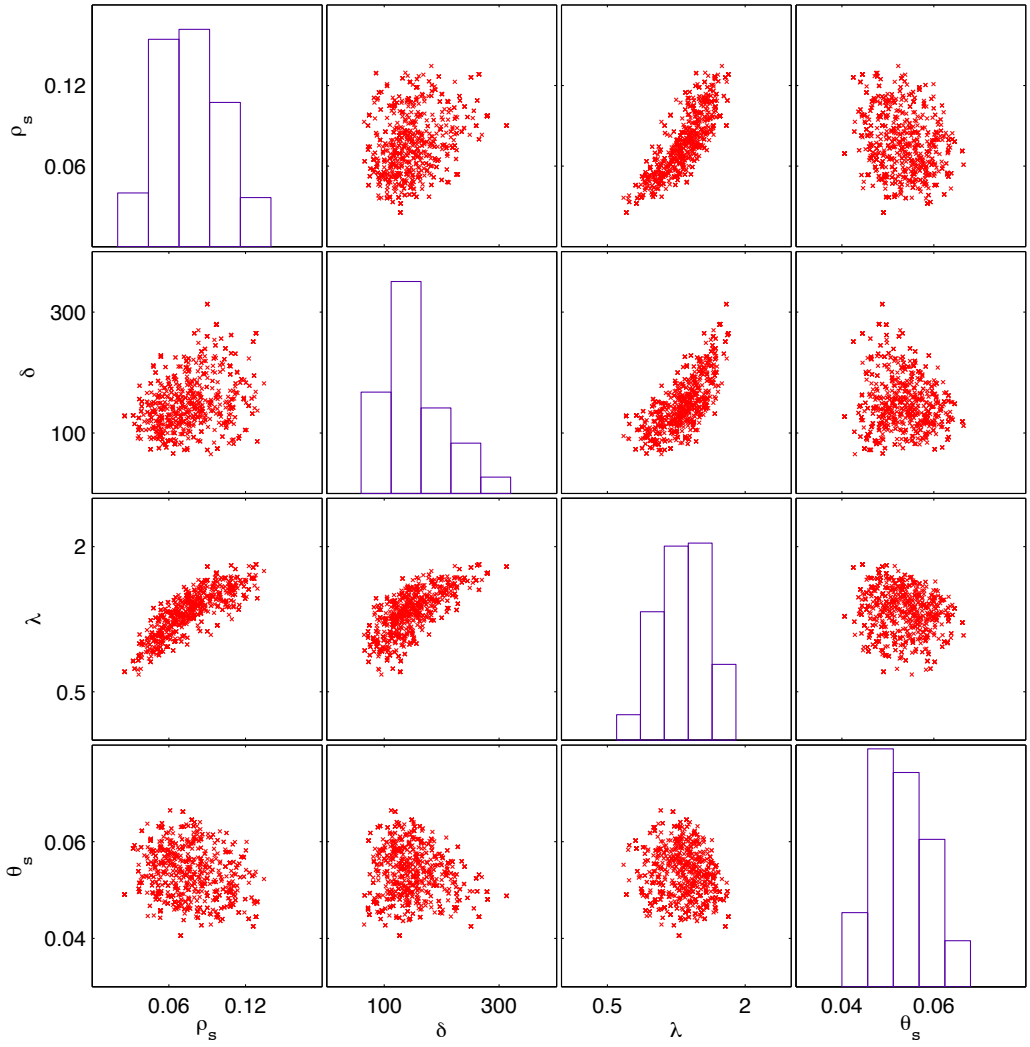


Figure 3.14: Posterior distributions of model parameters for the *Bacillus* dataset. The histograms show the marginal posterior distributions of each parameter whereas the scatter plots show their joint posterior distributions. The posterior distribution of the model parameters are significantly correlated. Higher values of λ translate into a higher bias in recombination and therefore a smaller effect of recombination. Therefore it is logical to have a trade-off between smaller values of ρ_s and λ on one hand and larger values of ρ_s and λ on the other hand. The same relationship would hold for δ and λ.

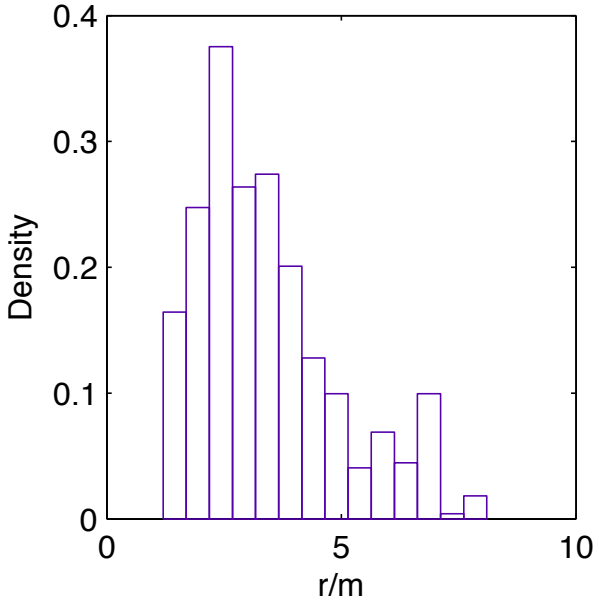


Figure 3.15: Posterior density of the relative impact of recombination to mutation r/m for the *Bacillus* dataset. It was estimated using Monte Carlo simulation as outlined in section 3.2.6. The mean of r/m is 3.4 with a 95% credible interval of 1.6-6.7. On average recombination adds 3.4 times more polymorphism into the dataset than mutations.

In order to test the fit of our model with biased recombination to the observed data, we considered the posterior predictive distribution of three additional summary statistics, i.e. their distribution when parameters are drawn from the posterior sample (Gelman *et al.*, 1996). These summary statistics had not been used in inference, but were similar in principle to the clade homoplasy statistic previously defined. The *Bacillus* clonal genealogy was divided into four distinct clades. One of these clades had a single member which was ignored. We measured the amount of clade homoplasy between the other three clades and used them as posterior predictive summary statistics. This Bayesian model criticism approach has been used in several previous ABC studies (Thornton and Andolfatto, 2006; Morelli *et al.*, 2010). We found that the observed values of the three summary statistics were contained within the boundaries of the posterior predictive distributions (Figure 3.16). Our model with biased recombination is therefore able to reproduce the observed summary statistics and represents a good fit to the data. However the observed values of clade homoplasies in the *Bacillus* data set is on the lower end of the predictive posterior distribution. This could have several reasons. Firstly, our model is an approximation and gross simplification of the reality and therefore it will not capture all aspects of the data well. In addition, assuming that the model parameters are well estimated and the model is a good fit, if the branch lengths of the genealogy have been underestimated, we would expect this pattern. Underestimated branch lengths would result in more recombination between clades than in the observed data. In this study the clonal genealogy was estimated using ClonalFrame which uses a free recombination model and ignores within population recombination.

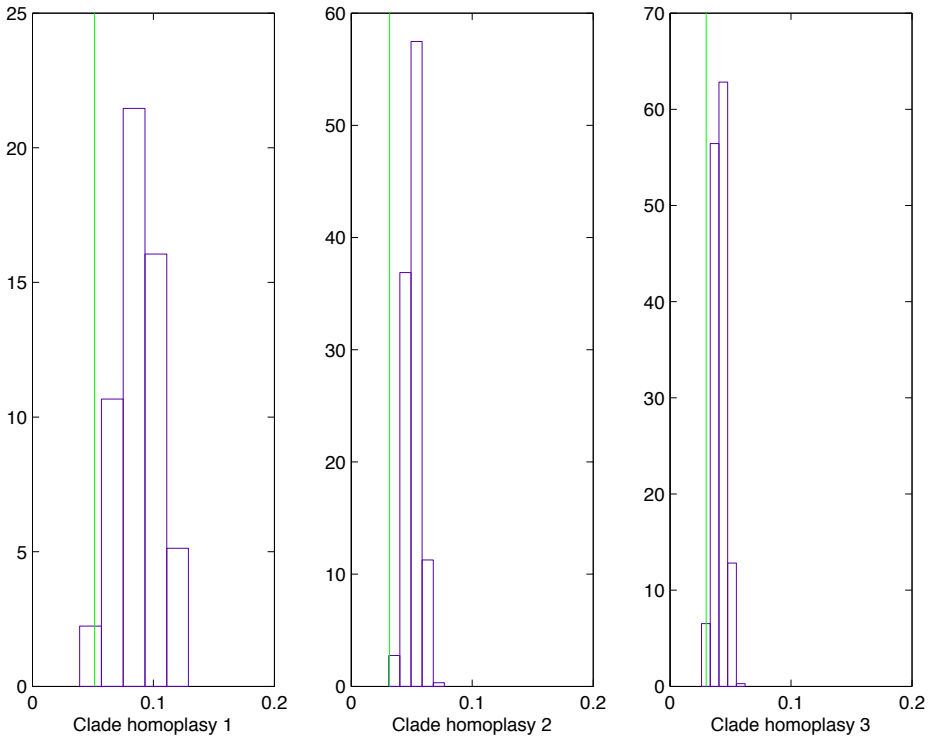


Figure 3.16: Posterior predictive distributions of the three additional clade homoplasy summary statistics for the application to the *Bacillus* dataset. The green lines represent the observed values. The observed values of the three summary statistics are contained within the boundaries of the posterior predictive distributions.

3.3.4 Comparison with experimental studies

Several experimental studies have demonstrated a log-linear relationship between sequence divergence and frequency of recombination (Roberts and Cohan, 1993; Zawadzki *et al.*, 1995; Vulić *et al.*, 1997; Majewski *et al.*, 2000). These results are summarized in Figure 1A of Fraser *et al.* (2007), which shows that different bacterial species have a similar log-linear relationship, with a coefficient around 20. To compare our results on biased recombination to these previous experimental studies, we need to compute the relative rate of recombination between two isolates as a function of their homology. Since this equates to considering recombination between two cells living at the same time, the first part of Equation 3.3 is equal to one and the probability of recombination is proportional to $\exp(-\lambda D)$ where D is the distance between the donor and the recipient cells in coalescent unit of time. The expected amount of sequence divergence π between two genomes separated by a branch of length D is $\pi = \theta_s D/2$ which implies that $D = 2\pi/\theta_s$, and therefore we obtain that the rate of recombination between two cells is proportional to $\exp(-2\lambda\pi/\theta_s)$. The frequency of recombination has therefore a log-linear relationship with sequence divergence in our biased recombination model, with coefficient (measured on a log of base 10 as in previous studies) equal to $2\lambda/(\theta_s \ln(10))$. In the case of the *Bacillus* application above, this coefficient had mean equal to 22.1 with CI=12.5-32.0. Our estimate for the rate of bias in recombination is therefore slightly higher than the rate of homology dependency of recombination that was found in previous experimental studies.

3.4 Discussion

Linkage disequilibrium, G4 and homoplasy are often interpreted informally as evidence of recombination. We have introduced a flexible statistical framework to interpret the values of these statistics calculated from whole bacterial genomes. Our underlying model is based on an approximation to the coalescent with gene-conversion (Didelot *et al.*, 2010) which has the advantage to be sequentially Markovian along the genomes. This allows to simulate patterns of LD and G4 through sampling of many pairs of sites at given distances, which takes only a small fraction of the computational power that would be needed to simulate large segments of DNA. Approximate Bayesian Computation (Pritchard *et al.*, 1999; Beaumont *et al.*, 2002) was used to perform inference under this bacterial population genomic model. This approach offers great flexibility to implement extensions of the model like the one we presented in Equation 3.3 to account for the biased recombination, simply by modifying the way simulation is performed without the need to compute a new likelihood function. We applied our method to simulated datasets and a real dataset consisting of 13 whole genomes of *Bacillus*. We showed that this data contains evidence that the recombination process depends on the evolutionary distance between donors and recipients, and measured the strength of this relationship. Our model is robust to slight misspecification of the clonal genealogy, but gross inaccuracies would lead to misleading results.

Evidence for a higher rate of recombination within than between the three major clades of *Bacillus* was first presented using multi-locus sequence typing data, by searching *a posteriori* for the most likely origin of ClonalFrame recombination segments (Didelot *et al.*, 2009a). This

approach was also applied to genomic data from *Salmonella enterica*, and more recombination was found within five lineages than between them (Didelot *et al.*, 2011a). However, this method is not very powerful, because ClonalFrame does not look for potential donors of the recombination events, and therefore is better able to detect recombination coming from further away (Didelot and Falush, 2007). A better approach is the one implemented in ClonalOrigin (Didelot *et al.*, 2010), where the source of recombined fragments is inferred jointly with the recombination events rather than relying on a post-processing step. By comparison of the number of recombination events observed between pairs of branches and expected under the prior model, recombination was found to happen more often between members of the same *Bacillus* clades (Didelot *et al.*, 2010). Similar results have been obtained using the same technique in other organisms, such as *Sulfolobus islandicus* (Cadillo-Quiroz *et al.*, 2012) or *Escherichia coli* (Didelot *et al.*, 2012). However, this is still not fully satisfactory from a statistical point of view, since the analysis is done using a prior model where recombination does not depend on evolutionary distance which is proved to be incorrect by the posterior distribution of events. For this reason, this approach does not allow to estimate the strength of bias in recombination, since the posterior depends on both the prior (where this parameter is zero) and the observed data (which contains evidence that this parameter is non-zero). The best statistical approach is therefore the one we presented here, where the model explicitly incorporates this important parameter, so that we can use Bayesian statistics to formally test whether it is significantly different from zero, if and if so estimate its value.

We estimated the coefficient for the log-linear relationship between

recombination rate and the effective sequence divergence to be around 22 in *Bacillus*. This is slightly higher than previous estimates based on laboratory experiments which were around 20 (Fraser *et al.*, 2007). This higher coefficient could be due to the fact that laboratory experiment only measure the rate of recombination between two bacteria when they are brought into contact, whereas there are factors in nature, such as geographical or ecological structuring of the population, that would increase the sexual isolation between distantly related bacteria (Majewski, 2001). Yet, this coefficient is far lower than the value of 300 predicted by population genetics models to be required in order for recombination to be on one hand a strong cohesive force between highly homologous bacteria and on the other hand very rare between diverged bacteria, thus resulting in clusters of diversity which could be considered to represent separate species (Falush *et al.*, 2006; Hanage *et al.*, 2006; Fraser *et al.*, 2007; Achtman and Wagner, 2008). To test this hypothesis further, we used the Monte-Carlo simulation detailed in section 3.2.7 to see the effect of the next evolutionary events likely to happen to any one of the *Bacillus* genomes in our dataset. We found that for all except the most closely related pairs of genomes, future recombination events would result in convergence, i.e. a reduction of the genetic distance (Figure 3.17, bottom part). However, we also found that mutation would increase the genetic distance between any pair of genomes at a much higher rate than recombination would reduce it (Figure 3.17, top part). We conclude that all pairs of genomes are likely to diverge in the near future, since the convergence effect of recombination will not be sufficient to compensate the divergence effect of mutation. Convergence via recombination is likely to be restricted to rare situations where strong

selective or ecological factors are involved, such as found in the convergence of *Salmonella enterica* serovars Typhi and Paratyphi A (Didelot *et al.*, 2007) or the convergence of *Campylobacter jejuni* and *coli* (Sheppard *et al.*, 2008).

3.5 Acknowledgments

An abridged version of this chapter appeared as Ansari and Didelot (2014). With much thanks to Rory Bowden, Alison Etheridge, Richard Everitt, Daniel Falush, Simon Myers, Daniel Wilson and two anonymous reviewers for providing helpful comments and suggestions.

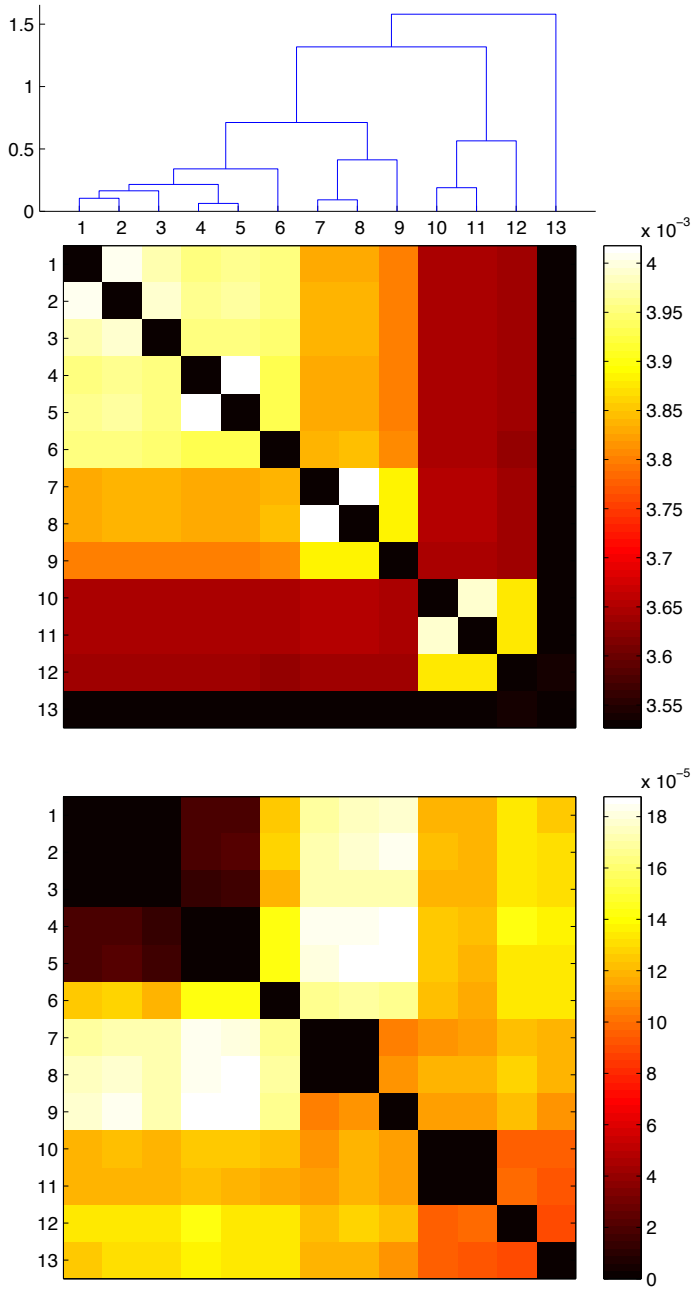


Figure 3.17: Prediction of the future effect of mutation and recombination on the genetic distance between pairs of *Bacillus* genomes. The heat map at the top indicates the rate at which mutation will increase the distance between all pairs of genomes (i.e. pairwise divergence). The heat map at the bottom indicates the rate at which recombination will decrease these same distances (i.e. pairwise convergence). For closely related isolates recombination leads to divergence which is shown as zero divergence. The rate which mutation causes divergence is an order of magnitude higher than the rate at which recombination leads to convergence. Thus in these isolates, the overall short term impact of recombination and mutation is divergence of the isolates.

Chapter 4

Ecologically Distinct Clades

4.1 Introduction

The concept of species and speciation in prokaryotes is controversial (Doolittle and Papke, 2006; Cohan and Perry, 2007; Achtman and Wagner, 2008; Fraser *et al.*, 2009). Firstly, majority of the bacterial species cannot be cultivated at present (Amann *et al.*, 1995; Hugenholtz *et al.*, 1998) and thus cannot be studied in laboratory. Secondly, bacterial diversity is huge (Dykhuizen, 1998; Gans *et al.*, 2005) and only a small fraction of them are known to us. Thirdly, although bacteria are clonal organisms, recombination can potentially occur between distantly related strains (Levin and Cornejo, 2009) and genetically homogenise populations. Currently bacteria are assigned to the same species if their reciprocal, pairwise DNA re-association values are bigger than 70% in DNA-DNA hybridisation experiments (Achtman and Wagner, 2008). The 70% is an arbitrary cutoff threshold and this definition does not take into account the evolutionary forces that are important in demarcating species.

Recently incipient sympatric speciation has been demonstrated in bacteria (Ferris *et al.*, 2003; Sikorski and Nevo, 2005; Johnson *et al.*, 2006). To infer incipient sympatric speciation, one has to be able to distinguish between ecologically distinct populations in the same community. Differential evolutionary forces acting on such populations can increase the genetic distance among them that can be interpreted as incipient speciation.

Intuitively, one would expect genetically similar strains to have similar functions in a community. This is because bacteria are clonal and usually the recombination rate is much lower than the mutation rate. Therefore strains with a common ancestor nearer to the present share more of their genetic material relative to a group of strains with common ancestor further back in time. Thus closely related isolates are more likely to share the genetic elements that causes adaptation to a specific niche or habitat. There is evidence for phylogenetic clusters filling distinct ecological niches in the community (Ferris *et al.*, 2003; Sikorski and Nevo, 2005; Horner-Devine and Bohannan, 2006; Johnson *et al.*, 2006). However, theoretical studies have suggested that neutral evolution alone can produce phylogenetic clusters within a population (Falush *et al.*, 2006; Fraser *et al.*, 2009). These studies assume extremely high rates for the log-linear relationship between recombination rate and sequence divergence which does not seem realistic for most bacterial species. In addition if phylogenetic clusters are due to neutral evolution, one would not expect to see significant differences in the preference for habitats in different clusters.

Assuming clades reasonably correlate with ecology and function within a community, it is unclear what level of the hierarchy delineates between such distinct ecological populations. Two recently developed software

packages that demarcate phylogenetic clusters as ecologically distinct populations are AdaptML (Hunt *et al.*, 2008) and Ecotype Simulation (Koeppel *et al.*, 2008). AdaptML takes phylogeny and information about habitat as input and builds a hidden Markov model for the evolution of habitat associations. It then uses maximum likelihood estimation and some ad hoc rules to infer ecologically distinct populations. Ecotype Simulation models the sequence diversity within a bacterial clade as the evolutionary result of three stochastic processes: net rate of creation of new ecotype, periodic selection (whereby all linages but one within an ecotype are eliminated) and genetic drift(the coalescence of a pair of lineages within an ecotype into one lineage). It then uses Monte Carlo simulation to estimate the model parameters that produce sequence diversity similar to the ones in the observed data and to estimate the number of ecologically distinct populations within a clade. It is important to note that Ecotype Simultion does not use any information about the habitats and the putative ecotypes identified are subsequently investigated to detect ecological distinctness.

Here we propose a new statistical model for occurrence of adaptive events on a phylogenetic tree and how these events affect the distribution of observed environmental categories. This parametrisation allows us to demarcate the boundaries of phylogenetic clusters that have distinct distribution over the observed environmental categories. We use MCMC to sample from the posterior distribution of the model parameters. We conduct several simulation studies to measure the power of our method and its sensitivity to model parameters. We then present the application of our method to two real datasets.

4.2 Model

In most realistic ecological sampling situations for bacteria, true habitats are not known and can only be associated with projections into measurable environmental categories. Therefore we distinguish between a true habitat and an environmental category. The habitats are not known and are associated with distinct distribution of isolates among the observed environmental categories.

Given a random sample of isolates, their phylogenetic tree and observed environmental categories from a community, we would like to know if the environmental categories distinctly distribute among the phylogenetic clusters and what level of the hierarchies on the tree partitions these clusters.

We assume habitat adaptation events happen as a Poisson process with rate λ on the branches of the tree. Given that there are K observed environmental categories in the community, we model each habitat adaptation event as a probability mass function (PMF) $\boldsymbol{q} = (q_1, \ldots, q_K)$ which gives the probability of a strain with the given adaptation to be found in each of the K environmental categories (the probabilities can be interpreted as preference for the environmental categories). In addition we also make the reasonable assumption that given the adaptive events on the tree, isolates are independent of each other in their preference for environmental categories.

Figure 4.1 illustrates the model. There are two observed environmental categories. The observed environmental categories for each isolate is shown on the tips of the tree. An adaptation event has occurred on branch 1 and another adaptation event on branch 2. These two events have divided the tree into three sections (white, green and red). All isolates in the same

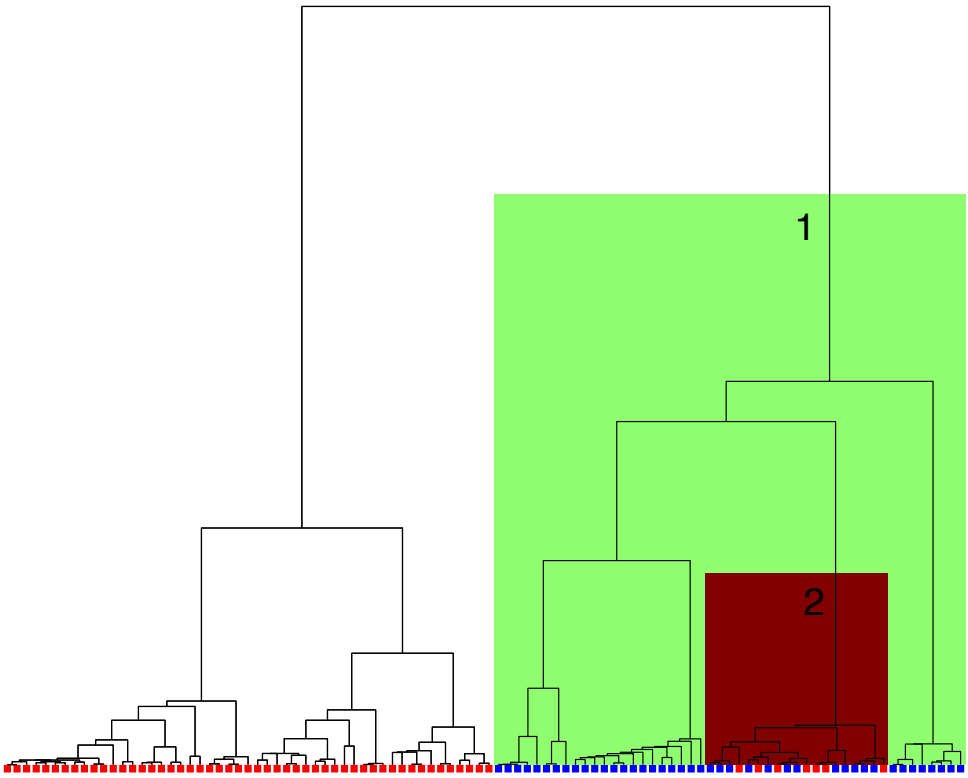


Figure 4.1: Illustration of habitat adaptation model. Branches 1 and 2 have adaptation events on them. These two event have divided the tree into three sections (White, green and red). Isolates in the same section (with the same adaptation) have the same distribution on the environmental categories. The green section is the parent of the red section and it is the child of the white section.

section have the same adaptation and therefore the same distribution on the environmental categories. The red section is a child of the green section and the green section is the parent of the red section. In absence of any adaptive events all isolates on the tree will have the same distribution over the observed environmental categories.

4.2.1 Single ecological lineage

Let us assume that we have a random sample of N isolates from a bacterial community with K observed environmental categories. In

addition we assume the phylogenetic tree of the isolates is known. If there are no adaptation events on the branches of the phylogenetic tree, all isolates independent of their location on the tree have the same distribution over the K environmental categories. Likelihood of the observed environmental categories of the isolates is given by:

$$p(D|\boldsymbol{q}) = \prod_{i=1}^{K} q_i^{x_i} \qquad \sum_{i=1}^{K} q_i = 1, \quad \sum_{i=1}^{K} x_i = N \tag{4.1}$$

Where D is the observed environmental categories, $\boldsymbol{q} = (q_1, \ldots, q_K)$ is the distribution of the isolates over the K environmental categories and $\boldsymbol{x} = (x_1, \ldots, x_k)$ is the number of observed isolates in each environmental category. The posterior distribution of the model parameters is given by:

$$p(\boldsymbol{q}|D) = \frac{p(D|\boldsymbol{q})p(\boldsymbol{q})}{p(D)} \tag{4.2}$$

We use a Dirichlet distribution as prior for $\boldsymbol{q}$ with hyper parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ as it is the conjugate prior for the multinomial likelihood:

$$p(\boldsymbol{q}) = \frac{1}{Z(\boldsymbol{\alpha})} \prod_{i=1}^{K} q_i^{\alpha_i - 1} \tag{4.3}$$

Where $Z(\boldsymbol{\alpha})$ is defined as:

$$\frac{1}{Z(\boldsymbol{\alpha})} = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \tag{4.4}$$

Where $\Gamma(.)$ is the Gamma function and we define $\alpha_0 = \sum_{i=1}^{K} \alpha_i$. Therefore the posterior density is given by:

$$\begin{aligned} p(\boldsymbol{q}|D) &= \frac{\prod_{i=1}^{K} q_i^{x_i} \prod_{i=1}^{K} q_i^{\alpha_i - 1}}{Z(\boldsymbol{\alpha})p(D)} \\ &= \frac{\prod_{i=1}^{K} q_i^{x_i + \alpha_i - 1}}{Z(\boldsymbol{\alpha})p(D)} \\ \boldsymbol{q}|D &\sim \mathrm{Dir}(\boldsymbol{x} + \boldsymbol{\alpha}) = \mathrm{Dir}(x_1 + \alpha_1, \ldots, x_K + \alpha_K) \end{aligned} \tag{4.5}$$

The evidence for the model (marginal likelihood) can be calculated as:

$$\begin{aligned} p(D) &= \int_{\boldsymbol{q}} p(D|\boldsymbol{q})p(\boldsymbol{q})d\boldsymbol{q} \\ &= \frac{1}{Z(\boldsymbol{\alpha})} \int_{q_1} \cdots \int_{q_K} q_1^{x_1+\alpha_1-1} \cdots q_K^{x_K+\alpha_K-1} dq_1 \cdots dq_K \\ &= \frac{Z(\boldsymbol{x} + \boldsymbol{\alpha})}{Z(\boldsymbol{\alpha})} \\ &= \frac{\frac{\prod_{i=1}^{K} \Gamma(x_i+\alpha_i)}{\Gamma(N+\alpha_0)}}{\frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}} \\ &= \frac{\Gamma(\alpha_0) \prod_{i=1}^{K} \Gamma(x_i + \alpha_i)}{\Gamma(N + \alpha_0) \prod_{i=1}^{K} \Gamma(\alpha_i)} \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(N + \alpha_0)} \prod_{i=1}^{K} \frac{\Gamma(x_i + \alpha_i)}{\Gamma(\alpha_i)} \end{aligned} \tag{4.6}$$

We will assume a uniform Dirichlet prior on $\boldsymbol{q}$ where all the hyper parameters are set to one. Therefore the marginal likelihood is given by:

$$p(D) = \frac{\Gamma(K)}{\Gamma(N + K)} \prod_{i=1}^{K} \Gamma(x_i + 1) \tag{4.7}$$

4.2.2 Two or more ecologically distinct lineages

Given that N isolates have been sampled from the community, there are $2N-2$ branches on the phylogenetic tree that could have adaptive events on them. We define $\boldsymbol{b} = (b_1, \ldots, b_{2N-2})$ as a binary vector with $2N-2$ elements which represent the branches of the tree. If a branch on the tree has at least one adaptive event on it, the corresponding element in $\boldsymbol{b}$ will be one and zero otherwise. It is important to note that under our model there can be more than one adaptive event on a branch, but only the last adaptive event on the branch will have an effect on the likelihood. Assuming that $\boldsymbol{b}$ divides the tree into m sections as in shown in Figure 4.1, the likelihood of the observed environmental categories of the isolates is given by:

$$p(D|\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m, \boldsymbol{b}) = \prod_{j=1}^{K} q_{1j}^{x_{1j}} \cdots \prod_{j=1}^{K} q_{mj}^{x_{mj}} \tag{4.8}$$

where

$$\sum_{i=1}^{m}\sum_{j=1}^{K} x_{ij} = N \qquad \text{and} \qquad \sum_{j=1}^{K} q_{ij} = 1 \text{ for } i = 1, \ldots, m \tag{4.9}$$

Where $\boldsymbol{q}_i = (q_{i1}, \cdots, q_{iK})$ is the distribution of isolates in section i over the K environmental categories and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK})$ is the number of observed isolates in each category in section i.

As adaptive events are Poisson distributed on branches of the tree with rate λ, the prior probability of branch i of length l_i to have no adaptation and at least one adaptation event is given by:

$$\Pr(b_i = 0|\lambda) = e^{-\lambda l_i}$$

$$\Pr(b_i = 1|\lambda) = 1 - e^{-\lambda l_i}$$

Thus

$$\Pr(\boldsymbol{b}|\lambda) = \prod_{i=1}^{2N-2} (e^{-\lambda l_i})^{1-b_i}(1 - e^{-\lambda l_i})^{b_i} \tag{4.10}$$

We will assume Dirichlet uniform priors for $\boldsymbol{q}_i$s:

$$p(\boldsymbol{q}_i) = \Gamma(K) \tag{4.11}$$

Furthermore we will assume an exponential prior on λ. Using parsimony argument (acceptance of the simplest explanations that fits the data) we assume the prior expectation of the number of adaptive events on the tree to be one. As the number of adaptive events M are Poisson distributed on the branches of the tree with rate λ, expectation of M is given by:

$$\mathbb{E}(M) = \mathbb{E}(\lambda T) = 1 \implies \mathbb{E}(\lambda) = \frac{1}{T} \tag{4.12}$$

Where T is the total branch length of the tree. Equation 4.12 gives the hyper parameter of the exponential distribution for λ:

$$\lambda \sim \text{Exp}(T) \tag{4.13}$$

Now we are in a position to describe the posterior distribution of the model parameters. Let us assume that $\boldsymbol{b}$ divides the tree into m sections,

then the posterior distribution for $(\boldsymbol{q}_i, \ldots, \boldsymbol{q}_m), \boldsymbol{b}$ and λ is given by:

$$\begin{aligned} p(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m, \boldsymbol{b}, \lambda | D) &= \frac{p(D|\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m, \boldsymbol{b}) p(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m, \boldsymbol{b}, \lambda)}{p(D)} \\ &\propto p(D|\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m, \boldsymbol{b}) p(\boldsymbol{q}_1) \ldots p(\boldsymbol{q}_m) p(\boldsymbol{b}|\lambda) p(\lambda) \\ &\propto (\Gamma(K))^m \prod_{i=1}^{m} \prod_{j=1}^{K} q_{ij}^{x_{ij}} \prod_{s=1}^{2N-2} (e^{-\lambda l_s})^{1-b_s} (1 - e^{-\lambda l_s})^{b_s} T e^{-T\lambda} \end{aligned} \tag{4.14}$$

The dimensionality of the problem changes with $\boldsymbol{b}$. If $\boldsymbol{b}$ divides the tree into two sections then there are four parameters $\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{b}$ and λ to infer while if $\boldsymbol{b}$ divides the tree into three sections then there are five parameters $\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3, \boldsymbol{b}$ and λ to infer. This issue potentially can be addressed using reversible jump MCMC (Green, 1995). Instead we marginalise $\boldsymbol{q}_i$s which turns the problem into a constant dimension one. The marginal posterior density for $\boldsymbol{b}$ and λ assuming that $\boldsymbol{b}$ divides the tree into m sections is given by:

$$\begin{aligned} p(\boldsymbol{b}, \lambda | D) &= \frac{\int_{\boldsymbol{q}_1} \cdots \int_{\boldsymbol{q}_m} p(D|\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m, \boldsymbol{b}) p(\boldsymbol{q}_1) \cdots p(\boldsymbol{q}_m) d\boldsymbol{q}_1 \cdots d\boldsymbol{q}_m p(\boldsymbol{b}|\lambda) p(\lambda)}{p(D)} \\ &\propto (\Gamma(K))^m \prod_{i=1}^{m} \prod_{j=1}^{K} \int_0^1 q_{ij}^{x_{ij}} dq_{ij} T e^{-T\lambda} \prod_{s=1}^{2N-2} (e^{-\lambda l_s})^{1-b_s} (1 - e^{-\lambda l_s})^{b_s} \\ &\propto (\Gamma(K))^m \prod_{i=1}^{m} \frac{\prod_{j=1}^{K} \Gamma(x_{ij} + 1)}{\Gamma(K + \sum_{j=1}^{K} x_{ij})} T e^{-T\lambda} \prod_{s=1}^{2N-2} (e^{-\lambda l_s})^{1-b_s} (1 - e^{-\lambda l_s})^{b_s} \end{aligned} \tag{4.15}$$

Although we have integrated out the $\boldsymbol{q}_i$s from the posterior distribution, we can easily sample from their posterior distribution given $\boldsymbol{b}$. Assuming that we have a sample from the posterior distribution of $\boldsymbol{b}$ and λ, the posterior distribution of $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m | D, \boldsymbol{b}, \lambda$, assuming that $\boldsymbol{b}$ divides

the tree into m sections is given by:

$$\begin{aligned} p(\boldsymbol{q}_1,\ldots,\boldsymbol{q}_m|\boldsymbol{b},\lambda,D) &= \frac{p(D|\boldsymbol{b},\lambda,\boldsymbol{q},\ldots,\boldsymbol{q}_m)p(\boldsymbol{b},\lambda|\boldsymbol{q}_1,\ldots,\boldsymbol{q}_m)p(\boldsymbol{q}_1,\ldots,\boldsymbol{q}_m)}{p(D,\boldsymbol{b},\lambda)} \\ &\propto p(D|\boldsymbol{b},\boldsymbol{q}_1,\ldots,\boldsymbol{q}_m) \\ &\propto \prod_{i=1}^{m}\prod_{j=1}^{K} q_{ij}^{x_{ij}} \end{aligned} \tag{4.16}$$

Or in other words, given $\boldsymbol{b}$ each $\boldsymbol{q}_i$ is Dirichlet distributed with known parameters equal to one plus the number of observed isolates in each category in the section i of the tree.

4.2.3 MCMC moves

To sample from the posterior distribution of $\boldsymbol{b}$ and λ we will use MCMC. To do this we have to specify the proposals for updating $\boldsymbol{b}$ and λ and calculate the acceptance ratio. We will use a component-wise MCMC which updates each parameter separately. We will use a symmetric proposal for $\boldsymbol{b}$ where the proposed value $\boldsymbol{b}^\star$ is the same as $\boldsymbol{b}$ apart from a randomly chosen branch i for which $b_i^\star = 1 - b_i$. Therefore if the randomly chosen branch i has an adaptive event on it in $\boldsymbol{b}$, it will not have an adaptive event on it in $\boldsymbol{b}^\star$ and vice versa. Assuming that $\boldsymbol{b}^\star$ divides the tree into n sections the Metropolis-Hastings ratio for this move is given by $h(\boldsymbol{b},\boldsymbol{b}^\star)$:

$$\begin{aligned} h(\boldsymbol{b},\boldsymbol{b}^\star) &= 1 \wedge \frac{(\Gamma(K))^n \prod_{i=1}^{n} \frac{\prod_{j=1}^{K}\Gamma(x_{ij}^\star+1)}{\Gamma(K+\sum_{j=1}^{K}x_{ij}^\star)} Te^{-T\lambda}\prod_{s=1}^{2N-2}(e^{-\lambda l_s})^{1-b_s^\star}(1-e^{-\lambda l_s})^{b_s^\star}}{(\Gamma(K))^m \prod_{i=1}^{m} \frac{\prod_{j=1}^{K}\Gamma(x_{ij}+1)}{\Gamma(K+\sum_{j=1}^{K}x_{ij})} Te^{-T\lambda}\prod_{s=1}^{2N-2}(e^{-\lambda l_s})^{1-b_s}(1-e^{-\lambda l_s})^{b_s}} \\ &= 1 \wedge \frac{(\Gamma(K))^{n-m} \prod_{i=1}^{n} \frac{\prod_{j=1}^{K}\Gamma(x_{ij}^\star+1)}{\Gamma(K+\sum_{j=1}^{K}x_{ij}^\star)} \prod_{s=1}^{2N-2}(e^{-\lambda l_s})^{1-b_s^\star}(1-e^{-\lambda l_s})^{b_s^\star}}{\prod_{i=1}^{m} \frac{\prod_{j=1}^{K}\Gamma(x_{ij}+1)}{\Gamma(K+\sum_{j=1}^{K}x_{ij})} \prod_{s=1}^{2N-2}(e^{-\lambda l_s})^{1-b_s}(1-e^{-\lambda l_s})^{b_s}} \end{aligned} \tag{4.17}$$

Please note that the prior on λ cancelled out from the Metropolis-Hastings ratio. To update λ we propose from a normal density with the mean at the current value of λ and variance of 0.1:

$$\lambda^{\star}|\lambda \sim \mathcal{N}(\lambda, 0.1) \tag{4.18}$$

This proposal distribution is symmetric and therefore the Metropolis-Hastings ratio is given by:

$$\begin{aligned} h(\lambda, \lambda^{\star}) &= 1 \wedge \frac{(\Gamma(K))^m \prod_{i=1}^{m} \frac{\prod_{j=1}^{K} \Gamma(x_{ij}+1)}{\Gamma(K+\sum_{j=1}^{K} x_{ij})} T e^{-T\lambda^{\star}} \prod_{s=1}^{2N-2} (e^{-\lambda^{\star} l_s})^{1-b_s} (1-e^{-\lambda^{\star} l_s})^{b_s}}{(\Gamma(K))^m \prod_{i=1}^{m} \frac{\prod_{j=1}^{K} \Gamma(x_{ij}+1)}{\Gamma(K+\sum_{j=1}^{K} x_{ij})} T e^{-T\lambda} \prod_{s=1}^{2N-2} (e^{-\lambda l_s})^{1-b_s} (1-e^{-\lambda l_s})^{b_s}} \\ &= 1 \wedge \frac{e^{-T\lambda^{\star}} \prod_{s=1}^{2N-2} (e^{-\lambda^{\star} l_s})^{1-b_s} (1-e^{-\lambda^{\star} l_s})^{b_s}}{e^{-T\lambda} \prod_{s=1}^{2N-2} (e^{-\lambda l_s})^{1-b_s} (1-e^{-\lambda l_s})^{b_s}} \end{aligned} \tag{4.19}$$

When the proposal $\lambda^{\star}$ is less than zero, the move is rejected and the chain stays at λ.

4.2.4 Model selection

To compare our null model of a single ecological lineage (model 1) against the alternative model of two or more ecologically distinct lineages (model 2) on the tree, we calculate the Bayes factor (Kass and Raftery, 1995) for the two models. To do this we use reversible jump MCMC (Green, 1995) to sample from the joint distribution of $p((k, \boldsymbol{\theta}_k)|D)$. k is the index of the model and $\boldsymbol{\theta}_k$ is the parameters of model k. For each model k by marginalising over its parameters $\boldsymbol{\theta}_k$, its marginal likelihood $p(D|k)$ can be estimated. Marginal likelihood of model 1, $p(D|1)$ is calculated analytically using Equation 4.7. For the purpose of RJMCMC there are no parameters

in model 1 to infer, however there are two parameters in model 2 to infer (λ and $\boldsymbol{b}$).

For a move from model 1 to model 2, to match dimensions we generate two random variable u and $\boldsymbol{v}$ and map them such that $(\lambda^\star, \boldsymbol{b}^\star) = (u, \boldsymbol{v})$. In addition we set the proposal distribution for u and $\boldsymbol{v}$, $q(u, \boldsymbol{v})$ in model 1 to be the same as the prior distribution on λ and $\boldsymbol{b}$ in model 2. Thus for a proposed move from model 1 to model 2 we have:

$$q(u, \boldsymbol{v}) = q(u)q(\boldsymbol{v}|u) = -Te^{-Tu} \prod_{i=1}^{2N-2} (e^{-ul_i})^{1-v_i}(1 - e^{-ul_i})^{v_i} \tag{4.20}$$

Where we set

$$(\lambda^\star, \boldsymbol{b}^\star) = (u, \boldsymbol{v})$$

Where T is the total branch length of the tree and l_i is the length of branch i. Therefore the probability of acceptance of this move is given by:

$$\begin{aligned}
\alpha\left((1) \to (2, (\lambda^\star, \boldsymbol{b}^\star))\right) &= 1 \wedge \frac{p(2, (\lambda^\star, \boldsymbol{b}^\star)|D)p(2 \to 1)}{p(1|D)p(1 \to 2)q((u, \boldsymbol{v})|1)} \left|\frac{\partial(\lambda^\star, \boldsymbol{b}^\star)}{\partial(u, \boldsymbol{v})}\right| \\
&= 1 \wedge \frac{p(2, (u, \boldsymbol{v})|D)p(2 \to 1)}{p(1|D)p(1 \to 2)q((u, \boldsymbol{v})|1)} \times 1 \\
&= 1 \wedge \frac{p(D|(u, \boldsymbol{v}), 2)p((u, \boldsymbol{v})|2)p(2)p(2 \to 1)}{p(D|1)p(1)p(1 \to 2)q((u, \boldsymbol{v})|1)} \\
&= 1 \wedge \frac{p(D|(u, \boldsymbol{v}), 2)p(2 \to 1)}{p(D|1)p(1 \to 2)}
\end{aligned} \tag{4.21}$$

Where $p(D|1)$ is given by Equation 4.7. A move from model 2 with parameters $(\lambda, \boldsymbol{b})$ to model 1 is made deterministically and is accepted with probability:

$$\alpha((2, (\lambda, \boldsymbol{b})) \to (1)) = 1 \wedge \frac{p(D|1)p(1 \to 2)}{p(D|(\lambda, \boldsymbol{b}), 2)p(2 \to 1)} \tag{4.22}$$

Where we set $p(2 \to 1) = 0.05$ and $p(1 \to 2) = 0.5$ and we assume the prior

probabilities of the two models are equal $p(1) = p(2) = 0.5$.

4.2.5 Representation of the MCMC output

Consensus representation of $\boldsymbol{b}$

As $\boldsymbol{b}$ is a vector which can have thousands of element, we need a concise way of summarising the samples from the posterior distribution of $\boldsymbol{b}|D$. As we are interested in the branches with an adaptive event on them, for each branch i we marginalise over all other branches by counting the number of times branch i is tagged as having an adaptive event on it in the sampled $\boldsymbol{b}$s. The proportion of times that each branch is tagged approximates the marginal posterior probability of the branch having an adaptive event on it. This consensus representation of $\boldsymbol{b}$ can be used to draw the tree such that line thickness or colour shade or both is proportional to the marginal posterior probability of having an adaptive event on the branch. Figure 4.2 shows this for a toy simulated dataset. The binary environmental categories are shown in red and blue on the tips of the tree and the line thickness is proportional to the marginal posterior probability of having an adaptive event on a branch. In this example there are three branches with posterior probability of having an adaptive event which is around 0.9 and a few branches with posterior probability of having an adaptive event which is around 0.1 and for the rest of the branches the posterior probability of having an adaptive event is near zero.

To have a point estimate of $\boldsymbol{b}$ we use a cutoff threshold on the consensus representation of $\boldsymbol{b}$. The value of the threshold will determine the rate of type I and II errors. As the threshold rises, the rate of false positives decreases, while the rate of false negatives increases. Depending on application one can choose a suitable threshold. For the toy example of Figure 4.2 to get a

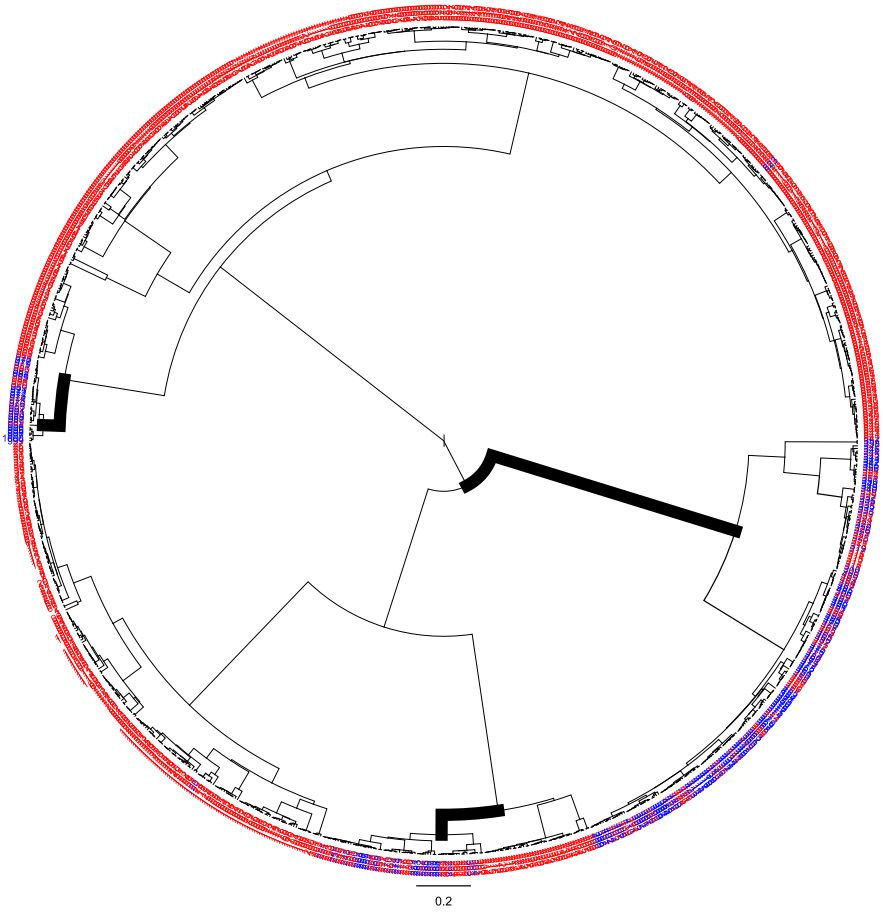


Figure 4.2: Consensus representation of $\boldsymbol{b}$. Marginal posterior probability of a branches having adaptive events is the consensus representation of $\boldsymbol{b}$. The thickness of a branch is proportional to the posterior probability of having an adaptive event on that branch. In this instance there are three branches which their posterior probability of having an adaptive event is around 0.9. A few branches with posterior probability of having an adaptive event of around 0.1 and all other branches with posterior probability of having an adaptive event which is almost zero.

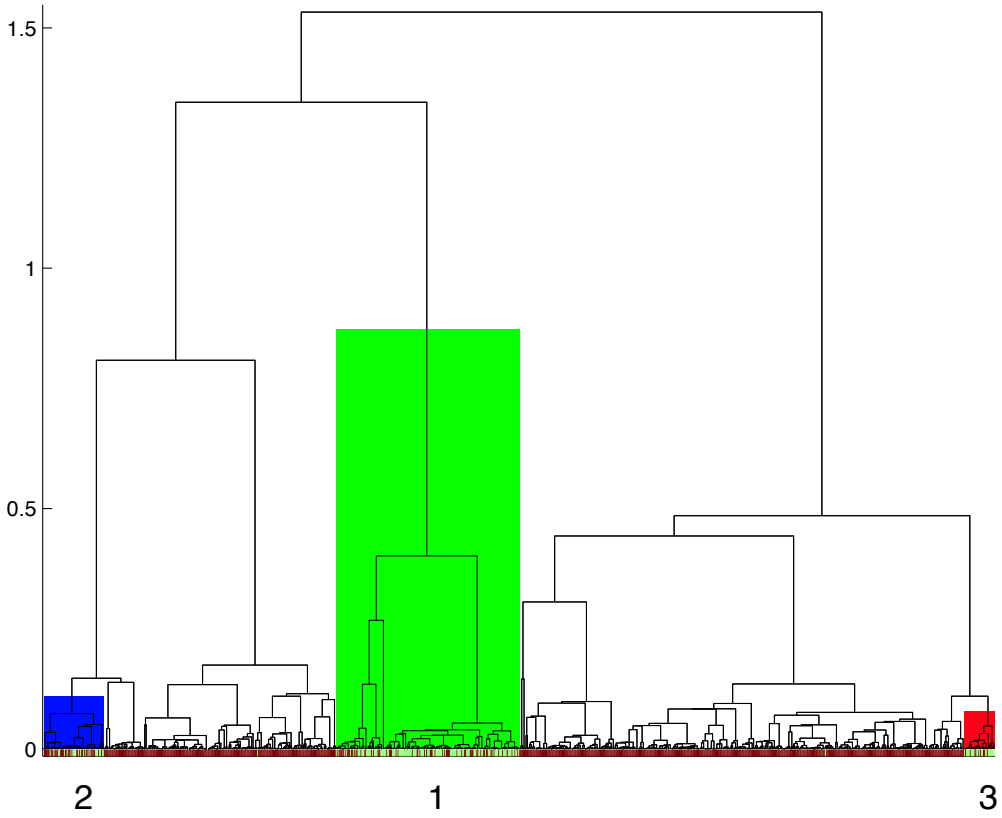
point estimate of $\boldsymbol{b}$ we chose a cutoff threshold of 0.5. Figure 4.3a shows how the point estimate of $\boldsymbol{b}$ has divided the tree into four ecologically distinct lineages highlighted by different colours. Our model implicitly assumes an adaptive event at the root of the tree. All isolates that are not part of the sections delineated by the inferred adaptive events are part of the root section. In Figure 4.3a, all isolates with white background are part of the root section.

Distribution of environmental categories

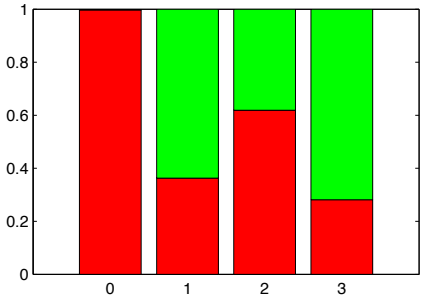
Given the point estimate of $\boldsymbol{b}$, we can estimate the most likely distribution over the environmental categories for each ecologically distinct lineage. Equation 4.8 gives the likelihood of the observed environmental categories of the isolates subject to constraints given in Equation 4.9. Assuming that the point estimate of $\boldsymbol{b}$ divides the tree into m sections, each section has a multinomial distribution on the environmental categories. Given the point estimate of $\boldsymbol{b}$ the values of q_{ij} that maximise the data likelihood are given by:

$$\hat{q}_{ij} = \frac{x_{ij}}{\sum_{j=1}^{K} x_{ij}} \tag{4.23}$$

Figure 4.3b shows the estimated distribution over the environmental categories for each ecologically distinct lineage for the toy example of Figure 4.3a. The root section is named section zero and all isolates in section zero have the same adaptive event as the one present at the root of the tree. In this example isolates in section zero, are almost exclusively found in the red environmental category, while the strains in the other sections are found in both red and green categories.



(a) Point estimate of $\boldsymbol{b}$ divides the tree into four ecologically distinct lineages.



(b) Estimated distribution of strains over the environmental categories for each distinct lineage.

Figure 4.3: Representation of point estimate of $\boldsymbol{b}$ for a toy example. (a) The point estimate of $\boldsymbol{b}$ has divided the tree into four ecologically distinct lineages. The binary environmental category are shown in red and green. (b) The distribution over the environmental categories in each distinct lineage. Strains in section zero are almost exclusively found in the red environmental category while the strains in the other sections are found in both categories.

4.3 Simulation studies

To investigate the performance of our method, we performed three simulation studies. In all of these simulations for simplicity we assumed the environmental categories were binary and to sample from the posterior distribution of the model parameters our MCMC chain was run for 10^7 iterations. All of these simulations were implemented for a single genealogy simulated using coalescent with 1000 strains shown in Figure 4.4 unless otherwise stated. In the first simulation study we tested how the number of isolates in the section and the change in distribution over the environmental categories due to the adaptive event affects the power to infer an adaptive event. In the second simulation study we tested the model selection procedure and the relationship between the posterior expectation of number of adaptive events against the true numbers of adaptive events. We also quantified the effect of threshold on the point estimate of $\boldsymbol{b}$. In the final simulation study, we tested the relationship between the posterior expectation of the rate of adaptation against the true rate of adaptation.

4.3.1 Simulation study A

This simulation study was designed to assess the power of the method to detect adaptive events on the branches of the tree. The power depends on two factors p and n. The first factor p is the amount of change in the distribution over the environmental categories due to the adaptive event from the parent section to the child section. For binary environmental categories, we define p such that an adaptive event that changes the distribution from (0.5, 0.5) to (0.25, 0.75) has a $p = 0.25$. Adaptive events that lead to small changes in the distribution are difficult to infer as they

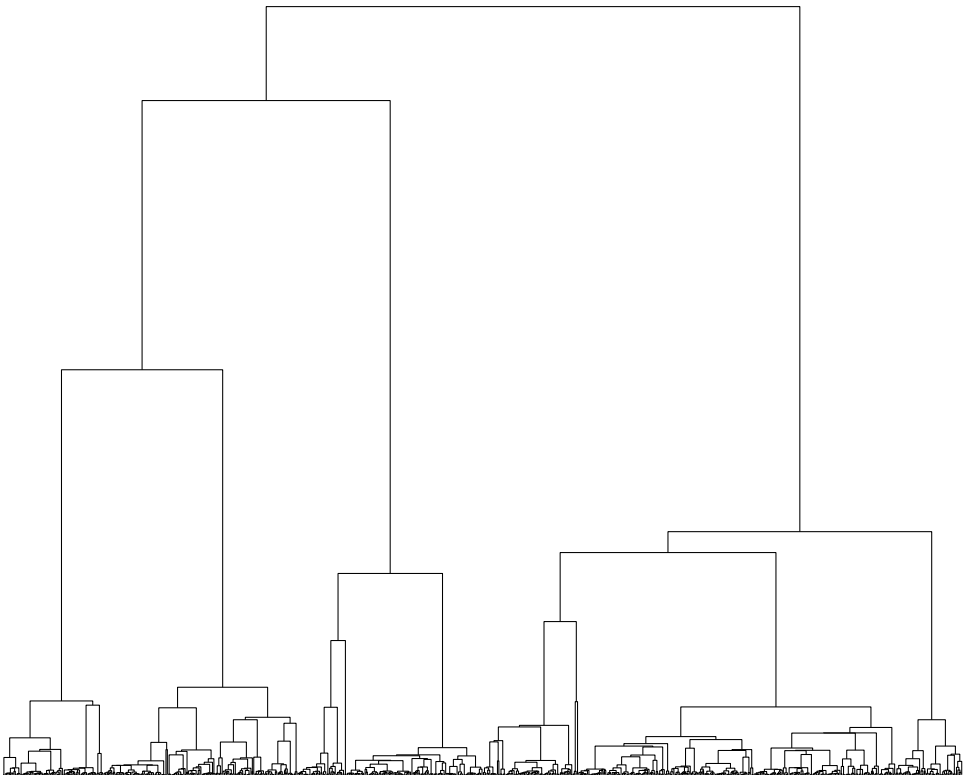


Figure 4.4: Tree used for the simulation studies. This tree was simulated using coalescent model with 1000 strains.

result in small changes to the observed pattern of distribution over the environmental categories that are likely to happen by chance alone. The second factor n is the number of isolates in the section which is demarcated by the adaptive event. Adaptive events delineating sections that contain only a few isolates are difficult to distinguish as lack of data makes the inference more uncertain. We expect that adaptive events that produce large changes in distribution or demarcate sections with large number of isolates or both to result in higher posterior probabilities.

First we investigated the effect of p the amount of change in the distribution over the environmental categories. We simulated 110 datasets each with a single adaptive event which divided the tree of Figure 4.4 into

two sections. The adaptive event demarcated a section which contained 500 isolates. We assumed the distribution over the categories for the root section to be (0.5, 0.5). We assumed the change in distribution over the categories due to the adaptive event to be $p = 0.0, 0.05, \ldots, 0.5$ and for each case simulated 10 datasets, 110 in total. For each p the result of the 10 datasets were summarised by calculating the mean of marginal posterior probability of having an adaptive event for the branch with the adaptive event on it. The result is shown in Figure 4.5. Bigger changes in the distribution over the categories, leads to larger posterior probability of having an adaptive event. There is a sharp transition threshold between $p = 0.05$ and $p = 0.15$ for this simulation study.

Next we examined the relationship between p and n and how they affect detection of adaptive events. We divided the space of $n \times p$ into a grid where $n = (10, 50, 100, 250)$ and $p = (0.1, 0.2, 0.3, 0.4, 0.5)$. For each node of the grid (p_i, n_j) a branch on the tree of Figure 4.4 was chosen to have an adaptive event on it. For each node of the grid we simulated 20 datasets with a single adaptive event on the chosen branch of the tree. We assumed that the root section distribution over the categories was $(0.5, 0.5)$. For each node of the grid the results of the 20 datasets were summarised by calculating the mean marginal posterior probability of having a adaptive event on the branch with the adaptive event on it. Figure 4.6 shows the result. An adaptive event that causes large changes to the distribution over the categories and delineates a section with large number of isolates results in larger posterior probability of an adaptive event. Adaptive events that cause small changes in the distribution or demarcate sections that contain few isolates or both result in small posterior probability of having an adaptive event.

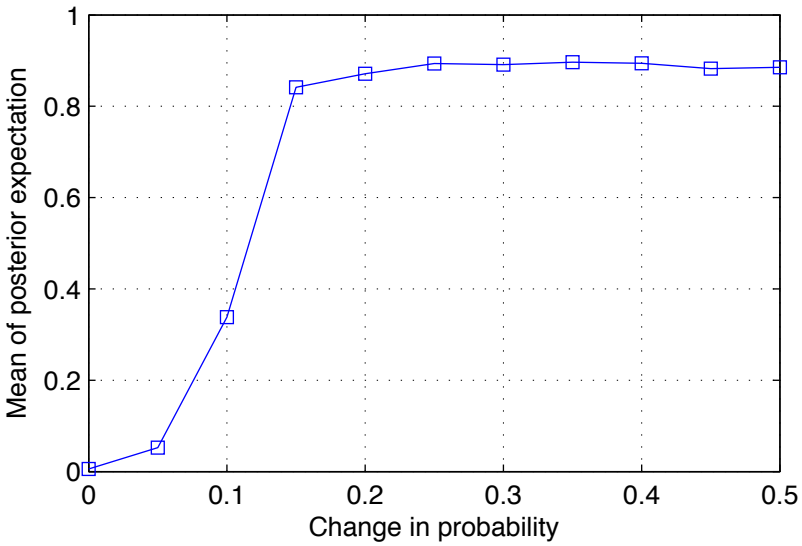


Figure 4.5: Mean of marginal posterior probability of having an adaptive event against the change in the distribution over the categories. 110 datasets were simulated each with a single adaptive event which divided the tree of Figure 4.4 into two sections. The adaptive event demarcated a section that contained 500 isolates. The root section distribution over the categories was $(0.5, 0.5)$. We set the change in the distribution due to adaptive event to be $p = 0.0, 0.05, \ldots, 0.5$ and for each case 10 datasets were simulated. The results were summarised for each case by estimating the mean posterior probability of having an adaptive event for the branch with the adaptive event on it for the 10 simulated datasets. Bigger changes in the distribution result in larger posterior probability of having an adaptive event.

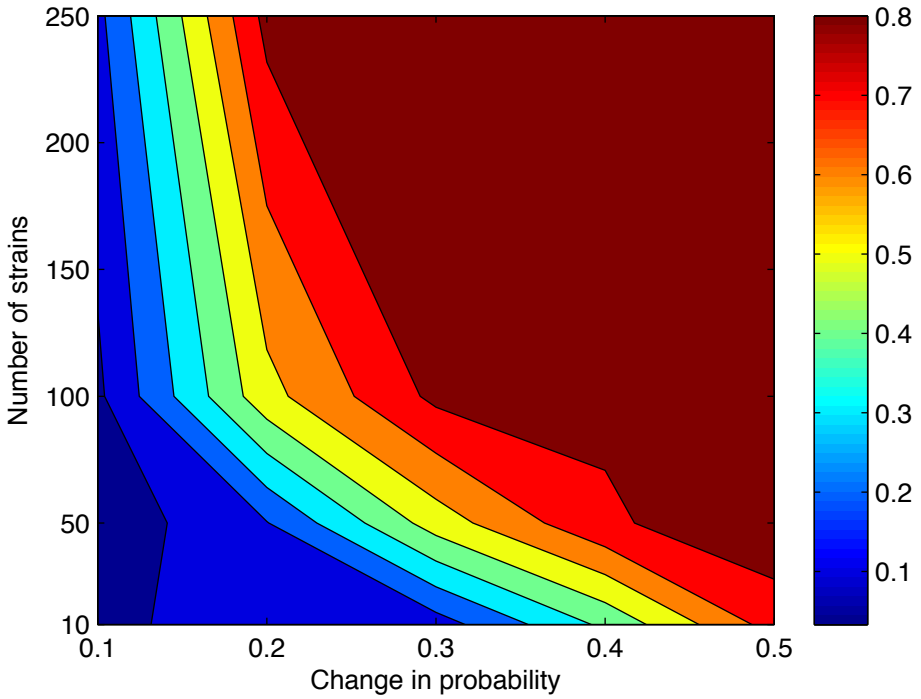
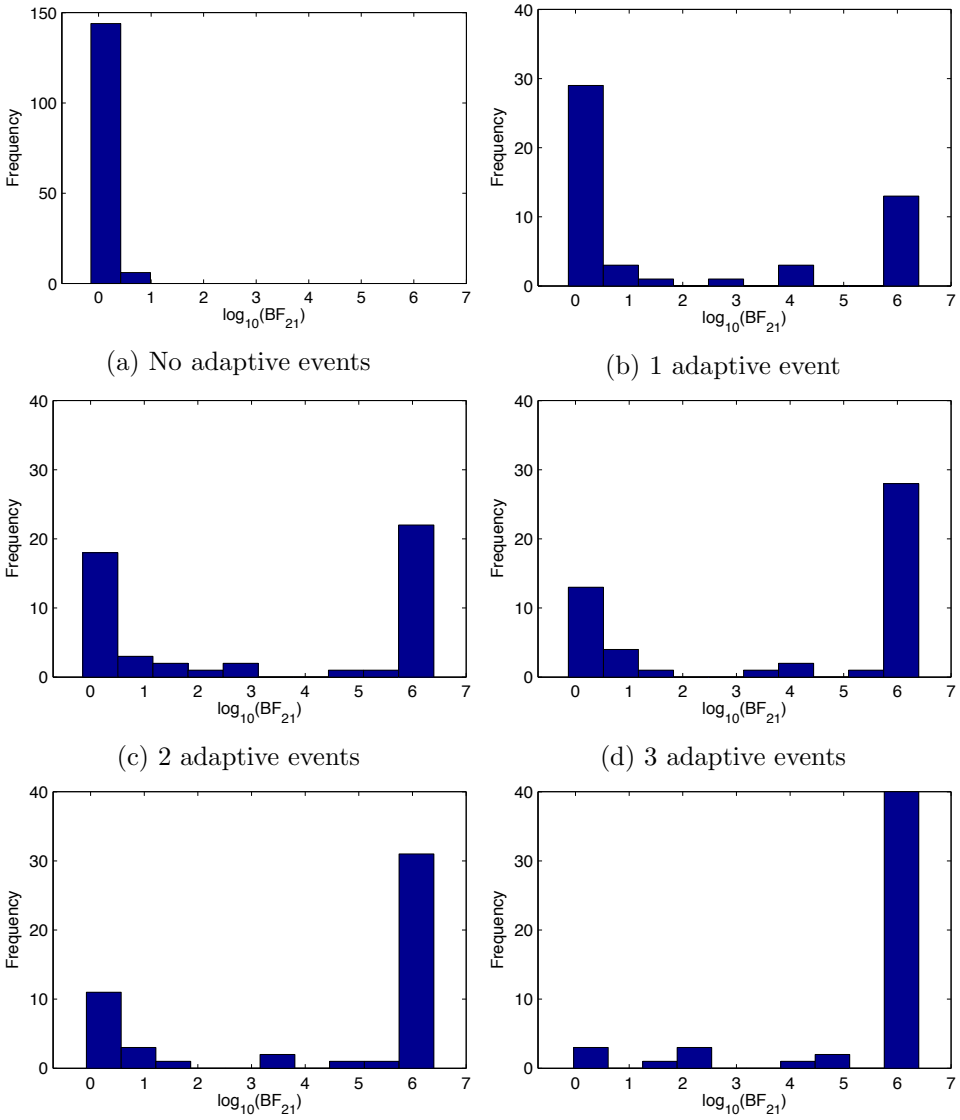


Figure 4.6: Contour plot of the mean posterior probability of having an adaptive event against number of isolates and the change in distribution. We divided the space of $n \times p$ into a grid where $n = (10, 50, 100, 250)$ and $p = (0.1, 0.2, 0.3, 0.4, 0.5)$. For each node of the grid (p_i, n_j) we simulated 20 datasets each with a single adaptive event using the tree of Figure 4.4. For each node of the grid to summarise the results, from the 20 datasets the mean posterior probability of having an adaptive event was calculated for for the branch with the adaptive event on it. An adaptive event that results in large changes in distribution or delineates a section that contains large number of isolates or both leads to higher posterior probabilities of having an adaptive event.

4.3.2 Simulation study B

We designed this simulation study to assess our model selection procedure, the effect of number of adaptive events on the inference and the effect of cutoff threshold on the point estimate of $\boldsymbol{b}$. First, we simulated 150 datasets with no adaptive event on the tree. In the absence of any adaptive events, all isolates follow a single distribution over the categories. As our model selection could potentially depend on the distribution over the environmental categories we used 3 different distributions for the environmental categories. The following three distributions $(0.1, 0.9), (0.25, 0.75), (0.5, 0.5)$ were used and 50 datasets were simulated for each of the distributions, 150 in total. We also simulated 50 datasets for each case of one, two, three, four and five adaptive events on the tree. For each simulated dataset the Bayes factor of model 2 against model 1, BF_{21} was estimated. Figure 4.7 shows the distribution of the estimated Bayes factors for the simulated datasets. For the 150 datasets with no adaptive event on the tree, all the estimated Bayes factors were below 10 i.e. there was no significant evidence against model 1 (single ecological lineage) for any of datasets. Adaptive events that result in small changes in the distribution or demarcate sections with small number of isolates will not be detected. Therefore for many datasets with a single adaptive event there is no significant evidence against model 1. As the number of adaptive events on the tree rises, the number of datasets with significant evidence for model 2 increases. Hence, our method is conservative and will not result in significant evidence for model 2 unless there is substantial data to support it.

Next, we used the simulations to gauge the effect of the true number of adaptive events on the posterior expectation of number of adaptive events.



(a) No adaptive events

(b) 1 adaptive event

(c) 2 adaptive events

(d) 3 adaptive events

(e) 4 adaptive events

(f) 5 adaptive events

Figure 4.7: $\log_{10}(\mathrm{BF}_{21})$. For the 150 datasets with no adaptive event on the tree, all the estimated Bayes factors are below 10 i.e. there is no significant evidence against model 1 (no adaptive event) for any of datasets. Adaptive events that result in small changes in the distribution or demarcate sections with small number of isolates will not be detected. Therefore for many datasets with a single adaptive event there is no significant evidence against model 1. As the number of adaptive events on the tree rises, the number of datasets with significant evidence for model 2 increases.

To estimate the posterior expectation of number of adaptive events, we used Bayesian model averaging (HOETING *et al.*, 1999). Figure 4.8 illustrates the results. In absence of adaptive events, the mean of posterior expectation of number of adaptive events is close to zero and almost for all datasets the posterior expectation of number of adaptive events is below one. When there are adaptive events on the tree, mean of posterior expectation of number of adaptive events is lower than the true numbers. This is expected as our method cannot detect an adaptive event that results in small changes in the distribution or demarcates a section which contains few isolates or both. As a result our method is conservative in estimating the number of adaptive events on the tree.

In addition we used the simulation results to assess the effect of the cutoff threshold on the point estimate of $\boldsymbol{b}$. For each of the datasets we inferred a point estimate for $\boldsymbol{b}$ by applying a threshold to the consensus representation of $\boldsymbol{b}$ (marginal posterior probability of having an adaptive event for all branches of the tree). The threshold was changed from 0.05 to 0.95 with increments of 0.05. For each threshold value, the number of branches which were correctly inferred to have adaptive events on them and the number of branches that were incorrectly inferred to have adaptive events on them was counted. We then normalised these values by the total number of true adaptive events for all datasets. Figure 4.9 shows the ratio of number of correctly and incorrectly inferred adaptive events to the total number of adaptive events as a function of the threshold value. As expected with increasing the threshold value, the rate of incorrectly inferred adaptive events decreases, but at the same time the rate of correctly inferred adaptive events decreases too. The choice of the cutoff threshold is a trade off between minimising the number of incorrectly

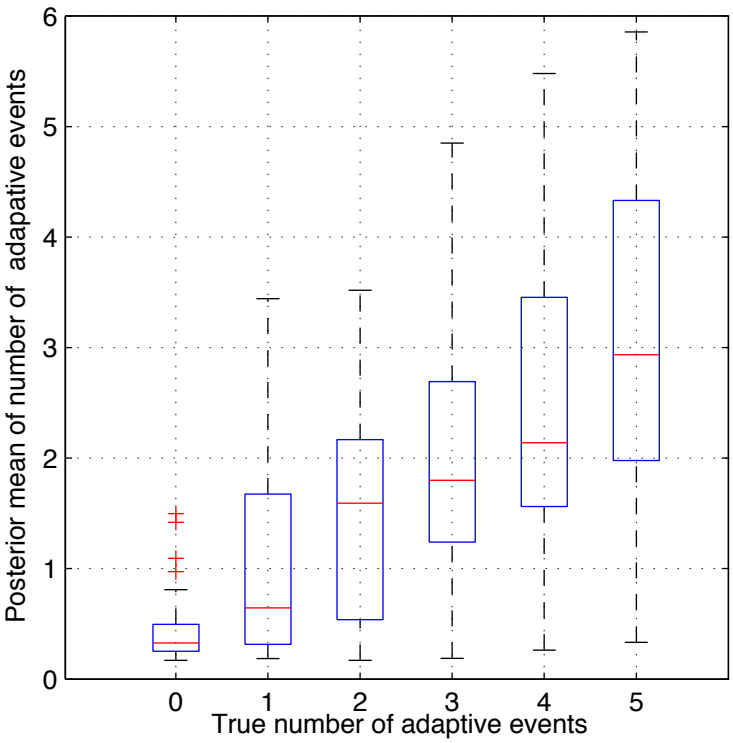


Figure 4.8: Distribution of posterior expectation of number of adaptive events against the true number of adaptive events. Apart for the case of no adaptive events, on average the method underestimates the true number of adaptive events. This is because adaptive events that lead to small changes in distribution and or are carried by few strains cannot be detected.

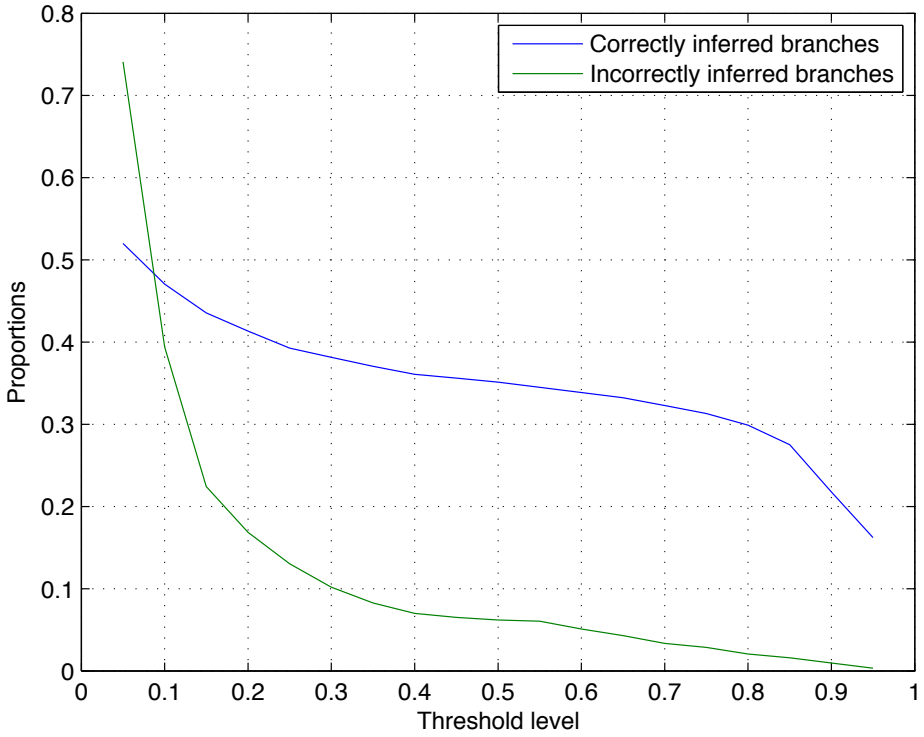


Figure 4.9: Ratio of the number of correctly and incorrectly inferred adaptive events to the total number of adaptive events against the threshold value. By increasing the threshold value, the rate of incorrectly inferred events decreases, but the rate of correctly inferred adaptive events decreases too. The choice of threshold value is a trade off between reducing the number of incorrectly inferred adaptive events against increasing the number of correctly inferred adaptive events.

inferred adaptive events and maximising the number of correctly inferred adaptive events. Depending on application we would recommend a threshold value between 0.4 and 0.6.

4.3.3 Simulation study C

In this simulation study we investigated the performance of the method on inferring the rate of adaptation λ on the branches of the tree. In our model the number of adaptive events on the tree is Poisson distributed with rate

λ. A priori we assumed that $\mathbb{E}(\lambda) = 1/T$ where T is the total branch length of the tree. We simulated 100 datasets using tree of Figure 4.4 for which $1/T = 0.067$. For each dataset λ was sampled from the uniform distribution on $[0, 0.25]$. We used 0.25 as upper limit as it is almost four times the prior expectation of λ. To estimate the posterior expectation of λ, we used Bayesian model averaging (Hoeting *et al.*, 1999). Results are shown in Figure 4.10. The observed pattern is similar to that of Figure 4.8. This would be expected because in our model the inferred number of adaptive events influences the inferred rate of adaptation. The method is conservative in estimating λ as some adaptive events are not detected. On average this results in smaller estimates of λ than the true value. Another point to notice is that when true value of λ is close to zero, the method overestimates the value of λ. This is because in the case of no adaptive event on the tree, λ can be non-zero. In this situation model 2 with a small λ and model 1 where $\lambda = 0$ are equally likely. Averaging between the two model results in an overestimate of λ.

4.4 Application to empirical datasets

4.4.1 *Salmonella enterica* subspecies *enterica* dataset

Salmonella genus is divided into two species, *Salmonella bongori* and *Salmonella enterica*. *S. enterica* is further divided into six subspecies. From these, *S. enterica* subspecies *enterica* (referred to as *enterica* subsequently) causes 99% of human and animal infections (Achtman *et al.*, 2012). It is the major cause of typhoid and paratyphoid fever in humans and gastroenteritis in humans and animals. *Enterica* is the most diverse subspecies of *S. enterica* as it contains about 60% of the

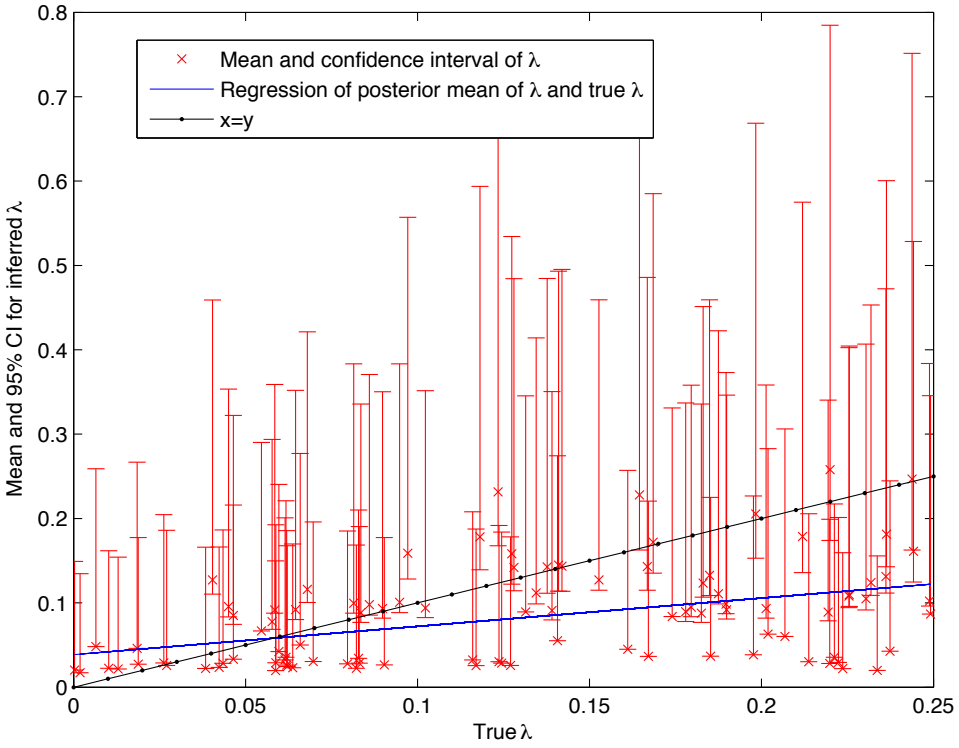


Figure 4.10: Posterior expectation of λ and its CI against true λ. We simulated 100 datasets where λ was sampled from a Uniform distribution on $[0, 0.25]$. Model averaging was used to calculate the posterior expectation of λ and its confidence interval which are shown in red. The black line shows the $y = x$ line and the blue line shows the regression line of posterior expectation of λ against true λ. The observed pattern is similar to that of Figure 4.8. This would be expected because in our method the inferred number of adaptive events influences the rate of adaptation λ. As the blue line shows, the method is conservative in estimating posterior expectation of λ as some adaptive events are not detected. On average this results in smaller estimates of λ than the true value. Another point to notice is that when true value of λ is close to zero, the method overestimates the posterior expection of λ. This is because in the case of no adaptive event on the tree, λ can be non-zero. In this situation model 2 with a small λ and model 1 where $\lambda = 0$ are equally likely. Averaging between the two model results in an overestimate for λ.

approximately 2500 serovars defined in the *Salmonella* genus (Achtman *et al.*, 2012). These serovars differ widely in their host range and diseases they cause. For instance, serovars Typhi and Paratyphi only infect humans and cause enteric fever. Serovars Dublin and Choleraesuis cause disease in cattle and pigs respectively, but can be carried asymptomatically by humans and other animals. Other serovars such as Typhimurium and Enteritidis can infect humans and a wide range of animals and cause gastroenteritis (Uzzau *et al.*, 2000).

Adaptive events in the history of *enterica* have resulted in these serovar host association patterns. Our method is suitable for detecting these adaptive events and demarcating phylogenetic clusters of isolates that have differential distribution over human and animal hosts. We applied our method to the *Salmonella enterica* pubMLST database to detect which lineages are differentially distributed over human and animal hosts. In September 2010, the database included 4137 isolates. After cleaning the data, there were 2169 isolates of *enterica* where the host was either human or animal. From these isolates, 1455 were from human hosts and 714 were from animal hosts. These isolates were assigned to 636 ST types, 393 serovars and 123 eBurstGroups (eBG) (some isolates' serovars and eBG numbers were not known). The sequences for the seven house keeping genes used in the MLST typing scheme were concatenated to make a sequences of length 3336 bp. A UPGMA tree was built from these sequences (Fitch and Margoliash, 1967). The resulting tree and the binary host type of human or animal was used as input to our method.

To sample from the posterior distribution of model parameters an MCMC with 10^8 iterations was performed. The Bayes factor of model 2 against model 1 was of order 10^8 i.e. decisive evidence that there are

lineages that have distinct distribution over the host categories. Figure 4.11 shows the UPGMA tree for the isolates with the human and animal hosts indicated at the tips of the tree by red and blue colours respectively. In addition, branches are coloured proportional to their marginal posterior probability of having an adaptive event. Darker shades are equivalent to higher posterior probabilities. Furthermore branches with posterior probability of having an adaptive event bigger than 0.6 are highlighted by green and purple colours. This point estimate has 13 branches with adaptive events on them and divides the tree into 14 sections. The sections demarcated by the adaptive events are numbered from 1 to 13. All other isolates will be in the root section which is numbered as zero. Figure 4.12 shows the distribution of each section over the host categories. Clusters 1, 9, and 11 are only associated with humans hosts while clusters 2, 3, 4 and 8 are only associated with animal hosts. Clusters 5, 6 ,10 and 12 are mostly associated with human hosts but they are present in animal hosts too while cluster 7 is mostly associated with animal hosts but is present in human hosts too. Cluster 13 is associated more with animal than with human hosts. Section zero (which is made up of all isolates not in any of the highlighted sections) is associated more with human hosts than animal hosts. The information about these clusters are summarised in table 4.1. The table shows for each cluster the number of human and animal hosts in that cluster as well as their ST types, serovars and eBurstGroups (eBG).

Cluster 1, 9 and 11 that are only associated with human hosts are made up of Typhi, Isangi and mostly Paratyphi C serovars respectively. Section 10 is almost exclusively associated with humans and is made up of Paratyphi A, Sendi, Derby and other closely related serovars. Sections 2, 3 and 4 and 9 are exclusively associated with animals and are made up of

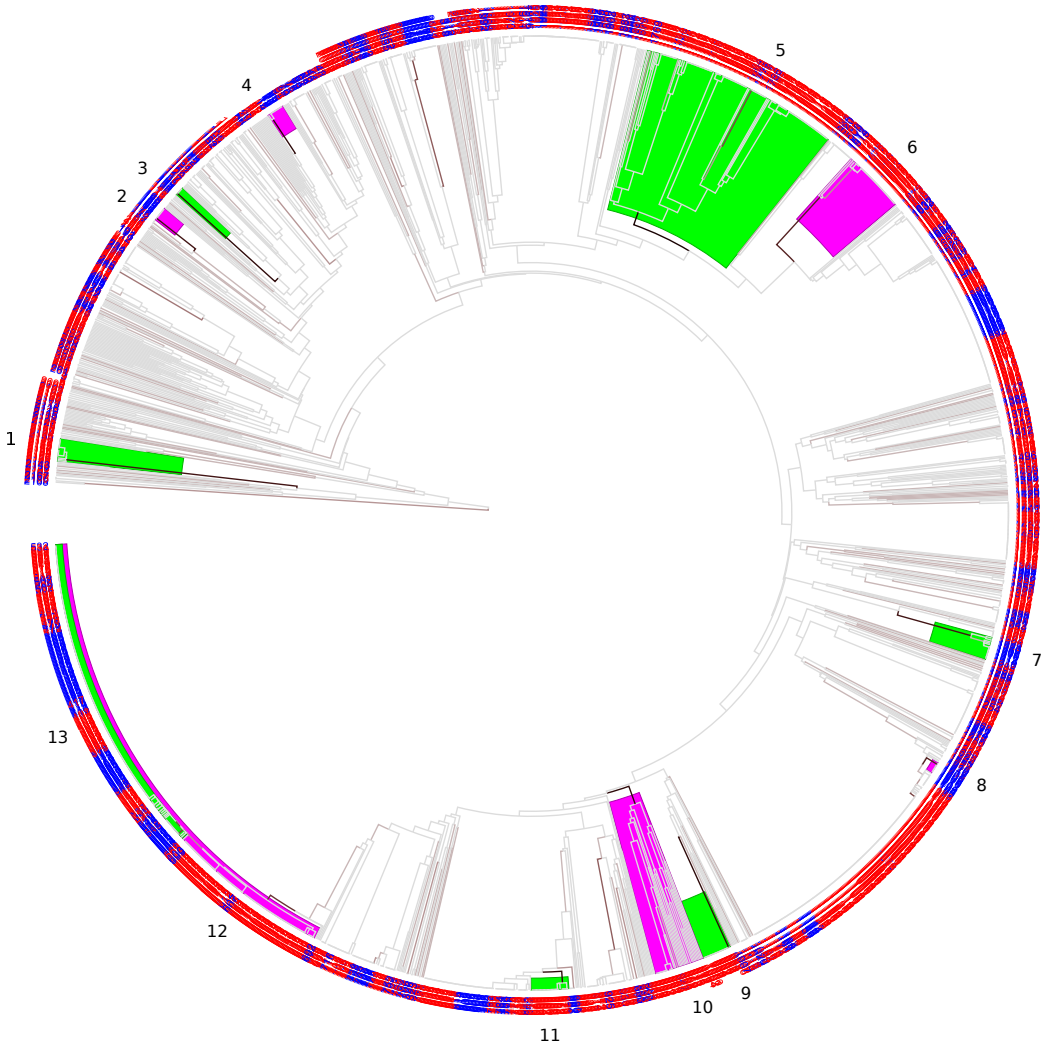


Figure 4.11: UPGMA tree of *enterica* dataset. Isolates from human and animal hosts are coloured in red and blue respectively. Marginal posterior probability of a branch having an adaptive event is proportional to the branch colour shade. Darker branches have a higher posterior probability. There were 13 branches with posterior probability which is bigger than 0.6 and they have been highlighted by green and purple colours. These lineages are differentially distributed over the host categories.

Bahrenfeld and Oranienburg serovars. An interesting feature of the data is about the Typhimurium isolates in clusters 12 and 13. These clusters are almost exclusively made up of Typhimurium isolates. Section 13 is a subcluster of section 12 and the two sections according to the UPGMA tree are closely related. However the two sections are distinctly distributed over the host categories. Isolates of section 12 are mostly associated with human hosts while isolates of section 13 are more associated with animal hosts. It seems that an adaptive event made Typhimurium isolates adapted to human hosts and subsequently another adaptive event made a subcluster of Typhimurium isolates to be more associated with animal hosts. Typhimurium is usually referred to as the typical broad host range serovar, although there are some evidence that different variants are distinctly associated with hosts (Rabsch *et al.*, 2002).

All three form of serovar host association in *enterica* described previously (Uzzau *et al.*, 2000) have been detected by our method. There are clusters that are associated only with human or animal hosts. There are clusters that are associated mostly with human or animal hosts, but can be found in the other host too. There are also clusters of isolates that are associated with human and animal host almost in equal proportions.

4.4.2 *Vibrionaceae* dataset

Hunt *et al.* (2008) investigated differential distribution of coastal ocean bacteria of the family *Vibrionaceae* with temporal and spatial factors using AdaptML software. They argue that coastal ocean is a suitable place to test for ecological preference of isolates, as waves and currents will ensure there is no geographic barrier to movement and migration of isolates.

The family *Vibrionaceae* is one of the most studied marine bacteria.

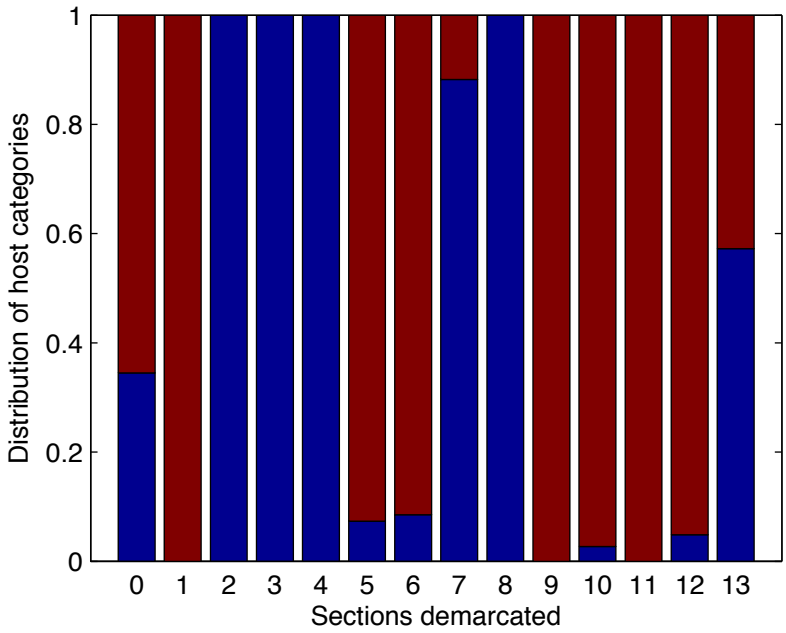


Figure 4.12: Distribution over the host categories. Thirteen adaptive events were inferred for *Enterica* dataset that divided the tree into 14 sections. Distribution of human and animal hosts in each section are shown in red and blue colours respectively. Sections one to thirteen are as marked on the tree of Figure 4.11. The root section numbered as zero is made up of all isolates that are not in section one to thirteen.

Section	# of isolates	Human	Animal	ST (number)	Serovar (number)	eBG (number)
1	17	17	0	2(8) 1(6) 3(1) 892(1) 890(1)	Typhi(17)	13(17)
2	11	0	11	859(11)	Bahrenfeld(11)	131(11)
3	7	0	7	179(7)	Oranienburg(7)	52(7)
4	13	0	13	174(12) 169(1)	Oranienburg(13)	44(13)
5	150	139	11	533(37) 16(22) 166(21) 15(14) 26(14) ⋮	Concord(36) Virchow(27) Newport(24) Unknown(18) Thompson(14) Heidelberg(13) Stanleyville(2) ⋮	86(52) 9(29) 35(24) 26(15) Unknown(9) 28(4) 79(2) ⋮
6	47	43	4	31(38) 132(5) 191(2) 193(1) 200(1)	Newport(47)	7(46) 3(1)
7	17	2	15	112(11) 82(2) 170(1) 173(1) 176(1) 177(1)	Muenchen(16) Valdosta(1)	8(17)
8	9	0	9	92(9)	Unknown(9)	4(9)
9	22	22	0	216(22)	Isangi(22)	25(22)
10	37	36	1	85(10) 129(4) 71(2) 365(2) ⋮	Paratyphi A(10) Sendai(4) Derby(3) Weltevreden(2) ⋮	Unknown(21) 11(14) 205(2)
11	28	28	0	146(21) 90(6) 114(1)	Paratyphi C(27) Limete(1)	20(28)
12	123	117	6	313(55) 213(33) 34(26) 302(6) 35(2) 394(1)	Typhimurium(105) Typhimurium Mono(11) Unkown(7)	1(116) Unkown(7)
13	241	138	103	19(204) 128(16) 568(5) 376(3) 98(2) ⋮	Typhimurium(227) Unknown(8) Typhimurium Mono(5) Hato(1)	1(235) Unknown(6)

Table 4.1: Analysis of *enterica* dataset. Sections are numbered according to Figure 4.11. Columns 3 and 4 indicate the number of human and animal hosts in each section. In columns 5, 6 and 7 the values in parentheses gives the number of isolates in that type.

They are important players in surface water as they cycle organic and inorganic compounds. As bacterioplanktons, they could either use dissolved nutrients which are of low concentration but more evenly distributed or attach to the larger particles that are suspended in the water and degrade them (Nishiguchi and Jones, 2005).

To investigate differentiation in water columns along temporal (season) and spatial (free living or particle associated) lines, Hunt *et al.* (2008) sampled coastal ocean water on two days representing spring and autumn near Plum Island Estuary (NE Massachusetts). The water samples were sequentially filtered with decreasing filter sizes and the resulting media was cultivated on *Vibrio* selective media. The largest size fraction (media and particle > 63 μm) is enriched in zooplankton and detrital matter. The next fraction size (63-5 μm) likely contains zooplankton faecal pellets and algae. The size fraction 5-1 μm is a buffer between free living bacteria and particle associated bacteria as it can contain both large free living bacteria or bacteria associated with smaller particles. The smallest size fraction ($<$ 1 μm) contains free living bacteria which likely live on dissolved organic matter. This procedure resulted in 1023 *Vibrionaceae* cultures for which part of the *hsp60* gene was sequenced. A maximum likelihood tree was constructed from the partial gene sequence and was rooted using an outgroup (Hunt *et al.*, 2008).

We applied our method to the dataset published by Hunt *et al.* (2008). The four size fraction and the two seasons resulted in eight environmental categories. To sample from the posterior distribution of model parameters an MCMC with 10^8 iterations was performed. The estimated Bayes factor in favour of model 2 was of order 10^8 i.e. decisive evidence that some of the lineages are differentially distributed over the environmental categories.

Figure 4.13 shows the distribution of size fractions and seasons of the isolates on the tree. The tree branches were drawn such that the shade and the line thickness on the tree represents the posterior probability of having an adaptive event on that branch. Thicker lines and darker shades indicate larger posterior probabilities.

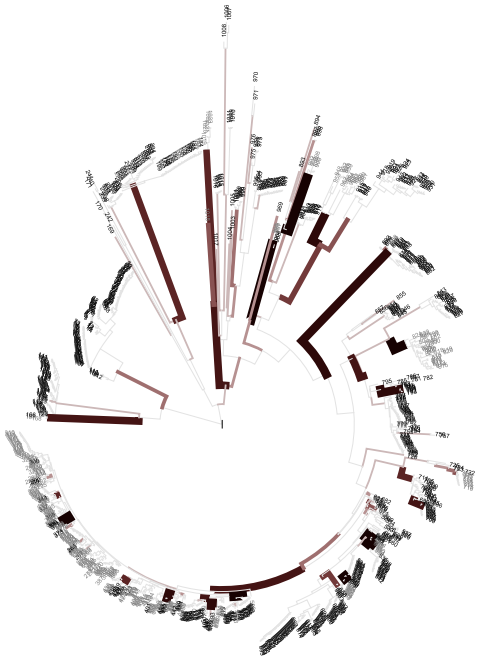
Figure 4.14 shows the same tree where branches with posterior probability of having an adaptive event bigger than 0.6 are highlighted. This point estimate has 21 branches with adaptive events on them and divides the tree into 22 sections. These sections are numbered from 1 to 21 on the tree and are highlighted. The root section numbered as zero contains all isolates that are not in one of the highlighted sections. This Figure also indicates the branches inferred by AdaptML by red dots. AdaptML has inferred 26 branches with adaptive events on them. Table 4.2 shows a comparison of the result of AdaptML and our method. There are eight branches that are common to both analysis. AdaptML has inferred another two branches which have posterior expectation of close to 0.5 under our model. Furthermore, AdaptML has inferred another 12 branches which under our model have a posterior expectation of less than 0.1.

Figure 4.15 shows the distribution of size fraction and season for each section marked on the tree of Figure 4.14. Most sections are strongly associated with one or the other season. Most sections are also highly associated with being free living or particle associated. We assume size fraction $< 1\mu m$ is free living bacteria and size fraction $> 5\mu m$ is particle associated. Our results confirms Hunt *et al.* (2008) conclusion that the single bacterial family of *Vibrionaceae* coexisting in coastal ocean's water divides into many ecologically distinct populations. Some of the branches

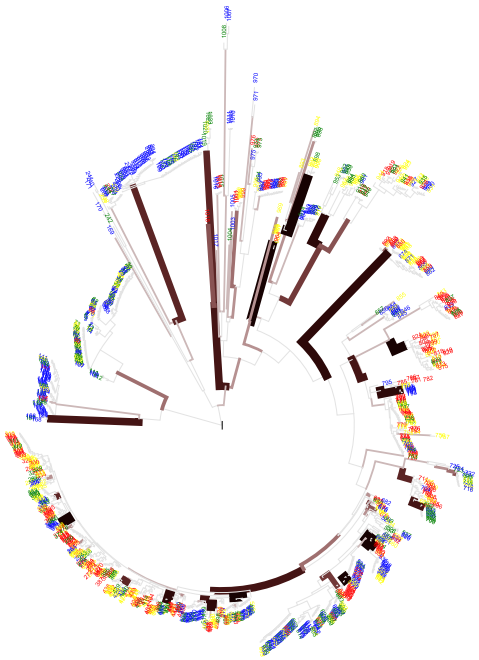
inferred by the two methods are the same however there are significant differences between our results as there are clusters which have been detected in one but not in the other analysis. These differences would be expected as our models are very different as well as the inference procedures.

AdaptML Inferred Branch	Our Posterior Expectation
1920	0.889
1586	0.887
1811	0.810
1911	0.810
1980	0.760
1534	0.738
1862	0.682
1246	0.606
1981	0.464
1739	0.439
1872	0.086
1185	0.065
2011	0.060
1860	0.058
1861	0.058
1370	0.019
1504	0.006
1190	0.004
1686	0.000
1305	0.000
1352	0.000
1378	0.000
1412	0.000
1435	0.000
1931	0.000
1707	0.000

Table 4.2: Comparison of the AdaptML results to our model. The first column shows branches that create ecologically distinct populations with an empirical p-value of less than 0.0001 which is inferred by AdaptML. The second column shows the posterior expectation of those branches having an adaptive event by our method. As it can be seen 10 of the branches inferred by AdaptML have posterior expectation of near or above 0.5 by our method. On the other hand there are 16 branches inferred by AdaptML that have very small posterior expectation under our model.



(a) Seasonal distribution of the *Vibrionaceae* isolates on the tree. Black and grey colours indicate autumn and spring respectively.



(b) Size distribution of the *Vibrionaceae* isolates on the tree. Red, yellow, green and blue indicate $> 63\ \mu m$, 63–5 μm, 5–1 μm and $< 1\ \mu m$ size fractions respectively.

Figure 4.13: Marginal posterior probability of having an adaptive event indicated by colour shade and line thickness. Thicker lines and darker shades indicate higher posterior probabilities. (a) Seasonal distribution of the isolates on the tree. (b) Distribution of size fraction of the isolates on the tree.

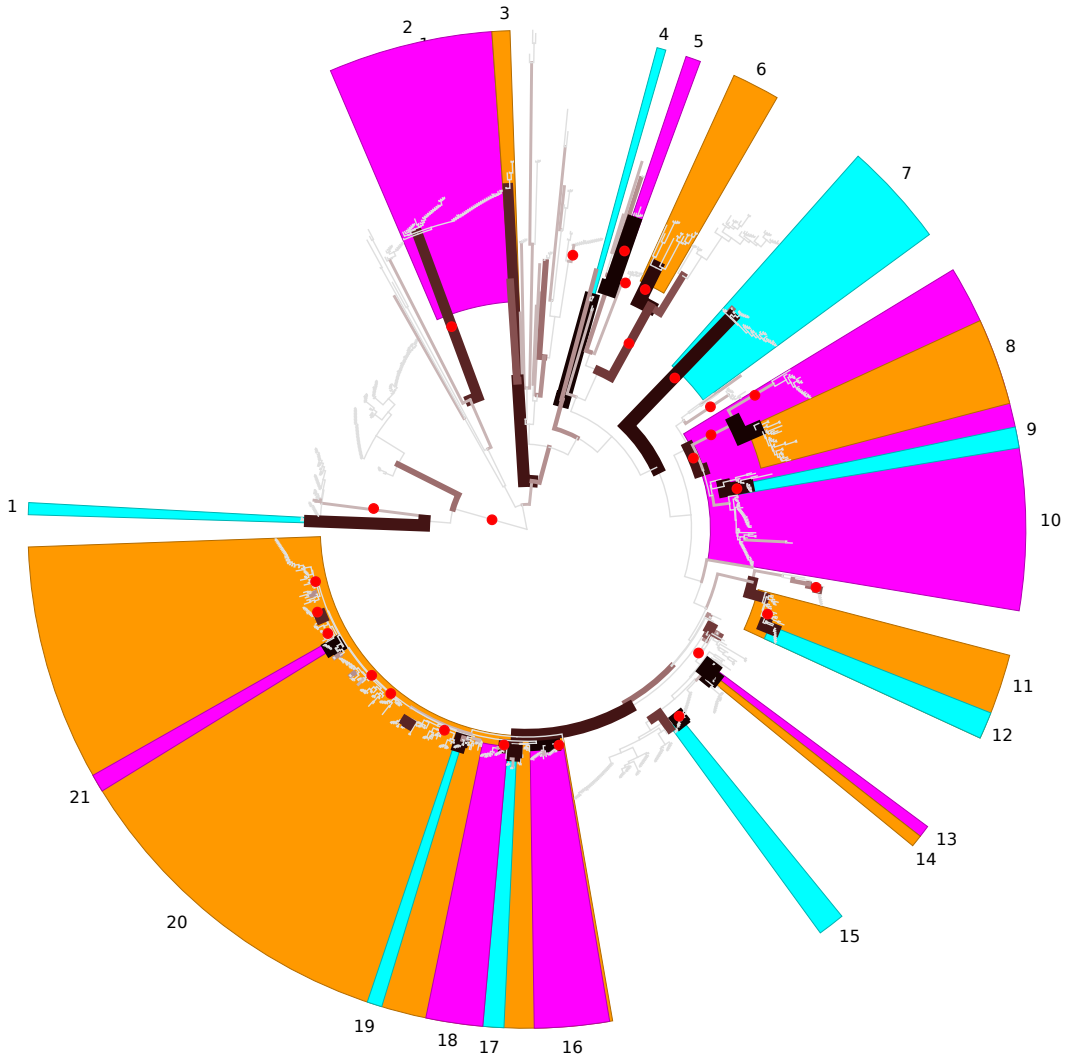


Figure 4.14: *Vibrionaceae* tree with posterior probability of a branch having an adaptive event indicated by line thickness and colour shade. Branches with posterior probability of having an adaptive event bigger than 0.6 are highlighted. This point estimate has resulted in 21 branches with adaptive events on them which has divided the tree into 22 sections. AdaptML has inferred 26 branches as having adaptive events on them which are indicated by a red dot. As it can be seen 8 of the branches inferred to have an adaptive event on them by our method have also been inferred by AdaptML to create ecologically distinct populations. 12 of the branches inferred by AdaptML have very small posterior expectation under our model.

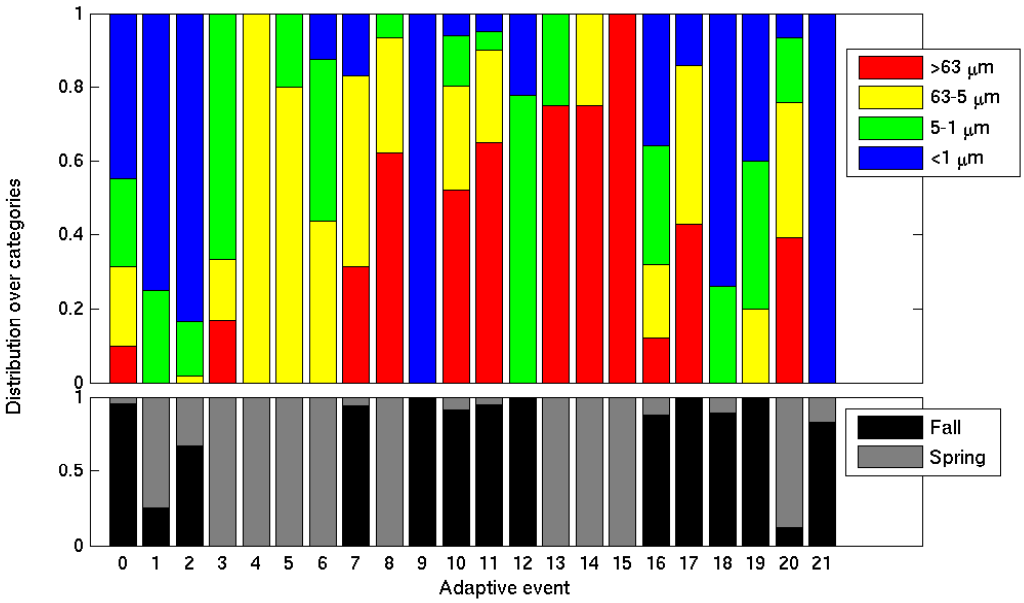


Figure 4.15: Distribution of size fraction and season for the sections that are distinctly distributed over the environmental categories. Most sections have a strong seasonal association. Furthermore most clusters are strongly associated with being free living or particle associated.

4.5 Discussion

We have presented a method to demarcate ecologically distinct lineages on a phylogenetic tree given a random sample from a community with their observed environmental categories. We use Bayesian statistics to take uncertainty in the data into account and to infer model parameters.

In our null model there are no adaptive events on the tree and all isolates have the same distribution over the environmental categories. In our alternative model there is at least one adaptive event on the tree that results in distinct distribution over the environmental categories. To measure the power of our method, we applied it to simulated datasets. Branches with adaptive events that demarcate section which contain many isolates will have a higher posterior probability of having an adaptive

event. In addition if the adaptive event on the branch results in a large change in the distribution over the categories, the branch will have a higher posterior probability of having an adaptive event. We used reversible jump MCMC to perform model selection. We also used model averaging to ensure that we take uncertainty in the model into account. Using simulated datasets, we showed that our method is conservative and underestimates the number of adaptive events and the rate of adaptation on the tree. We applied our method to *Salmonella enterica* subspecies *enterica* pubMLST dataset. Our method inferred 13 branches that had distinct distributions over the host ranges. We detected that Typhimurium serovars are made up of two lineages with distinct distributions over the host ranges. We also applied our method to a dataset that was analysed by AdaptML previously. Our results is broadly similar to AdaptML results.

In our model we assume that there is no uncertainty about the tree. However in reality the tree will not be known exactly. This uncertainty about the tree can be easily accounted for by applying our method to a sample of trees from the posterior distribution of the trees that are produced by Bayesian phylogenetic programs such as MrBayes and BEAST (Huelsenbeck and Ronquist, 2001; Drummond *et al.*, 2012). Another potential issue is recombination. In bacteria recombination moves genetic elements across lineages and could potentially move adaptive elements between genomes in the population. If the recombination rate is extremely high, the population will be homogeneous in terms of the distribution over the categories and it will be very difficult to infer any adaptive events. In addition in such populations there is little signal to infer the tree. Thus our method is not suitable for highly recombining species. If the rate of recombination is not extremely high, our method

sees the movements of adaptive elements as new adaptive events. Another point to note is that our method only detects differentiation along the observed environmental categories. There might be further differentiations along other categories that have not been observed.

Chapter 5

Discussion and Further Work

5.1 Summary of previous chapters

5.1.1 Biased recombination

We defined biased recombination in contrast to free recombination where the likelihood of recombination between a donor and recipient is independent of the amount of evolutionary distance between them. In population genetics recombination is either ignored or assumed to be free despite empirical evidence to the contrary (Roberts and Cohan, 1993; Zawadzki *et al.*, 1995). Biased recombination can be caused by several factors. Geographical and ecological structuring due to clonal reproduction of bacteria, implies that isolates that are more closely related are more likely to recombine (Feil and Spratt, 2001; Didelot and Maiden, 2010). Purifying selection can potentially purge recombinant isolates that are due to recombination between distantly related cells. Empirical evidence shows that recombination in bacteria is homology dependent and there is a log linear relationship between relative rate of recombination and the amount of sequence divergence between donor and recipient cells

(Majewski *et al.*, 2000; Majewski, 2001). As the effect of these forces is hard to disentangle we have lumped them together and called them biased recombination.

To understand recombination and speciation in bacteria, biased recombination has to be taken into account. Using a free recombination model when recombination is actually biased will result in underestimating recombination rate. In addition theoretical studies have shown that under neutral model of evolution biased recombination can potentially be important in speciation. In these models biased recombination can be a cohesive force between similar strains, while it is rare between distantly related strains so that isolates that pass a divergence threshold will not converge due to recombination (Falush *et al.*, 2006; Hanage *et al.*, 2006; Fraser *et al.*, 2007).

We introduced a novel model of biased recombination in bacteria based on the ClonalOrigin model (Didelot *et al.*, 2010). As the model likelihood is intractable, we used approximate Bayesian computation (Pritchard *et al.*, 1999; Beaumont *et al.*, 2002) to infer model parameters including the rate of bias in the recombination process. Informative summary statistics of the recombination process including the four gamete test (Hudson and Kaplan, 1985), LD (Maynard Smith *et al.*, 1993) and homoplasy (Maynard Smith and Smith, 1998) were used to infer the model parameters. We used the Markovian nature of the model to introduce an efficient method for simulating from the model. Instead of simulating whole genomes, we simulate the summary statistics directly which leads to fast simulations that is essential for inference under ABC.

Application to simulated datasets indicated that the chosen summary statistics were highly informative about the parameters of interest. We

also showed that parameters in a large range can be inferred and that the method is robust to slight misspecification in the clonal genealogy. Finally the method was applied to 13 whole *Bacillus cereus* genomes and the rate of bias in the recombination process for the dataset was estimated. We used posterior predictive distribution to show that biased recombination model is a good fit to the data. We used Monte Carlo simulation to estimate the relative effect of recombination to mutation r/m for the dataset. We also estimated the instantaneous rate of convergence and divergence for the dataset and concluded that all pairs of genomes in this dataset are likely to diverge in the near future due to much stronger divergence effect of mutation.

The framework introduced here is highly flexible. The model can easily be changed to probe other aspects of recombination which are less studied. There is no need to calculate likelihoods and as long as simulation from the model is relatively fast, questions can be answered. There are several avenues for further work in this project. The clonal genealogy could potentially be inferred as a parameter of the model. This can be hampered by the high computational cost of the inference procedure. The inference procedure already takes several days to be completed on a server with 12 cores. Adding genealogy as an extra parameter could substantially increase the time taken for inference. However it should be possible to infer the clonal genealogy for four or five isolates in our model. In phylogenetics, closely related isolates are usually the source of uncertainty as recombination obscures the relationship between them. Our method takes recombination between closely related isolates into account, hence it can be used to infer the clonal genealogy between few closely related isolates.

Another aspect of recombination that is usually ignored is the directional flow of DNA material. Gene flow can be asymmetric in terms of the direction

of movement of DNA and there is empirical evidence for instances of it in eukaryotes (Gornall *et al.*, 1998; Schunter *et al.*, 2011). One can imagine many scenarios for which this would be the case for bacteria too. Our model can be easily extended for the recombination between different clades to be directional. Although the extra number of parameters can increase the computational cost of the model.

5.1.2 Ecologically distinct lineages

We introduced a novel model to demarcate ecologically distinct lineages on a phylogenetic tree. In prokaryotes due to their clonal mode of reproduction one would expect for closely related isolates to share functions within a community. There is growing empirical evidence to support this hypothesis, but it is unclear what level of hierarchy delineates such clusters (Ferris *et al.*, 2003; Sikorski and Nevo, 2005).

We proposed a model for the occurrence of adaptive events on a phylogenetic tree and how these events affect the distribution over the environmental categories. We then used Bayesian methodology to account for uncertainty in the data and to infer the boundaries of lineages that have differential distribution over the categories. We used RJMCMC (Green, 1995) to perform model selection and averaging (Kass and Raftery, 1995; Hoeting *et al.*, 1999). We used extensive simulation studies to characterize the power of the method. As expected adaptive events that demarcate section that contain many isolates or result in large changes in the distribution over the categories or both are easier to detect. Our model is conservative and underestimates the number of adaptive events as some of them cannot be detected. We also estimated the rate of adaptive events on the branches of the tree.

We applied our method to two real datasets. We first applied our method to the *Salmonella enterica* subspecies *enterica* PUBMLST dataset. We showed that there is decisive evidence that some of the lineages are distinctly distributed over the host categories. The detected lineages where in agreement with the known literature on serovar host association. We also detected two lineages in Typhimurium serovar with distinct distributions over the host categories although Typhimurium is typically known as the broad host serovar.

One potential issue with our method is that we assume the tree is known. However the uncertainty about the tree can be easily taken into account. Given a sample from the posterior distribution of trees, we can apply our method to each tree separately. This can be done in parallel on many cores to speed up the process but for large number of trees could computationally become too expensive.

5.2 Further work

In this section we present two projects to be developed that are related to the work presented in this thesis. In the first project we propose to model transmission events on top of a phylogenetic tree. We present an overview of the current methods on inferring transmission trees and how it could be improved on by modelling it on top of a tree. In the second project we propose to develop computationally efficient methods for inferring clonal genealogy in presence of recombination. The *de facto* method for inferring clonal genealogy in presence of recombination for bacteria is ClonalFrame (Didelot and Falush, 2007), which can take a long time to produce results for large datasets. We propose to develop a computationally efficient version of ClonalFrame.

5.2.1 Inferring transmission trees

Observing temporally and or spatially clustered cases of infections, clinicians usually assume an outbreak, but inferring whom infected whom is done using ad hoc rules based on the data available (Ou *et al.*, 1992; Holmes *et al.*, 1993; Walker *et al.*, 2012). When an outbreak is suspected, many questions arise. Is this a real outbreak? Do the cases have different sources and are they distinct from each other? If it is an outbreak, is there a common source for all the cases? Or, are they the result of transmission between different individuals?

Reconstructing transmission trees will help to detect the source of outbreaks (Cottam *et al.*, 2008; Spada *et al.*, 2004), to design and evaluate intervention policies (Ferguson *et al.*, 2001b; Donnelly *et al.*, 2003), to predict how the outbreak is going to develop (Ferguson *et al.*, 2001a), to understand how transmission occurs (Spada *et al.*, 2004; Cottam *et al.*, 2008) and to examine evolutionary forces affecting the pathogen (Leitner and Albert, 1999).

In presence of epidemiological data such as location and timings of the infections one can construct possible transmission trees, but there are likely to be many possible transmission trees that are consistent with these data. An additional source of information is the genetic data. The last decade has seen huge advances in sequencing technologies and the cost of sequencing has reduced substantially (Zhou *et al.*, 2010). Front-line organisations such as hospitals have started to sequence pathogen genomes. There has been attempts to sequence the genome of sampled isolates in real time during an outbreak (Eyre *et al.*, 2012; Bertelli and Greub, 2013). These genomic data will add substantial information that can be used in constructing the transmission trees.

Ou *et al.* (1992) was the first use of sequence data and phylogenetics to identify possible transmission events between an HIV-positive dentist and his patients. Over the last two decades phylogenetic reconstruction has often been used to find possible transmission trees (Holmes *et al.*, 1993; Ferguson *et al.*, 2001b; Cottam *et al.*, 2008). However the phylogenetic tree is usually used in an ad hoc manner to postulate possible transmission events. Statistical methodology to infer transmission trees using sequence data has been lacking.

Jombart *et al.* (2011) made the first attempt in constructing transmission trees from the sequence data. They highlight the issue that transmission trees are not the same as phylogenetic trees and one should distinguish between them. Their method uses the sequence data and the time of collection of the samples to find the best transmission tree describing the data. They use other epidemiological data if the tree is not resolved. Ypma *et al.* (2012) introduces a more unified method that assumes the temporal, spatial and the genetic data are independent of each other and constructs likelihood function for each source of data and uses a Markov Chain Monte Carlo (MCMC) methodology to infer the transmission trees. Morelli *et al.* (2012) point out that genetic and epidemiological data are not independent of each other and introduces a framework that models the correlation between data sources and uses MCMC methodology to infer the transmission tree. Teunis *et al.* (2013) models the transmission tree as a transmission matrix and assumes genetic and epidemiological data are independent of each other and constructs likelihood functions for the transmission matrix which is estimated using a MCMC methodology. The transmission matrix is then converted to a transmission tree.

Method

All of the methods on construction of transmission trees use the sequence data in a simplistic way (Jombart *et al.*, 2011; Ypma *et al.*, 2012; Morelli *et al.*, 2012; Teunis *et al.*, 2013). These methods count the number of differences between sequences and assume samples with fewer differences have higher probability of transmission between each other. They all ignore the diversity of the pathogen within the host and assume a single host is equivalent to a single sequence. When a host is infected, the pathogen will evolve and given time a certain amount of diversity will develop and different pathogen samples from the host may have different sequences (Wright *et al.*, 2011; Morelli *et al.*, 2013). This host can infect other individuals at different times and the strains passed on may not be identical.

The reason that a phylogenetic tree is different from a transmission tree is that a phylogenetic tree shows the evolutionary relationship between pathogen sequences (isolated from hosts), but transmission trees shows the direction of infection from one host to another. Here we argue that modelling the transmission events and epidemiological data on top of the phylogenetic tree is more appropriate and will lead to better and more robust estimates for transmission trees rather than counting number of SNPs between sequences as is done in previous methods.

Figure 5.1 shows a transmission tree between 8 different hosts. Each host is numbered from 1 to 8 and in brackets the time of infection of the host and the time that the pathogen isolate was taken is given. The phylogenetic tree for the pathogen sequences sampled from each host is shown in Figure 5.2. Figure 5.3 displays the phylogenetic tree with transmission events shown as stars and pathogen evolution in each host is indicated by colour. For instance the dark blue colour indicates evolution of the pathogen in host

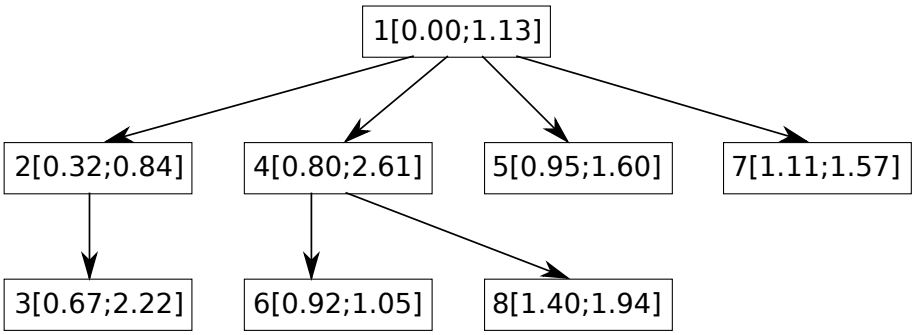


Figure 5.1: Illustration of a transmission tree between 8 hosts. Each host is numbered from 1 to 8. Time that the host was infected and the time that the pathogen isolate was taken from the host are shown in brackets.

1 while the green colour indicates the evolution of the pathogen in host 4. This representation of the phylogenetic tree clearly shows the transmission tree while highlighting within host evolution. Phylogenetic trees captures more information from the sequence data than a simple counting of SNPs. This extra information can be used to improve the inference of transmission trees.

We will start with assuming that phylogenetic tree is known and model the transmission events and the epidemiological data on top of the phylogenetic tree. In absence of recombination and using whole genomes, there will be little uncertainty about the phylogenetic tree. Given the model, we can use Bayesian methodology to infer the possible transmission trees accounting for uncertainty in the data. To take the uncertainty in the phylogenetic tree estimation into account, as a second stage one can infer the phylogenetic and transmission trees in one statistical framework. This can potentially be implemented as a plug-in for BEAST (Drummond and Rambaut, 2007) which has a big user community and due to its graphical user interface is accessible by non-expert users.

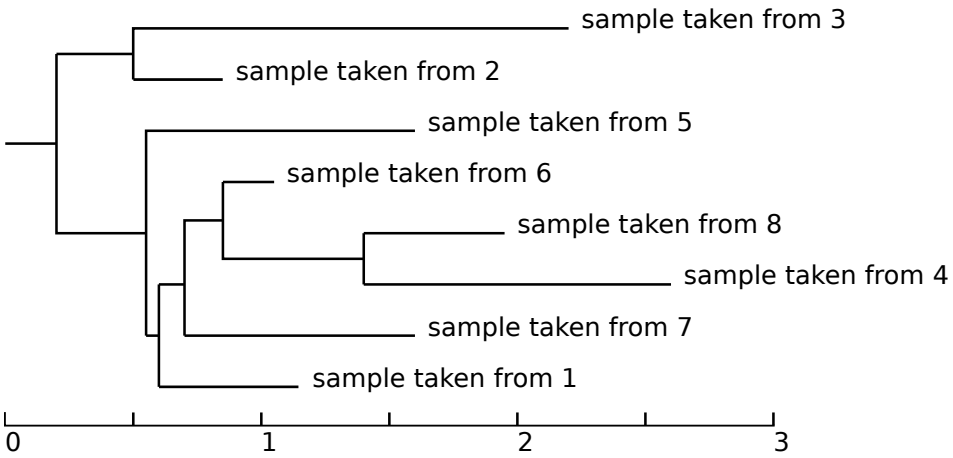


Figure 5.2: Phylogenetic tree of the sequences obtained from the pathogen isolates taken from each host. Using intuition, one might think that host 6 infected hosts 8 and 4.

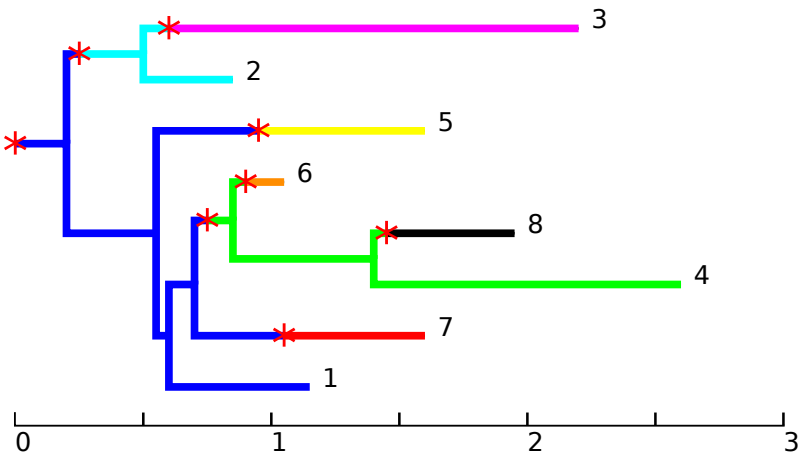


Figure 5.3: Phylogenetic Tree of the pathogen sequences with transmission events shown as stars and pathogen evolution within each host shown as a different colour. For instance evolution of pathogen within host one is shown with blue colour. In this figure it is clear which host infected which other host indicating how much of the pathogen evolution happened within each host.

5.2.2 Efficient inference of clonal genealogy from whole bacterial genomes

The clonal genealogy is the fundamental structure that reveals the evolutionary relationship between bacterial isolates from a population and is the basis for many evolutionary analysis (Awadalla, 2003). Many aspects of bacterial biology can be better understood in the context of their clonal genealogy (Sikorski and Nevo, 2005; Harris *et al.*, 2010; Croucher *et al.*, 2011). Bacteria are clonal organisms that divide by binary fission, but they do recombine. These recombination events have to be taken into account for accurate reconstruction of genealogies (Schierup and Hein, 2000). ClonalFrame (Didelot and Falush, 2007) is the *de facto* method for inferring the clonal genealogy between bacterial isolates accounting for recombination. It was developed for MLST sequence data, but has been applied to whole genome datasets. It can handle 30-40 whole bacterial genomes and takes several weeks to run. With the explosion of whole genome sequencing, there is a need for methods that can handle hundreds to thousands of whole bacterial genomes and produce results in relatively short amount of time.

Method

In this project we propose to develop computationally efficient models and methods to infer the clonal genealogy of large samples of bacterial isolates from their whole genomes accounting for recombination. ClonalFrame uses an MCMC to infer the clonal genealogy and uses an HMM to infer the recombination events. Most of the computational time is spent inferring recombination events. First research avenue is to reparametrise the ClonalFrame model to remove the estimation of ancestral sequences at the

internal nodes of the genealogy. In this reparametrisation it may be possible to integrate out the recombination events and only infer the clonal genealogy. This should speed up the inference procedure by several orders of magnitude. In this reparametrisation it may also be possible to use an efficient proposal for recombination events that speeds up the inference procedure. Another possibility is to divide the genome into short chunks where we assume either recombination is present or not. This would reduce the computational burden on inferring recombination events by several folds. A completely different approach would be to use maximum likelihood estimation or some form of heuristics to estimate the clonal genealogy. This could for instance be an iterative method that infers recombinant segments at each step and removes them until it has converged to the correct clonal genealogy.

Bibliography

Achtman, M. and M. Wagner, 2008 Microbial diversity and the genetic nature of microbial species. Nat. Rev. Microbiol. **6**: 431–40.

Achtman, M., J. Wain, F.-X. Weill, S. Nair, Z. Zhou *et al.*, 2012 Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. PLoS Pathog. **8**: e1002776.

Amann, R. I., W. Ludwig and K. H. Schleifer, 1995 Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. **59**: 143–69.

Ansari, M. A. and X. Didelot, 2014 Inference of the properties of the recombination process from whole bacterial genomes. Genetics **196**: 253–65.

Arias, C. a. and B. E. Murray, 2009 Antibiotic-resistant bugs in the 21st century–a clinical super-challenge. N. Engl. J. Med. **360**: 439–43.

Awadalla, P., 2003 The evolutionary genomics of pathogen recombination. Nat. Rev. Genet. **4**: 50–60.

Beaumont, M. a., 2010 Approximate Bayesian Computation in Evolution and Ecology. Annu. Rev. Ecol. Evol. Syst. **41**: 379–406.

Beaumont, M. a., J.-M. Cornuet, J.-M. Marin and C. P. Robert, 2009 Adaptive approximate Bayesian computation. Biometrika **96**: 983–990.

Beaumont, M. a., W. Zhang and D. J. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics **162**: 2025–35.

Bertelli, C. and G. Greub, 2013 Rapid bacterial genome sequencing: methods and applications in clinical microbiology. Clin. Microbiol. Infect. : 1–11.

Brockwell, A. E., 2006 Parallel Markov Chain Monte Carlo Simulation by Pre-Fetching. J. Comput. Graph. Stat. **15**: 246–261.

Burrus, V., G. Pavlovic, B. Decaris and G. Guédon, 2002 Conjugative transposons: the tip of the iceberg. Mol. Microbiol. **46**: 601–610.

Cadillo-Quiroz, H., X. Didelot, N. L. Held, A. Herrera, A. Darling *et al.*, 2012 Patterns of gene flow define species of thermophilic Archaea. PLoS Biol. **10**: e1001265.

Chen, G. K., P. Marjoram and J. D. Wall, 2009 Fast and flexible simulation of DNA sequence data. Genome Res. **19**: 136–42.

Chen, I., P. J. Christie and D. Dubnau, 2005 The ins and outs of DNA transfer in bacteria. Science **310**: 1456–60.

Cohan, F. M., 2002 Sexual isolation and speciation in bacteria. Genetica **116**: 359–70.

Cohan, F. M. and E. B. Perry, 2007 A systematics for discovering the fundamental units of bacterial diversity. Curr. Biol. **17**: R373–86.

Cottam, E. M., J. Wadsworth, A. E. Shaw, R. J. Rowlands, L. Goatley *et al.*, 2008 Transmission pathways of foot-and-mouth disease virus in the United Kingdom in 2007. PLoS Pathog. **4**: e1000050.

Courvalin, P., 1994 Transfer of antibiotic resistance genes between gram-positive and gram-negative bacteria. Antimicrob. Agents Chemother. **38**: 1447–51.

Croucher, N. J., S. R. Harris, C. Fraser, M. a. Quail, J. Burton *et al.*, 2011 Rapid pneumococcal evolution in response to clinical interventions. Science (80-.). **331**: 430–4.

Csilléry, K., M. G. B. Blum, O. E. Gaggiotti and O. François, 2010 Approximate Bayesian Computation (ABC) in practice. Trends Ecol. Evol. **25**: 410–8.

Darling, A. C. E., B. Mau, F. R. Blattner and N. T. Perna, 2004 Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. **14**: 1394–403.

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

Davison, J., 1999 Genetic exchange between bacteria in the environment. Plasmid **42**: 73–91.

Didelot, X., M. Achtman, J. Parkhill, N. R. Thomson and D. Falush, 2007 A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? Genome Res. **17**: 61–8.

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

Didelot, X., R. Bowden, T. Street, T. Golubchik, C. Spencer *et al.*, 2011a Recombination and population structure in *Salmonella enterica*. PLoS Genet. **7**: e1002191.

Didelot, X., R. G. Everitt, A. M. Johansen and D. J. Lawson, 2011b Likelihood-free estimation of model evidence. Bayesian Anal. **6**: 49–76.

Didelot, X. and D. Falush, 2007 Inference of bacterial microevolution

using multilocus sequence data. Genetics **175**: 1251–66.

Didelot, X., D. Lawson, A. Darling and D. Falush, 2010 Inference of homologous recombination in bacteria using whole-genome sequences. Genetics **186**: 1435–49.

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

Didelot, X. and M. C. J. Maiden, 2010 Impact of recombination on bacterial evolution. Trends Microbiol. **18**: 315–22.

Didelot, X., G. Méric, D. Falush and A. E. Darling, 2012 Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics **13**: 256.

Donnelly, C. a., A. C. Ghani, G. M. Leung, A. J. Hedley, C. Fraser *et al.*, 2003 Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. Lancet **361**: 1761–6.

Doolittle, W. F. and R. T. Papke, 2006 Genomics and the bacterial species problem. Genome Biol. **7**: 116.

Drummond, A. J. and A. Rambaut, 2007 BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. **7**: 214.

Drummond, A. J., M. a. Suchard, D. Xie and A. Rambaut, 2012 Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. **29**: 1969–73.

Dubnau, D., 1999 DNA uptake in bacteria. Annu. Rev. Microbiol. **53**: 217–44.

Dykhuizen, D. E., 1998 Santa Rosalia revisited: why are there so many species of bacteria? Antonie Van Leeuwenhoek **73**: 25–33.

Eyre, D. W., T. Golubchik, N. C. Gordon, R. Bowden, P. Piazza

et al., 2012 A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. BMJ Open **2**.

Falush, D., C. Kraft, N. S. Taylor, P. Correa, J. G. Fox *et al.*, 2001 Recombination and mutation during long-term gastric colonization by Helicobacter pylori: estimates of clock rates, recombination size, and minimal age. Proc. Natl. Acad. Sci. U. S. A. **98**: 15056–61.

Falush, D., M. Stephens and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics **164**: 1567–87.

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**: 2045–53.

Fearnhead, P., 2007 Computational methods for complex stochastic systems: a review of some alternatives to MCMC. Stat. Comput. **18**: 151–171.

Fearnhead, P. and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159**: 1299–1318.

Fearnhead, P. and P. Donnelly, 2002 Approximate likelihood methods for estimating local recombination rates. J. R. Stat. Soc. Ser. B (Statistical Methodol. **64**: 657–680.

Feil, E. J. and B. G. Spratt, 2001 Recombination and the population structures of bacterial pathogens. Annu. Rev. Microbiol. **55**: 561–90.

Ferguson, N. M., C. a. Donnelly and R. M. Anderson, 2001a The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. Science (80-.). **292**: 1155–60.

Ferguson, N. M., C. a. Donnelly and R. M. Anderson, 2001b

Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. Nature **413**: 542–8.

Ferris, M. J., M. Kühl, A. Wieland and D. M. Ward, 2003 Cyanobacterial ecotypes in different optical microenvironments of a 68 degrees C hot spring mat community revealed by 16S-23S rRNA internal transcribed spacer region variation. Appl. Environ. Microbiol. **69**: 2893–8.

Fisher, R. A., 1930 The genetical theory of natural selection.

Fitch, W. and E. Margoliash, 1967 Construction of phylogenetic trees. Science (80-.). **155**: 279–284.

Fraser, C., E. J. Alm, M. F. Polz, B. G. Spratt and W. P. Hanage, 2009 The bacterial species challenge: making sense of genetic and ecological diversity. Science (80-.). **323**: 741–6.

Fraser, C., W. P. Hanage and B. G. Spratt, 2007 Recombination and the nature of bacterial speciation. Science (80-.). **315**: 476–80.

Frost, L. S., 2009 Conjugation, Bacterial. In M. Schaechter, editor, *Encycl. Microbiol. (Third Ed.*. Academic Press, Oxford, third edit edition, 517–531.

Gans, J., M. Wolinsky and J. Dunbar, 2005 Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science (80-.). **309**: 1387–90.

Gelman, A., X.-L. Meng and H. Stern, 1996 Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. **6**: 733–759.

Goodman, S. D. and J. J. Scocca, 1988 Identification and arrangement of the DNA sequence recognized in specific transformation of Neisseria gonorrhoeae. Proc. Natl. Acad. Sci. U. S. A. **85**: 6982–6.

Gornall, R. J., P. M. Hollingsworth and C. D. Preston, 1998

Evidence for spatial structure and directional gene flow in a population of an aquatic plant, Potamogeton coloratus. Heredity (Edinb). **80**: 414–421.

Green, P. J., 1995 Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. Biometrika **82**: 711.

Grelaud, A., C. Robert, J. M. Marin, F. Rodolphe and J. F. Taly, 2009 ABC likelihood-free methods for model choice in Gibbs random fields. Bayesian Anal. **4**: 317–336.

Griffiths, R. C. and P. Marjoram, 1997 An ancestral recombination graph. In P. Donnelly and S. Tavaré, editors, *IMA Vol. Math. Popul. Genet.*, Progress in Population Genetics and Human Evolution. Springer, Verlag, New York, 257–270.

Guttman, D. and D. Dykhuizen, 1994 Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science (80-.). **266**: 1380–1383.

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

Hanage, W. P., B. G. Spratt, K. M. E. Turner and C. Fraser, 2006 Modelling bacterial speciation. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **361**: 2039–44.

Harris, S. R., E. J. Feil, M. T. G. Holden, M. a. Quail, E. K. Nickerson *et al.*, 2010 Evolution of MRSA during hospital transmission and intercontinental spread. Science (80-.). **327**: 469–74.

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**: 97–109.

Hill, W. G. and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38**: 226–231.

Hoeting, J. A., D. Madigan, A. E. Raftery and C. T. Volinsky,

1999 Bayesian Model Averaging : Tutorial. Stat. Sci. : 382–401.

Holmes, E. C., L. Q. Zhang, P. Simmonds, a. S. Rogers and a. J. Brown, 1993 Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. J. Infect. Dis. **167**: 1411–4.

Horner-Devine, M. C. and B. J. M. Bohannan, 2006 Phylogenetic clustering and overdispersion in bacterial communities. Ecology **87**: S100–8.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23**: 183–201.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159**: 1805–17.

Hudson, R. R. and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111**: 147–64.

Huelsenbeck, J. P. and F. Ronquist, 2001 MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**: 754–5.

Hugenholtz, P., B. M. Goebel and N. R. Pace, 1998 Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. **180**: 4765–74.

Hunt, D. E., L. a. David, D. Gevers, S. P. Preheim, E. J. Alm *et al.*, 2008 Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science (80-.). **320**: 1081–5.

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.

Johnson, Z. I., E. R. Zinser, A. Coe, N. P. McNulty, E. M. S.

Woodward *et al.*, 2006 Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. Science (80-.). **311**: 1737–40.

Johnston, C., B. Martin, G. Fichant, P. Polard and J.-P. Claverys, 2014 Bacterial transformation: distribution, shared mechanisms and divergent control. Nat. Rev. Microbiol. **12**: 181–96.

Jombart, T., R. M. Eggo, P. J. Dodd and F. Balloux, 2011 Reconstructing disease outbreaks from genetic data: a graph approach. Heredity (Edinb). **106**: 383–90.

Jukes, T. and C. Cantor, 1969 Evolution of Protein Molecules. In H. N. MUNRO, editor, *Mamm. Protein Metab.*. Academic Press, New York, 21–132.

Kass, R. E. and A. E. Raftery, 1995 Bayes Factors. J. Am. Stat. Assoc. **90**: 773.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**: 893–903.

Kimura, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**: 111–20.

Kimura, M., 1985 *The neutral theory of molecular evolution*. Cambridge University Press.

Kingman, J. F. C., 1982a On the Genealogy of Large Populations. J. Appl. Probab. **19**: 27.

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

Koeppel, A., E. B. Perry, J. Sikorski, D. Krizanc, A. Warner *et al.*,

2008 Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc. Natl. Acad. Sci. U. S. A. **105**: 2504–9.

Kuhner, M. K., J. Yamato and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156**: 1393–401.

Lawson, D. J., G. Hellenthal, S. Myers and D. Falush, 2012 Inference of population structure using dense haplotype data. PLoS Genet. **8**: e1002453.

Le Minor, L., 1988 Typing of *Salmonella* species. Eur. J. Clin. Microbiol. Infect. Dis. **7**: 214–8.

Leitner, T. and J. Albert, 1999 The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc. Natl. Acad. Sci. U. S. A. **96**: 10752–10757.

Levin, B. R. and O. E. Cornejo, 2009 The population and evolutionary dynamics of homologous gene recombination in bacterial populations. PLoS Genet. **5**: e1000601.

Li, N. and M. Stephens, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165**: 2213–33.

Lorenz, M. G. and W. Wackernagel, 1994 Bacterial gene transfer by natural genetic transformation in the environment. Microbiol. Rev. **58**: 563–602.

Low, K. B. and D. D. Porter, 1978 Modes of gene transfer and recombination in bacteria. Annu. Rev. Genet. **12**: 249–87.

Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell *et al.*, 1998 Multilocus sequence typing: a portable approach to the

identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U. S. A. **95**: 3140–5.

Majewski, J., 2001 Sexual isolation in bacteria. FEMS Microbiol. Lett. **199**: 161–9.

Majewski, J. and F. M. Cohan, 1999 DNA sequence similarity requirements for interspecific recombination in *Bacillus*. Genetics **153**: 1525–33.

Majewski, J., P. Zawadzki, P. Pickerill, F. M. Cohan and C. G. Dowson, 2000 Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. J. Bacteriol. **182**: 1016–23.

Marjoram, P., J. Molitor, V. Plagnol and S. Tavare, 2003 Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. U. S. A. **100**: 15324–8.

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

Martinko, J. M., D. A. Stahl and J. Parker, 1997 *Biology of Microorganisms*. Prentice Hall, Upper Saddle River, NJ, 8 edition.

Matic, I., F. Taddei and M. Radman, 1996 Genetic barriers among bacteria. Trends Microbiol. **4**: 69–72.

Maynard Smith, J., 1999 The detection and measurement of recombination from sequence data. Genetics **153**: 1021–7.

Maynard Smith, J. and N. H. Smith, 1998 Detecting recombination from gene trees. Mol. Biol. Evol. **15**: 590–9.

Maynard Smith, J., N. H. Smith, M. O'Rourke and B. G. Spratt, 1993 How clonal are bacteria? Proc. Natl. Acad. Sci. U. S. A. **90**: 4384–4388.

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. T. and N. J. Cardin, 2005 Approximating the coalescent with recombination. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **360**: 1387–93.

Medini, D., D. Serruto, J. Parkhill, D. A. Relman, C. Donati *et al.*, 2008 Microbiology in the post-genomic era. Nat. Rev. Microbiol. **6**: 419–30.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equation of State Calculations by Fast Computing Machines. J. Chem. Phys. **21**: 1087.

Metropolis, N. and S. Ulam, 1949 The Monte Carlo Method. J. Am. Stat. Assoc. **44**: 335.

Moran, P. a. P., 1958 Random processes in genetics. Math. Proc. Cambridge Philos. Soc. **54**: 60.

Morelli, G., X. Didelot, B. Kusecek, S. Schwarz, C. Bahlawane *et al.*, 2010 Microevolution of Helicobacter pylori during prolonged infection of single hosts and within families. PLoS Genet. **6**: e1001036.

Morelli, M. J., G. Thébaud, J. Chadœ uf, D. P. King, D. T. Haydon *et al.*, 2012 A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. PLoS Comput. Biol. **8**: e1002768.

Morelli, M. J., C. F. Wright, N. J. Knowles, N. Juleff, D. J. Paton *et al.*, 2013 Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. Vet. Res. **44**: 12.

Namouchi, A., X. Didelot, U. Schöck, B. Gicquel and E. P. C.

Rocha, 2012 After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. Genome Res. **22**: 721–34.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154**: 931–42.

Nishiguchi, M. and B. Jones, 2005 Microbial Biodiversity within the Vibrionaceae. In J. Seckbach, editor, *Orig. Evol. Biodivers. Microb. life*, volume 6 of *Cellular Origin, Life in Extreme Habitats and Astrobiology*. Springer Netherlands, 533–548.

Nübel, U., P. Roumagnac, M. Feldkamp, J.-H. Song, K. S. Ko *et al.*, 2008 Frequent emergence and limited geographic dispersal of methicillin-resistant *Staphylococcus aureus*. Proc. Natl. Acad. Sci. U. S. A. **105**: 14130–5.

Ochman, H., J. G. Lawrence and E. a. Groisman, 2000 Lateral gene transfer and the nature of bacterial innovation. Nature **405**: 299–304.

Ou, C. Y., C. A. Ciesielski, G. Myers, C. I. Bandea, C. C. Luo *et al.*, 1992 Molecular epidemiology of HIV transmission in a dental practice. Science (80-.). **256**: 1165–71.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16**: 1791–8.

Rabsch, W., H. L. Andrews, R. A. Kingsley, R. Prager, H. Tschäpe *et al.*, 2002 *Salmonella enterica* serotype Typhimurium and its host-adapted variants. Infect. Immun. **70**: 2249–55.

Rayssiguier, C., D. S. Thaler and M. Radman, 1989 The barrier to recombination between *Escherichia coli* and *Salmonella* is disrupted in mismatch-repair mutants. Nature **342**: 396–401.

Ripley, B., 2009 *Stochastic Simulation*. Wiley Series in Probability and Statistics. Wiley.

Robert, C. P., J.-M. Cornuet, J.-M. Marin and N. S. Pillai, 2011 Lack of confidence in approximate Bayesian computation model choice. Proc. Natl. Acad. Sci. U. S. A. **108**: 15112–7.

Roberts, M. S. and F. M. Cohan, 1993 The effect of DNA sequence divergence on sexual isolation in *Bacillus*. Genetics **134**: 401–8.

Sawasaki, Y., K. Inomata and K. Yoshida, 1996 Trans-kingdom conjugation between Agrobacterium tumefaciens and Saccharomyces cerevisiae, a bacterium and a yeast. Plant Cell Physiol. **37**: 103–6.

Schierup, M. H. and J. Hein, 2000 Consequences of recombination on traditional phylogenetic analysis. Genetics **156**: 879–91.

Schunter, C., J. Carreras-Carbonell, E. Macpherson, J. Tintoré, E. Vidal-Vijande *et al.*, 2011 Matching genetics with oceanography: directional gene flow in a Mediterranean fish species. Mol. Ecol. **20**: 5167–81.

Sheppard, S. K., N. D. McCarthy, D. Falush and M. C. J. Maiden, 2008 Convergence of Campylobacter species: implications for bacterial evolution. Science (80-.). **320**: 237–9.

Sikorski, J. and E. Nevo, 2005 Adaptation and incipient sympatric speciation of *Bacillus* simplex under microclimatic contrast at "Evolution Canyons" I and II, Israel. Proc. Natl. Acad. Sci. U. S. A. **102**: 15924–9.

Sisson, S. A., Y. Fan and M. M. Tanaka, 2007 Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. U. S. A. **104**: 1760–5.

Spada, E., L. Sagliocca, J. Sourdis, A. R. Garbuglia, V. Poggi *et al.*, 2004 Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus

infection. J. Clin. Microbiol. **42**: 4230–6.

Stumpf, M. P. H. and G. a. T. McVean, 2003 Estimating recombination rates from population-genetic data. Nat. Rev. Genet. **4**: 959–68.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–95.

Takuno, S., T. Kado, R. P. Sugino, L. Nakhleh and H. Innan, 2012 Population genomics in bacteria: a case study of *Staphylococcus aureus*. Mol. Biol. Evol. **29**: 797–809.

Tavaré, S., D. J. Balding, R. C. Griffiths, P. Donnelly and S. Tavare, 1997 Inferring coalescence times from DNA sequence data. Genetics **145**: 505–18.

Teunis, P., J. C. M. Heijne, F. Sukhrie, J. van Eijkeren, M. Koopmans *et al.*, 2013 Infectious disease transmission as a forensic problem: who infected whom? J. R. Soc. Interface **10**: 20120955.

Thomas, C. M. and K. M. Nielsen, 2005 Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat. Rev. Microbiol. **3**: 711–21.

Thornton, K. and P. Andolfatto, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. Genetics **172**: 1607–1619.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen and M. P. Stumpf, 2009 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. R. Soc. Interface **6**: 187–202.

Uzzau, S., D. J. Brown, T. Wallis, S. Rubino, G. Leori *et al.*, 2000 Host adapted serotypes of emphSalmonella enterica. Epidemiol. Infect. **125**: 229–55.

Vos, M., 2009 Why do bacteria engage in homologous recombination? Trends Microbiol. **17**: 226–32.

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

Vulić, M., F. Dionisio, F. Taddei and M. Radman, 1997 Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc. Natl. Acad. Sci. U. S. A. **94**: 9763–7.

Wakeley, J. and S. Lessard, 2003 Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. Genetics **164**: 1043–53.

Walker, a. S., D. W. Eyre, D. H. Wyllie, K. E. Dingle, R. M. Harding *et al.*, 2012 Characterisation of *Clostridium difficile* hospital ward-based transmission using extensive epidemiological data and molecular typing. PLoS Med. **9**: e1001172.

Wall, J. D., 2004 Estimating recombination rates using three-site likelihoods. Genetics **167**: 1461–73.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**: 256–76.

Weinbauer, M. G., 2004 Ecology of prokaryotic viruses. FEMS Microbiol. Rev. **28**: 127–81.

Weinbauer, M. G. and F. Rassoulzadegan, 2003 Are viruses driving microbial diversification and diversity? Environ. Microbiol. **6**: 1–11.

Whelan, S., P. Liò and N. Goldman, 2001 Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. **17**: 262–72.

Wiuf, C., 2000 A coalescence approach to gene conversion. Theor. Popul. Biol. **57**: 357–67.

Wiuf, C. and J. Hein, 1999 Recombination as a point process along sequences. Theor. Popul. Biol. **55**: 248–59.

Wiuf, C. and J. Hein, 2000 The coalescent with gene conversion. Genetics **155**: 451–62.

Wright, C. F., M. J. Morelli, G. Thébaud, N. J. Knowles, P. Herzyk *et al.*, 2011 Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. J. Virol. **85**: 2266–75.

Wright, S., 1931 Evolution in Mendelian Populations. Genetics **16**: 97–159.

Yahara, K., M. Kawai, Y. Furuta, N. Takahashi, N. Handa *et al.*, 2012 Genome-wide survey of mutual homologous recombination in a highly sexual bacterial species. Genome Biol. Evol. **4**: 628–40.

Ypma, R. J. F., a. M. a. Bataille, A. Stegeman, G. Koch, J. Wallinga *et al.*, 2012 Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc. R. Soc. B Biol. Sci. **279**: 444–50.

Zawadzki, P., M. S. Roberts and F. M. Cohan, 1995 The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust. Genetics **140**: 917–32.

Zhou, X., L. Ren, Y. Li, M. Zhang, Y. Yu *et al.*, 2010 The next-generation sequencing technology: a technology review and future perspective. Sci. China. Life Sci. **53**: 44–57.