SimBac: A Stochastic Simulator of Bacterial Evolution with Homologous Recombination

Tom Brown

Supervisors: Dr. Daniel Wilson & Dr. Nicola De Maio

Project 2

Abstract

Phylogenetic analysis of bacterial strains has been used in a number of clinical applications including outbreak detection of bacterial strains. The decrease in both time and cost of sequencing full genomes has resulted in much larger data sets, with higher resolution, which has generated the need for new and efficient computational phylogenetic methods. In particular, to match the change from multi-locus sequence typing (MLST) to full genome data, giving approximately 1000-times more information, efficient, genome-wide simulation methods are required. Recombination is known to complicate phylogenetic inference, therefore by developing an efficient bacterial genome evolution simulator with homologous recombination we can test the accuracy of a number of techniques aimed at reconstructing the ancestries of samples.

Introduction

Sequencing of bacterial genomes has rapidly become more cost-effective in recent years, allowing for greater genetic detail when analysing bacteria such as *Escherichia coli* and *Staphylococcus aureus* [1, 2]. With the increased facility with which DNA sequences can be extracted have come a number of phylogenetic tools designed to infer the genealogies underlying the ancestries of the samples. These genealogies have been further used in a number of clinical applications including outbreak detection [3] or detecting the phylogenetic context for emergence of levels of strain virulence [4, 5]. The inferred ancestries of species have a great number of uses when identifying the species or strain of unknown sequences and the age of strains, given by the phylogeny, offers insight into the origin of strains. Phylogenies can also be used to identify the geographical origin of strains, to study selective forces such as in drug resistance and phenotype-genotype associations via 'phylogeography' [6].

The Coalescent Model

The coalescent process has been established as a statistical model for genealogical relationships under which one can simulate the ancestries of lineages within a population. The Kingman Coalescent model [7] can be thought of as a generative model, simulating genealogies backward in time, with genealogies coalescing with rate $\binom{k}{2}$ per generation, where k is the number of lineages present in the current sample, with each pair of lineages equally likely to coalesce. Forward in time, a coalescent event is a reproduction event producing two separate lineages within the species. An example of a coalescent tree can be seen in Figure 1 with the three samples traced back to the Most Recent Common Ancestor (MRCA) of the three sampled lineages. By tracing the length of each branch of the tree, we are able to calculate the ancestral distance between present day lineages of different species. On each branch of the coalescent tree, mutations are often assumed to occur as a Poisson process with rate $\theta/2$, where $\theta = 2N_e\mu$ is the given mutation rate. Here, N_e is the effective population size and μ is the mutation rate per site per generation. This method of coalescence and mutation creates genetic diversity in the sampled population, with those lineages further from each other in the coalescent tree likely to be separated by more mutations than those lineages closer together.

Coalescent theory can be used to estimate key evolutionary parameters [8] such as mutation rate, θ, via the Watterson estimator [9], Tajima's D [10] or Bayesian likelihood-based inference, e.g. [11], the age of the MRCA [12, 13, 14] or the recombination rate, ρ [15, 16, 17]. Estimating these quantities for a phylogeny gives an indication of how much recombination certain species undergo, the different mutation rates experienced across species and can identify a strain as the source of certain phenotypic traits.

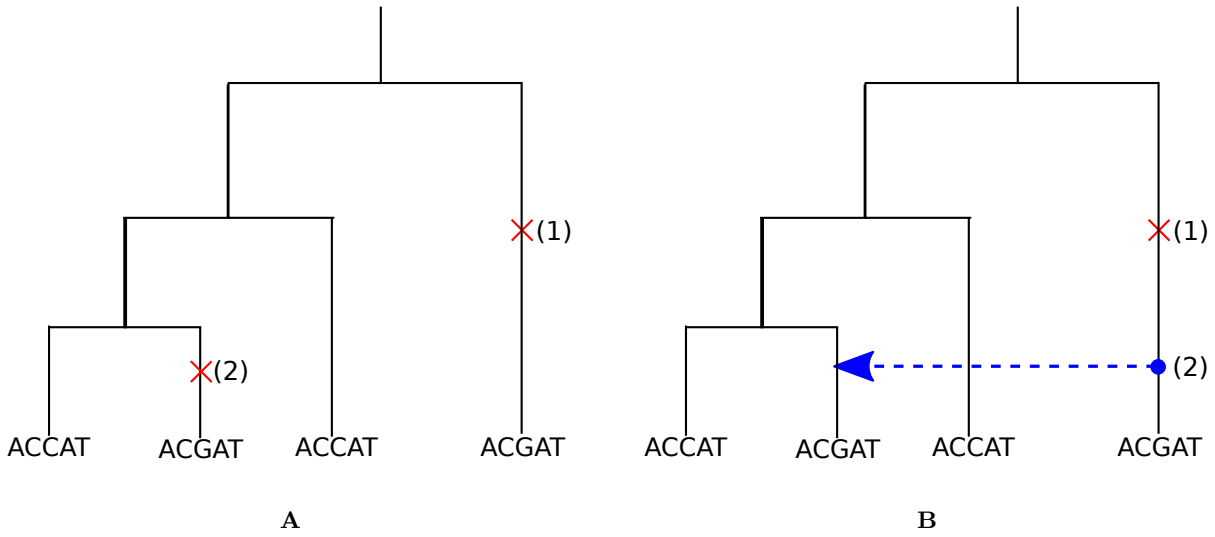

Fig. 1: Example of a homoplasy. **A** demonstrates the same mutation occurring on two lineages, resulting in two sequences with the same mutation from C→G at the third site arising through two separate mutations events (1) and (2). **B** demonstrates the same sequences arising through a single mutation (1) followed by a recombination event (2) where the fourth lineage acts as the donor and the second sequences receives the mutated site as part of the recombinant interval.

Tree Reconstruction

A number of techniques have been developed for reconstructing ancestral lineages using sequence data. Three techniques used here are Unweighted-Pair Group Method with Arithmetic Means (UPGMA) [18], Neighbor Joining (NJ) [19] and Maximum Likelihood (ML) [20]. Tree topology can be constructed backward in time by coalescing those lineages which have highest similarity at the sequence level. Under UPGMA and NJ this takes the form of a scoring matrix, where the distance between each sequence is determined by a metric calculating the pair-wise genetic distance between each sequence. The Maximum Likelihood approach attempts to find the tree that gives the highest probability of observing the data. RAxML [21] performs this by adding lineages one at a time and relaxing branch lengths, giving the tree as the phylogeny with highest likelihood. By estimating the mutation rate, the length of each branch can be further inferred by finding the value which best corresponds to the sequence data.

Homologous Recombination

Bacterial species do not undergo the same methods of gene sharing as is common in Eukaryotes as they do not reproduce sexually, meaning there is no mixing of DNA sequences between parents. Bacteria, however, reproduce clonally by 'binary fission', with all of the genetic material coming from one parent. Homologous recombination, however, allows for facultative exchange of DNA between two cells, whereby short fragments of DNA can be shared or transmitted. There are three main methods by which bacteria undergo homologous recombination: transduction (where a virus transfers DNA from one cell to another), transformation (where DNA is taken up by a recipient cell from its surroundings) and conjugation (which requires contact between the donor and recipient cells, where the DNA fragment is transferred from one cell to another via a pilus bridging the two cells) [22].

This is a key aspect of gene transfer in bacteria: Homologous recombination occurs not just between isolates from the same bacterial species at varying rates [23], but also between different bacterial strains, allowing for wider genetic diversity [24]. The rate at which bacteria undergo both within- and between-species recombination varies largely from species-to-species [23], with some bacteria sharing large portions of their genetic material with other bacterial strains.

The introduction of recombination to the standard coalescent model creates two events that can occur in the history: a coalescent event or a recombinant event. Backward in time, a recombinant event manifests itself as a lineage splitting into two parent lineages, a donor and a recipient. A larger number of recombination events results in more lineages in the

simulated graph, increasing the height of the graph and thus the time taken to simulate. We will refer to the coalescent tree with recombination as the Ancestral Recombination Graph (ARG).

The Clonal Frame

The clonal frame [25] is defined as the phylogeny of the sites that have not undergone recombination in each lineage. Under the model of homologous recombination described above, this can be described as the ancestry of the recipient lineages, excluding those lineages donating DNA fragments. Under the standard neutral model with bacterial recombination, the clonal frame follows the standard coalescent model [26]. If a bacterial species undergoes many recombination events, it is expected that the small number of sites contained in the clonal frame will result in a poor estimate of the clonal frame by conventional phylogenetic methods. Recent results, however, found that up to a moderate recombination rate of 1% per-site, per-generation, reconstruction of the clonal frame using standard methods was remarkably accurate with >97% accuracy in reconstructing the topology of the clonal frame [27]. Here, we aim to further test this finding by performing phylogenetic reconstruction under simulated datasets with higher recombination rates.

Homoplasies

Under the infinite-sites assumption [28], one expects only one mutation at any site in the genome in a species' lineage. This is motivated by the relatively small number of mutations evidenced in a species genealogy compared to its genome size. A number of phylogenetic tools incorporate this assumption, but is an approximation of the truth, given the finite length of the genomes. This implies that any site variation must be due to a single mutation at some point in the ancestry of the species of interest. Any site variation that requires two or more mutations on the clonal frame to explain the variation is called a homoplasy and can only be explained by a recombination event under the infinite-sites assumption (Fig. 1). Although only an approximation, this is useful for detecting recombinant events when both the sequences and the clonal frame are known. By detecting the number of homoplasic events that have occurred in the ancestry of a set of bacterial strains, we can infer the prevalence of recombinant events in the past.

Methods

To simulate the evolution of the bacterial population, an ARG is simulated to determine the recombinant breaks for recombinant lineages and coalescent events. The ARG includes both those lineages in the clonal frame, but also includes 'donor' lineages responsible for the DNA fragments imported during recombination events. Working backwards in time, starting with the lineages in the present-day sample, the time to next coalescent event is distributed as an exponential distribution with rate $\binom{k}{2}$ where k is the number of lineages currently in the sample. The time to next recombinant event is traditionally distributed exponentially with rate $k\rho/2$ where ρ is the recombination rate per generation per lineage.

Comparison to SimMLST

We wish to test the accuracy of existing phylogenetic tools under homologous recombination and as such require a simulator of bacterial genomes under a model of coalescence and recombination. The work carried out here was designed to extend the software SimMLST [29] which was built to simulate multilocus sequence typing (MLST) data of usually a few ($\sim$7) 450-500bp fragments [30]. As such, SimMLST takes a long time to simulate entire genomes. Here we present SimBac, which has been built with several new features, allowing efficient simulation of whole-genome data.

In SimBac, we assume a circular genome reflecting the usual state of affairs found in bacteria. If loci, rather than the entire genome are chosen to be simulated, the user can define a finite distance separating each fragment, as opposed to independent loci, as in SimMLST. A new addition is the introduction of recombination events from external species to the ARG. These events are simulated similarly to internal, within-species recombination events, however the recombinant intervals created this way undergo a high rate of mutation to simulate the foreign sequence being introduced to the lineage. Figure 2 shows an example ARG with external recombination. The imported fragment (shown in red) is inherited by all subsequent lineages in the ARG.

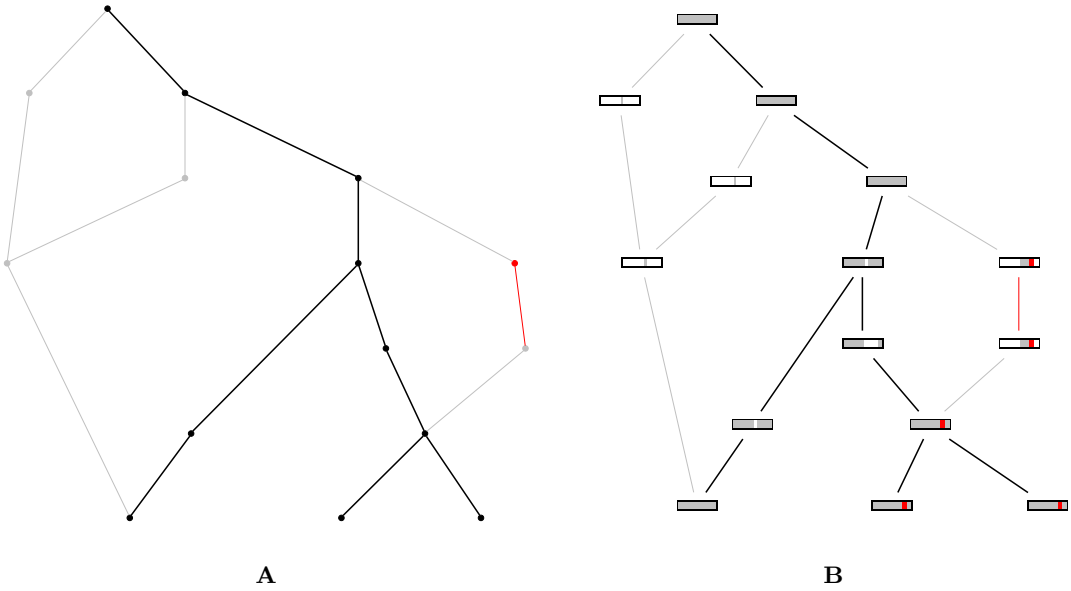

Fig. 2: Examples of Ancestral Recombination Graphs (ARGs). The Nodes represent lineages and the events that make up the ARG. **A** shows the clonal frame in black, with the non-clonal lineage in grey and a recombinant event involving an external species in red. **B** shows the ancestral material in each lineage of the ARG. The ancestral material is shown at each node in grey. Material shown in red represents genetic material imported from an external species. The graphs were written in the DOT language [34] and adapted from SimMLST [29].

To reduce the time and memory used storing the ancestral information at each lineage of the ARG, the ancestral material is stored as a set of intervals, instead of as boolean vectors with length equal to the length of the genome, which can be very large. Under this implementation each lineage requires less memory to store the genetic material currently present. In particular, it allows for fast checking of fully-coalesced material, which can be removed from the ARG. Finally, in order to avoid rejection sampling, as used in SimMLST, of recombinant events that arise when recombination events do not intersect the ancestral material present as the chosen lineage, the recombination rates and probabilities are calculated as follows.

Not all genetic material contained in each lineage will be ancestral to the genomes of the sampled individuals. As such, the ancestral material of each lineage is defined as the set of nucleotides in the genome that are ancestral to at least one of the sampled lineages. As we are only interested in the recombination events that affect the ancestral material in the present-day lineages, we wish to avoid simulating any recombinant events that do not fall into this criterion. For lineages in the clonal frame, we wish to include any recombinant intervals that intersect the ancestral material in the lineage. For non-clonal lineages we only include recombinant events where the recombinant interval intersects the ancestral material and splits the ancestral material in two non-empty sets, avoiding any events where the recombinant interval contains all of the ancestral material present in the lineage. We simulate recombination events in a clonal lineage in which the entire ancestral material is recombinant as we are still interested in clonal lineages which do not contain any of the ancestral material and indeed we expect this at very high recombination rates.

For each lineage, the ancestral material is made up of a set of intervals $I_1, ..., I_b$, where $I_i = [s_i, e_i]$ and the length of each interval is given by $L_i = e_i - s_i + 1$. To take into account the circularity of the bacterial genome, define $e_0 = e_b - G$, where G is the length of the genome. For a given site-specific recombination rate R and average recombinant break length

δ, the rate of a recombinant interval first affecting the first element of an ancestral interval, I_i, is given by:

$$\begin{aligned}\frac{R_{x,s_i}}{2} &= \frac{R}{2}\sum_{j=0}^{s_i-e_{i-1}-1}\left(1-\delta^{-1}\right)^j \\ &= \frac{R}{2}\left(\sum_{j=0}^{\infty}\left(1-\delta^{-1}\right)^j - \sum_{j=s_i-e_{i-1}}^{\infty}\left(1-\delta^{-1}\right)^j\right) \\ &= \frac{R}{2}\left(\delta - \delta\left(1-\delta^{-1}\right)^{s_i-e_{i-1}}\right) \\ \frac{R_{x,s_i}}{2} &= \frac{R\delta}{2}\left(1-\left(1-\delta^{-1}\right)^{s_i-e_{i-1}}\right) \end{aligned} \tag{1}$$

Where $\left(1-\delta^{-1}\right)^j$ for $j = 1, \ldots, (s_i - e_{i-1} - 1)$ is the probability that a recombinant interval has length which includes the site s_i, given that the recombinant interval begins j base pairs before s_i. This result follows from the cumulative distribution function of a geometric distribution, where:

$$\mathbb{P}\left\{G > g\right\} = (1-p)^g \tag{2}$$

For a geometric distribution with probability of success p.

The rate of a recombinant interval beginning at any other element of ancestral material in a lineage is given by $R/2$. Therefore if we define $L_x = \sum_{i=1}^{b_x} L_{x,i}$ to be the total amount of ancestral material in lineage x where b_x is the total number of ancestral blocks, the rate of recombination satisfying $r \cap a_x \neq \emptyset$, where r is the recombinant interval and a_x is the ancestral material in lineage x, is given by:

$$\frac{R_{x,a}}{2} = \left(\sum_{i=1}^{b_x}\frac{R_{s_i}}{2}\right) + \frac{R}{2}\left(L_x - b_x\right) \tag{3}$$

For a non-clonal lineage, we also satisfy $a - r \neq \emptyset$. The recombination rate in non-clonal lineages is:

$$\frac{R'_{x,a}}{2} = \frac{R_{x,a}}{2} - \left(\sum_{i=1}^{b_x}\frac{R_{x,s_i}}{2}\left(1-\delta^{-1}\right)^{G-(s_i-e_{i-1})}\right) - \frac{R}{2}\left(1-\delta^{-1}\right)^{G-1}\left(L_x - b_x\right) \tag{4}$$

Where $\left(1-\delta^{-1}\right)^{G-(s_i-e_{i-1})}$ is the probability of a recombinant interval starting at site s_i and ending beyond the point e_{i-1}, taking the entire ancestral material in the recombinant interval. Similarly, $\left(1-\delta^{-1}\right)^{G-1}$ gives the probability of a recombinant interval including the entire genome, i.e: starting at site i and having length of at least $G-1$, including all sites of the genome, ancestral or otherwise.

For a clonal lineage, the probability of the start site of a recombinant interval being the first site of an ancestral interval and satisfying $r \cap a_x \neq \emptyset$ is given by:

$$\mathbb{P}(s_i) = \frac{R_{x,s_i}}{R_{x,a}} \tag{5}$$

For i in $1 \ldots b_x$, and the probability of starting at any other site of the ancestral material is:

$$\frac{R}{R_{x,a}} \tag{6}$$

The end site of the recombinant interval is then chosen according to a geometric distribution with mean δ, incorporating the memoryless property of the geometric distribution to model a recombination interval commencing at the start of an ancestral block.

For a non-clonal lineage, the probability of a recombinant interval starting at the beginning of an ancestral interval and satisfying $r \cap a_x \neq \emptyset$ and $a_x - r \neq \emptyset$ is given by:

$$\mathbb{P}'(s_i) = \frac{R_{x,s_i}\left(1-\left(1-\delta^{-1}\right)^{G-(s_i-e_{i-1})}\right)}{R'_{x,a}} \tag{7}$$

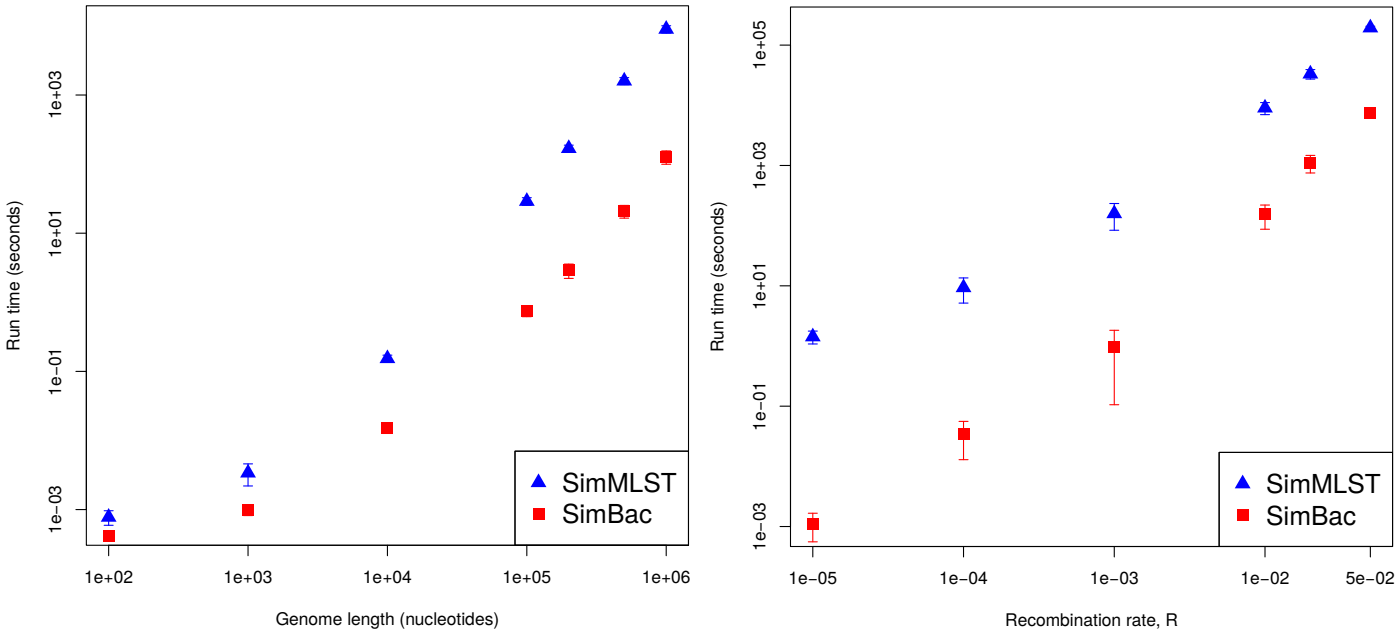

Fig. 3: Comparison of run-time between SimMLST and SimBac. **A** shows average time to simulate the ARG for a fixed recombination rate $R = 0.01$ and genome length from 100bp to 1Mbp. **B** shows the average time to simulate the ARG for a fixed genome length of 1Mbp and recombination rate increasing from $R = 0$ to $R = 0.05$. 100 Simulations were performed (10 for SimMLST $R = 0.02$ and $R = 0.05$) and error bars show ± 1 standard deviation.

The end site of the recombinant interval is then chosen from a truncated geometric distribution with mean δ conditional on $|r| < G - (s_i - e_{i-1})$. The probability of the recombinant interval beginning at another ancestral site is given by:

$$\frac{R\left(1 - \left(1 - \delta^{-1}\right)^{G-1}\right)}{R'_{x,a}} \tag{8}$$

and the end site is chosen via a truncated geometric distribution, conditioned on $|r| < G - 1$.

To verify the accuracy of the new implementation, simulations were performed in both SimMLST and SimBac to calculate the average height of the tree and the number of recombination events (Fig. S1). The similarity in tree height suggests correct simulation of the ARG and the number of recombination events implies correct adaptation of the rejection sampling through the equations above.

For a full description of the algorithmic implementation, see Supplementary Material: *Algorithmic Implementation.*

Results

To test the efficiency of SimBac in comparison to SimMLST, a series of simulations for a range of parameter values were tested (Fig. 3). Figure 3A shows, for a fixed recombination rate of $R = 0.01$, the time taken to simulate the ARG for increasing genome length. The time to simulate the ARG is significantly less for SimBac and in particular, when the length of the genome reached 1Mbp, comparable to bacterial genome length, there was an approximately 50-fold reduction in running time with the ARG requiring approximately 3 minutes to simulate in SimBac, compared to $2\frac{1}{2}$ hours in SimMLST. Figure 3B demonstrates that for a fixed genome of 1Mbp the time taken to simulate the ARG was reduced using SimBac. The time to simulate the ARG for 100 taxa and 1Mbp genomes with $R = 0.05$ was approximately 2 hours for SimBac and approximatly 53 hours for SimMLST.

Of particular interest when simulating bacterial evolution is the accuracy of phylogenetic tree construction under recombination. Three methods of phylogenetic tree inference were tested for increasing rates of recombination, namely UPGMA [18], NJ, [19] and ML [20]. Simulations were performed for a species with 100 isolates, with each having a

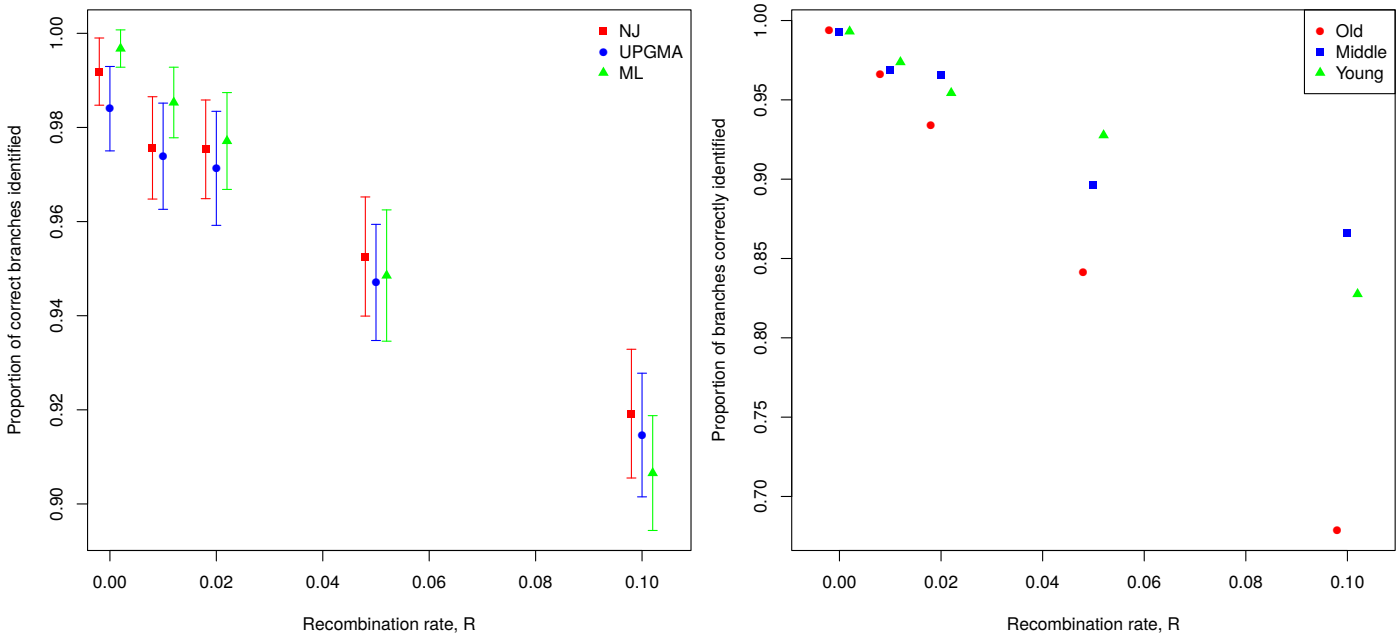

Fig. 4: Accuracy of constructing the clonal frame for increasing recombination rate, R. In **A**, phylogeny estimation methods used were Neighbour Joining (NJ), Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Maximum Likelihood (ML). The genome length and mutation rate, θ were fixed at 1Mbp and 0.01, respectively. 100 simulations (10 for $R = 0.1$) were performed and error bars represent ± 1 standard deviation. In **B**, the accuracy of branches are grouped by age into old, middle-aged and young branches so as to create equal-sized groups. Shown is proportion of accurately reconstructed branches for each age group by ML.

genome of length 1Mbp. The rate of mutation, θ was fixed at 0.01 per-site, per-generation and the recombination rate, R was increased from 0 to 0.1 per-site, per-generation.

The proportion of branches correctly identified was calculated by finding the symmetric Robinson-Foulds metric [35], giving the number of branches in each tree not present in the other tree, as such it gives twice the number of incorrectly inferred branches in the clonal frame. The proportion of correctly identified branches is then given by:

$$P = \frac{2(n-1) - \left(\frac{S_{RF}}{2}\right)}{2(n-1)} \tag{9}$$

Where n is number of lineages and S_{RF} is the Robinson-Foulds score. $2(n-1)$ is the number of branches in a tree with n tips, however given that the clonal frame and the predicted trees both include the same tips, all tip branches will be correctly identified, meaning that of $2(n-1)$ branches, n will always be correct. In figure 4A, an accuracy of 91% for ML with recombination rate $R = 0.1$ corresponds to 18 branches incorrect out of a possible 98 branches.

In [27], tree topology reconstruction was found to be accurate at a recombination rate of 1% per-site, per-generation, corresponding to $R = 0.01$. The simulations performed with SimMLST [29] restricted the rate of recombination that could be tested, with the time taken to simulate genomes growing too large. Figure 4A demonstrates that, by simulating genomes using SimBac, up to $R = 0.1$, or a recombination rate of 10% per-site, per-generation, the accuracy of tree topology construction is reasonably accurate, with only 8% or 9% error in reconstructing the clonal frame, which may be acceptable for many purposes.

The accuracy of reconstructing certain branches within the phylogenetic trees offers insight into how much faith one can place in tree reconstruction techniques. Figure 4B shows the accuracy of branch reconstruction by mean branch age. Branches were grouped into old (distance from root <1.32) middle-aged ($1.32 <$ distance < 2.09) and young (distance > 2.09), chosen to create equal-sized age-groups, and a branch was counted as correctly identified if it is contained

in the reconstructed tree. There is a striking difference in the accuracy of branch reconstruction when looking at the older branches, nearer the root of the tree. At a recombination rate of $R = 0.1$, less than 70% of older branches are correctly identified, compared with 85-90% accuracy for the younger branches. This result demonstrates that when using techniques such as RAxML [21], one can have a certain degree of confidence with the immediate grouping of taxa and the topologies of branches near the tips of the tree, but nearer the root, it is difficult to say with much confidence that the inferred topology is the true topology. This is consistent with the result found in [36] that recombination leads to 'star-like' phylogenies, with poorly resolved interval branching.

To offer an explanation for the difficulty found in reconstructing the clonal frame from DNA sequences, we investigated the proportion of site mutations that were homoplasies. Here, any site with more than one mutated lineage that could not be described by a single mutation on the clonal frame was defined as a homoplasy, in accordance with the infinite-sites assumption. Figure 5 demonstrates that even at relatively low levels of recombination (1%), the majority of mutations affecting multiple lineages are homoplasic. The high proportion of mutated sites that are explained by recombination even at a relatively modest level of recombination suggests that enough information about the true phylogeny can be recovered from the distribution of mutations across the population. By excluding homoplasic sites from the analysis, it is likely that the majority of informative mutated sites that affect a large number of sequences would be removed. This may offer an explanation for the distorted phylogenies recovered in [27] by removing homoplasies from the analysis.

Discussion

SimBac presents a significant reduction in time to simulate full bacterial genomes under homologous recombination and offers several new features. Simulating 100 isolates with 1Mbp length genome with recombination rate of 5% takes a couple of hours in SimBac, whereas the same simulations take approximately two days with SimMLST (Figure 3B). Given the ease with which long sequences with high recombination rates can now be simulated, testing existing phylogenetic software has revealed their effectiveness under high recombination rates (Figure 4). At recombination rates of $R = 0.05$, standard tree reconstruction programs are able to reconstruct the clonal frame with fairly high accuracy (>94%) but the techniques begin to fall down at $R = 0.1$.

The tree reconstruction techniques demonstrate the difficulty in detecting the true topology of the 'deep' branches of the tree, or those closer to the MRCA. This result suggests that caution should be exercised when inferring phylogenies of species which undergo high rates of recombination, for example in *Neisseria gonorrhoeae* [23].

The rate of recombination in a species is often given in terms of the proportion of variable sites in the genome due to recombination and mutation, r/m. Relating this fraction to the rates used in the model above, we incorporate the notation of [32], where r/m can be expressed as:

$$\frac{r}{m} = \frac{\rho\ \delta\ \nu}{\mu} \tag{10}$$

Where ν is the rate of mutation in an interval introduced to the genome through recombination and all rates are per-site, per-generation. In the model presented above, internal recombination results in mutations being introduced with rate μ, the same as the rate of mutation forward in time. As such, the ratio r/m can be calculated by finding $R\delta$. In Figure 4, as $\delta = 500$ for all simulations, a recombination rate of $R = 0.1$ corresponds to r/m= 50. This value is less than that seen for *Flavobacterium psychrophilum* (r/m = 63.6) or *Pelagibacter ubique* (SAR 11) (r/m = 63.1) [23] corresponding to $R = 0.13$. For such highly recombinant species, traditional score-based or maximum-likelihood phylogenetic techniques can not be used to reliably infer the true topology of the clonal frame and especially not for older branches of the tree. However, for bacterial species such as *Staphylococcus aureus* (r/m = 0.1), *Neisseria meningitidis* (r/m = 7.1) or *Helicobacter pylori* (r/m = 10.1), this corresponds to R in the range $(0.0002, 0.02)$. For rates of recombination in this range, phylogenetic inference is still reasonable reliable, with $> 97\%$ accuracy in reconstructing the clonal frame (Figure 4A) and $\sim 95\%$ accuracy of determining internal branches of all ages (Figure 4B).

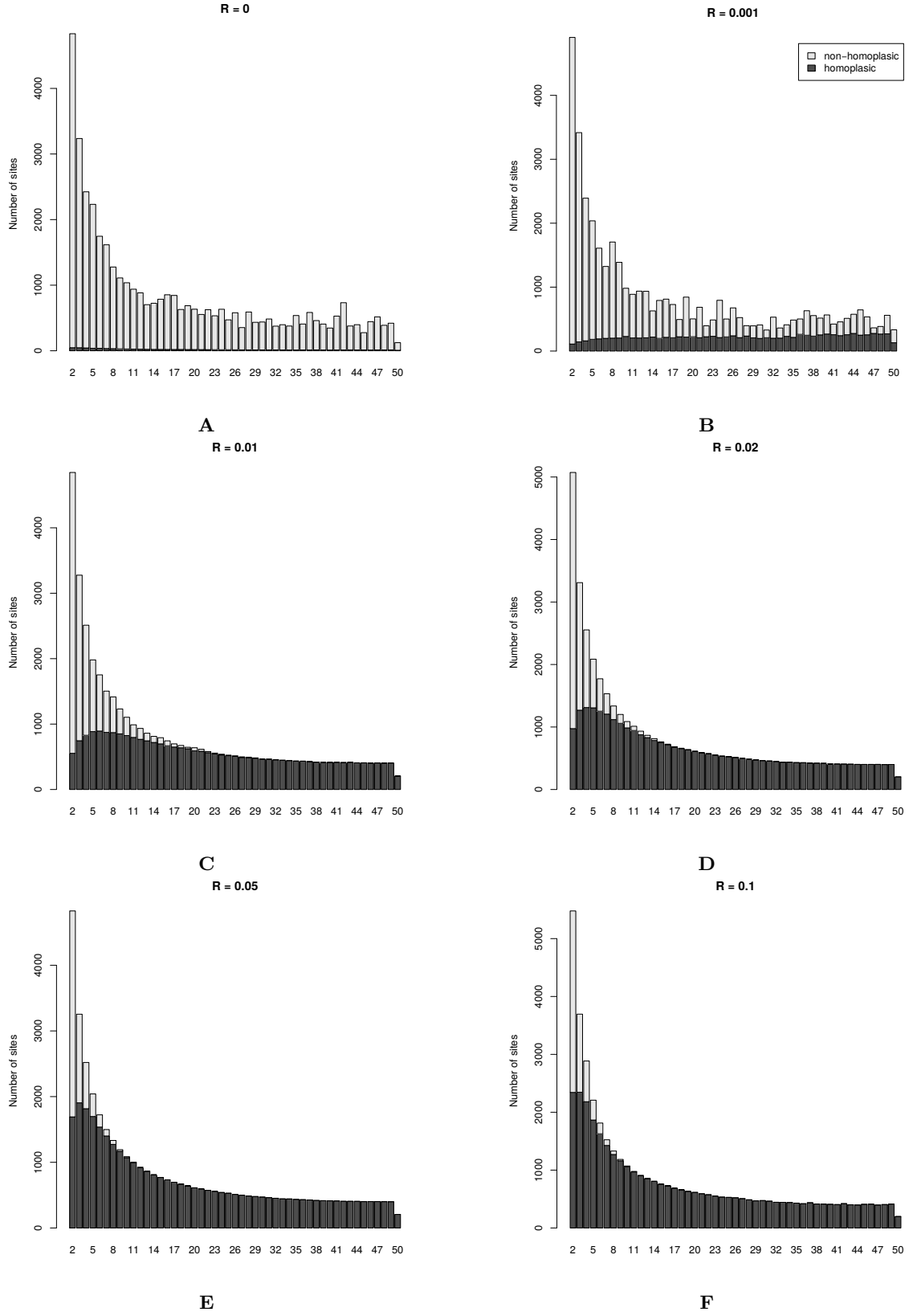

Fig. 5: Folded site frequency spectra for differing recombination rates with number of homoplasies highlighted. For a genome of length 1Mbp, 100 simulations were performed and the number of homoplasies was calculated for the generated sequences based on the true clonal frame. The separate figures show the proportion of homoplasies with: **A**: $R = 0$; **B**: $R = 0.001$; **C**: $R = 0.01$; **D**: $R = 0.02$; **E**: $R = 0.05$; **F**: $R = 0.1$

Future Work

The reduction in time to simulate bacterial genomes with high rates of recombination has allowed for more rigorous testing of phylogenetic methods aimed at reconstructing the clonal frame. Further work can be done to test the effectiveness of techniques that incorporate a model of recombination as part of their implementation. Software such as ClonalFrame [31] or ClonalOrigin [37] not only aim to reconstruct the clonal frame, but also detect the regions of the genome that have undergone recombination. In conjunction with SimBac, these techniques can be tested with the aim of finding at what thresholds of internal and external recombination these methods become non-viable. Of particular interest is the accuracy of the two techniques mentioned above as ClonalFrame models each recombination event as importing a genetic sequence from outside of the ARG, and ClonalOrigin models each recombination event as a node in the ARG. These two different modelling assumptions are likely to lead to differing levels of accuracy under different recombination scenarios.

By simulating population evolution under recombination, we can test hypotheses regarding estimation of rate parameters (e.g. ρ, θ, etc.) and we can detect selection in a population which contradicts the neutral model we have simulated under here. The simulations could be further used to detect demographic change or determine the structure of populations, applying the simulated techniques to whole-genome datasets.

Acknowledgements Algorithmic improvements were performed with Nicola De Maio and incorporated into the SimMLST code with Xavier Didelot. Many thanks to Jessica Hedge for her help with tree reconstruction and SimMLST implementation. The source code for SimBac is available from: `http://github/com/tbrown91/SimBac`. Relevant Python and R scripts for Figures 4 and 5 are available from: `http://www.dtc.ox.ac.uk/people/14/brownt/Research/Scripts_2`. NJ and UPGMA trees were constructed using the APE [38] and Phangorn [39] libraries of R, respectively. ML trees were constructed with RAxML [21]. Analysis and tree comparison was performed using the Dendropy [40] and Biopython [41] Python libraries.

References

[1] Didelot, X. Bowden, R. Wilson, D. Peto, T. Crook, D. (2012) Transforming clinical microbiology with bacterial genome sequencing *Nat. Rev. Genetic.*, **13**: 601–612.

[2] Köser et al. (2012) Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology *PLoS Pathog.*, *8*: e1002824.

[3] Paterson, G. et al. (2015) Capturing the cloud of diversity reveals complexity an heterogeneity of MRSA carriage, infection and transmission *Nature communications*, **6**: 6560.

[4] Laabei M. et al. (2014) Predicting the virulence of MRSA from its genome sequence *Genome Res.*, **24**: 839–849.

[5] Dingle, K. et al. (2014) Evolutionary History of the *Clostridium difficile* Pathogenicity Locus *Genome Biol. Evol.*, **6**(1): 36–52.

[6] Keim, P. Wagner, D. (2009) Humans and evolutionary and ecological forces shaped the phylogeography of recently emerged diseases *Nature Revies Microbiology*, **7**: 813–821.

[7] Kingman, J. (1982) On the Genealogy of Large Populations *Journal of Applied Probability*, **19**: 27–43.

[8] Fu, Y. Li, W. (1999) Coalescing into the 21st Century: An Overview and Prospects of Coalescent Theory *Theoretical Population Biology*, **56**: 1–10.

[9] Watterson, G. (1975) On the number of segregating sites in genetical models without recombination *Theoretical Population Biology*, **7**(2): 256–276.

[10] Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism *Genetics*, **123**(3): 585–595.

[11] Beerli, P. (2005) Comparison of Bayesian and maximum-likelihood inference of population genetic parameters *Bioinformatics*, **22**(3): 341–345.

[12] Fu, Y. (1996) Estimating the age of the common ancestor of a DNA sample using the number of segregating sites *Genetics*, **143**, 557–570.

[13] Tavaré, S. Balding, D. Griffiths, R. Donnelly, P. (1997) Inferring coalescence times from DNA sequence data *Genetics*, **145**, 505–518.

[14] Griffiths, R. Tavaré, S. (1996) Monte Carlo inference methods in population genetics *Math. Comput. Model.*, **23**: 141–158.

[15] Hey, J. Wakely, J. (1997) A coalescent estimator of the population recombination rate *Genetics*, **145**: 833–846.

[16] Hudson, R. Kaplan, N. (1994) Gene trees with background selection, "Non-neutral Evolution: Theories and Molecular Data" (B. Golding, Ed.), 140–153, Chapman and Hall, London.

[17] Griffiths, R. Marjoram, P. (1996) Ancestral inference from samples of DNA sequences with recombination *J. Comput. Biol.*, **3**: 479–502.

[18] Sneath, P. Sokal, R. (1973) Numerical taxonomy: the principles and practice of numerical classification. A series of books in biology. W. H. Freeman & Co, San Francisco, CA.

[19] Saito, N. Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees *Mol. Biol. Evol.*, **4**: 406–425.

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

[21] Stamatakis, A. (2014) RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies" *Bioinformatics*, **30**(9): 1312–1313.

[22] Ochman, H. Lawrence, J. Groisman, E. (2000) Lateral gene transfer and the nature of bacterial innovation *Nature*, **405**: 299–304.

[23] Vos, M. Didelot, X. (2009) A comparison of homologous recombination rates in bacteria and archaea *The ISME Journal*, **3**, 199–208.

[24] Majewski, J. Zawadzki, P. Pickerill, P. Cohan, F. Dowson, C. (2000) Barriers to Genetic Exchange between Bacterial Species: *Streptococcus pneumoniae* Transformation *J. Bacteriol.*, **182**(4), 1016–1023.

[25] Milkman, R. Bridges, M. (1990) Molecular evolution of the Escherichia coli chromosome. III. Clonal frames *Genetics*, **126**: 505–527.

[26] Wiuf, C. Hein, J. (2000) The Coalescent With Gene Conversion *Genetics*, **155**: 451–462.

[27] Hedge, J. Wilson, D. (2014) Bacterial Phylogenetic Reconstruction from Whole Genomes Is Robust to Recombination but Demographic Inference Is Not *mBio*, **5(6)**: e02158-14.

[28] Kimura, M. (1969) The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations *Genetics*, **61**: 893–903.438.

[29] Didelot, X. Lawson, D, Falush, D. (2009) SimMLST: simulation of multi-locus sequence typing data under a neutral model *Bioinformatics*, **25(11)**: 1442–1444.

[30] Maiden, M. et al. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms *Proc.Nat. Acad. Sci. USA*, **95**(6): 3140–3145.

[31] Didelot, X. Falush, D. (2007) Inference of Bacterial Microevolution using Multilocus Sequence Data *Genetics*, **175(3)**: 1251–1266.

[32] Didelot, X. Wilson, D. (2015) ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes *PLoS Computation Biology*, **11(2)**: e1004041.

[33] Croucher, N. et al. (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins *Nucleic Acids Research*, **43(5)**: e15.

[34] Gansner, D. Koutsofios, E. North, S. Vo, K. (1993) A technique for drawing directed graphs *IEEE Transactions on Software Engineering*, **19**: 214–230.

[35] Robinson, D. Foulds, L. (1981) Comparison of Phylogenetic Trees *Mathematical Biosciences*, **53**: 131–147.

[36] Schierup, M. Hein, J. (2000) Consequences of recombination on traditional phylogenetic analysis *Genetics*, **156**(2): 879–891.

[37] Didelot, X. Lawson, D. Darling, A. Falush, D. (2010) Inference of Homologous Recombination in Bacteria Using Whole-Genome Sequences *Genetics*, **186**: 1435–1449.

[38] Paradis, E. Claude, J. Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language *Bioinformatics*, **20**: 289–290.

[39] Schliep, K. (2011) phangorn: phylogenetic analysis in R *Bioinformatics*, **27**(4): 592–593.

[40] Sukumaran, J. Holder, M. (2010) DendroPy: a Python library for phylogenetic computing *Bioinformatics*, **26**(12): 1569–1571.

[41] Cock, P. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics *Bioinformatics*, **25**(11): 1422–1423.