Birth-Death Process with Recombination

Subtitle

Candidate Number: 123331

This dissertation is presented as part of
the requirement for the degree of
Master of Mathematics in Mathematics and Statistics

Department of Statistics
University of Oxford
United Kingdom
20 March, 2017

Abstract

Recombination is a well-studied topic in genetics. While it has been incorporated into a backwards-in-time Coalescent model, the feature has never been added to a forward-in-time Birth-death model. This forward-in-time model offers intuitive interpretation of the event of the recombination. This project aims to, through simulation, discover some of the differences between the genetic trees obtained by the two models. This is proceeded through comparison of calculation of differences in trees and some of the statistics from simulations of each of the models under different parameters. Through the simulations, it is discovered that even though the tree topologies generated by the two models are the same, the branch lengths of the trees are expected to be different. The role of importance sampling is also investigated, providing a method to analyse trees under the birth-death model with reference to recent advancements of the method to the coalescent model.

Contents

Part I
Introduction

Phylogenetic trees show different features of the population of living organisms. By studying the DNA and genetic materials, one can construct trees on speciations which is based on the similarity of organisms. Figure 1 shows an example of a phylogenetic tree on the evolution of species of organisms.

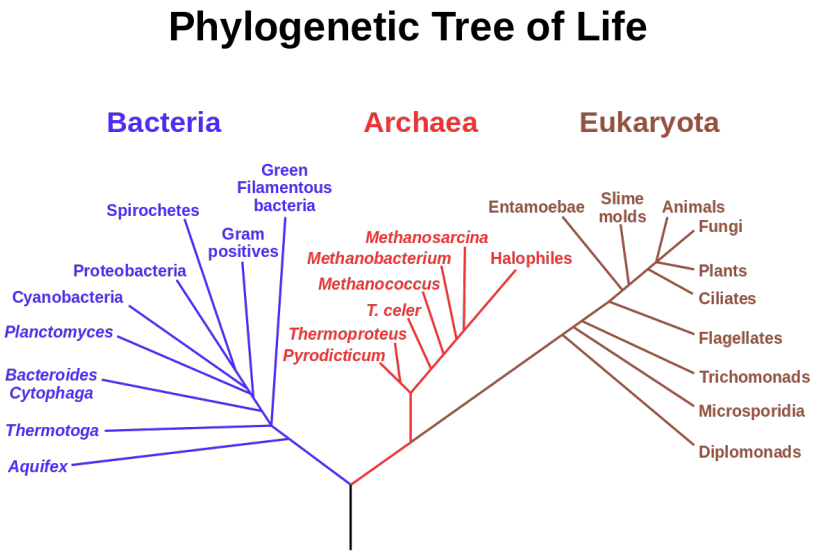

Figure 1: A Phylogenetic Tree showing speciation. (Photo Source: Wikipedia)

Phylogenetic trees are also effective in tracing genetic material and modelling behaviour of organisms, such as HIV. For these, computer simulations are performed to model actual growth and is able to provide information on how different strands of the DNA in the virus can interact with one another.

In the paper we shall refer to haploid trees, where the genetic material every daughter offspring of any event comes only from one parent. In further context regarding recombination in the future, we will restrict the trees to such that each part of the genetic material of daughter offspring must come only from one individual. This can be used to model cell divisions or viral growths. Whereas for individuals with diploid chromosomes (like humans), most genes will be present on two homologous chromosomes (each come from one biological parent) and hence will not be applicable to the models discussed.

The term tree is used in describing such a process since the whole process is equivalent to a directed graph which goes only one way, either forward or backward in time for the whole graph. The forward-in-time tree describes the process forward in time, providing a readily interpretable process. A backward-in-time tree is useful in illustrating mathematical concepts since all genetic material that is concerned is known. Figure 1 shows a reconstructed tree showing the speciation events of large kingdoms of species.

By simulating trees one is able to model events in the future. The trees can provide information such as estimating population sizes, times when individuals find a common ancestor, probability of occurrence of genes in a population.

Recombination has always been a topic of discussion among biologists as it appears in many populations and is vital to the spreading of genetic materials in a population. It refers to an event in the history of the phylogenetic tree in which as a result of the event the offspring has a different genetic material composition to their parent. The most common type of recombination is the crossing over recombination, also known as the reciprocal recombination. This is also the main focus of the paper.

Crossing-over recombination occurs during the replication of DNA during asexual reproduction or during meiosis in sexual reproduction. When two chromosomes are close enough in distance, genetic materials can be exchanged between the pair, resulting in chromosomes bearing parts of the chromosome from one parent and other parts from another parent. This can be illustrated by Figure 2.

Part II
Background

Birth-death models are used extensively in probability theory as the Markov Chain comparison with its birth and death rates is a naturally interpretable forward in time model. It models the birth and death probabilities by $Exp(\lambda)$ and $Exp(\mu)$ distributions respectively. The memoryless property and the Markov Chain property of the simple birth-death model have incurred various phenomena. [1] However, thus far there has not been birth-death models which are able to incorporate the effect of recombination.

Coalescent processes based on the Wright-Fisher model and also follows an exponential distribution. [2] Ways have been devised to include elements

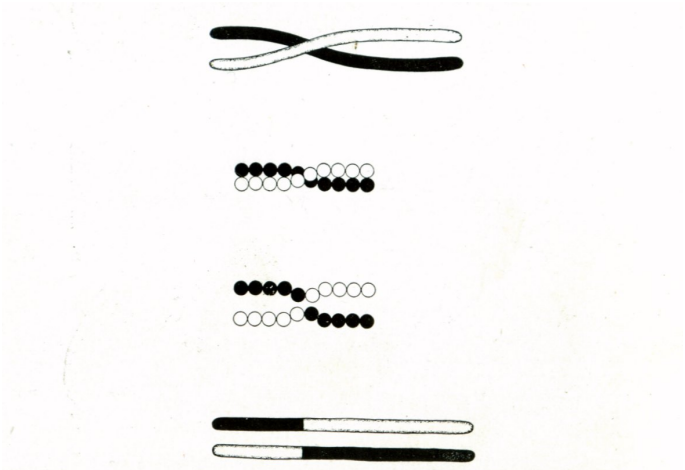

FIG. 64. Scheme to illustrate a method of crossing over of the chromosomes.

Figure 2: A recombination event. (Photo Source: Thomas Hunt Morgan's illustration in 1916)

such as mutation and recombination. When compared with a simple birth-death model, genealogical trees often share the same tree topology, but based on different sampling methods, differences may occur on the length of the stems and hence the distribution of the age of trees. [3]

Since the birth-death model is more intuitive for understanding, and that the incorporation of recombination has not been done to an extensive extent, the focus will be to seek a mathematical model which can accurately describe the effect of recombination. In our model assumptions, we assume that different events follow an exponential distribution.

We differentiate the two different models based on their nature of being forward-in-time (birth-death model) and backward-in-time (Coalescent). In particular, the Most Recent Common Ancestor (MRCA) in the coalescent model is determined from the process, while we simulate a birth-death tree from (an) individual(s). In particular, the process can die out when all surviving individuals have died out before the next birth event.

The realisations of the models are done via sampling Exponential jump times of events. This is done as an equivalent to simulating a Markov Chain.

The strong assumption that every generation has the same length in the coalescent model violates real-life situation, but nonetheless provides a good model for interpretation. The time-scaling in the coalescent model provide

a comparison with the Birth-Death model. The Birth-Death Process is very intuitive on what different events correspond to in real life. For instance, branching refers to the birth of an individual while a discontinued branch refers to the death of the individual. However, in the case of Coalescent, branching refers to a parent being chosen by more than one daughters, and a discontinued branch is non-existent in the sense that it occurs only at observation (without the information of the future).

Assumptions We assume that all processes and events follow Exponential Distribution with rates as stated in each case. In the Coalescent, all events are scaled according to the length of generation, instead of real time, and it also follows the assumption of constant population size in each generation. There have been theories that show that real populations converge to the coalescent under certain conditions [4], which will not be explored here. Birth-Death models follow exponential distributions on the memoryless property after each event.

1 Recombination

Figure 3 shows an example of a birth-death tree with recombination, where there is a recombination between the birth of the third individual and the fourth, whereby the different parts of the gene follow different paths in the cross, resulting in two different marginal trees when observing different genes. This is called the reciprocal recombination. Another case of recombination is called 'unidirectional' where instead of switching parts of the gene between two individuals, one individual remains the same as before the event, and the other individual has parts of the gene replaced by the segment corresponding to the same part of the first individual. This type of recombination is not studied in the simulations but the codes can easily be adapted to this case.

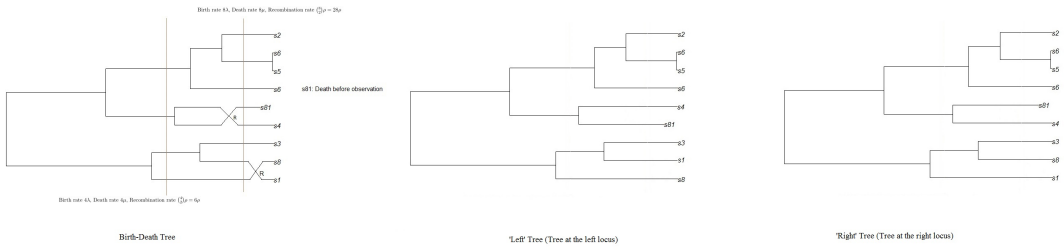

Figure 3: A Birth-Death Tree showing Recombination.

2 Statistical Tests on testing significance

In the following sections comparisons are made between models. Different aspects of the models are compared. The statistics of different models such as the time-to-coalescent showing the lengths of the time for two random individuals to coalesce are compared. This provides a comparison on the tree structure and how it behaves in the birth-death model and coalescent model.

The expected branches of a node show whether the tree topology is the same. For instance, whether the branches form or coalesce in a similar fashion. In this comparison, the time taken for branch forming or coalescing is not important.

Different ways to compare the statistics were attempted. Firstly, by computing the distribution functions and performing comparisons based on them was the most direct way of comparison. However, as noted in various papers by Stadler ([5],[3]), the calculations, even though only consists of exponential functions, often are very complicated and does not simplify easily. This is the case for pure Birth-Death models (Birth-death model with only birth and death, without mutations and recombinations). For this reason the route is not pursued for the Birth-death with recombination model. Similar is the case for the coalescent model.

Secondly, the method of estimating the empirical distribution through simulation and drawing samples and then performing the comparison was attempted. This is a non-parametric method in which we obtain the statistics from both models and perform a comparison between them. The Wilcoxon Paired Rank test [6] is applied throughout the simulation. The reason why the normal Rank test is not applied is because throughout the simulations, the coalescent model starts with the final number of individuals from the birth-death process and moves backwards in time. Hence there is a correlation between the times to coalescent of the two simulations.

For the two samples $X_{1,i}$ and $X_{2,i}$, the formulation of the rank test involves giving ranks for $R_i = |X_{1,i} - X_{2,i}|$ and summing the ranks for those positive $X_{1,i} - X_{2,i}$ where we label as N_r.

$$\sum_{i=1}^{N_r} R_i$$

is the mathematical way of expressing this statistic. The R function wilcox.test performs this test though takes the $p-$value from a Normal Approximation

which is a good estimation when the number of samples is large (which makes calculation of the exact statistic somewhat difficult).

The null hypothesis (H_0) that the data are drawn from the same distribution implies that the difference of the two has a symmetric distribution about 0, thereby allowing us to apply the signed rank test.

In the statistical tests in the simulations, $X_{1,i}$ would be the statistic from a birth-death tree under some parameters and $X_{2,i}$ would be the statistic from the coalescent tree with the corresponding parameters. As we aim to test whether the distributions are the same, the Wilcoxon Paired Rank test would be applicable.

3 Models

3.1 Kingman's n-Coalescent (Wright-Fisher Model)

The Coalescent is a backwards-in-time continuous time Markov Chain model [2]. The Coalescent model is motivated from the uniform probability of choosing a parent in each generation. Two individuals in the generation 'coalesce' mean that they have the same parent in the previous generation. In particular, each two individuals have a uniform probability of coalescing, and hence the exponential rate (which is inversely proportional to the probability), is $\binom{n}{2}$, upon scaling with the size of the population (which is assumed to be infinity).

In order to model real life biological examples, mutations and recombinations are added to the model. Each of the two events each have equal probability of happening to each individual. Hence by setting the rate of mutation and recombination to each individual during its lifetime as θ and ρ respectively, we obtain the rate at n individuals as $n\theta$ and $n\rho$ respectively.

The major assumption of the coalescent is that the events are actually scaled with the number of individuals and generation length. This meant that for comparison purposes, an adjustment has to be made to the rates and the population size for it to demonstrate characteristics observed in the birth-death model.

The representation of recombination in the coalescent model is different form that of the birth-death model. The Figure 4 shows four recombination events in a representation called the Ancestral Recombination Graph (ARG). Instead of crossing branches, we are only interested in the lineages of the genes of the individuals sampled. This implies that, different to the

Birth-Death model above, the second individual from the recombination event is not present in the sample at the time of sampling. By tracing different locations along the gene, we obtain different marginal trees (meaning the tree when looking at a specified locus). In particular, when we focus on the right part of the tree, we realise that tracing at any locus will give the same marginal tree for that part.

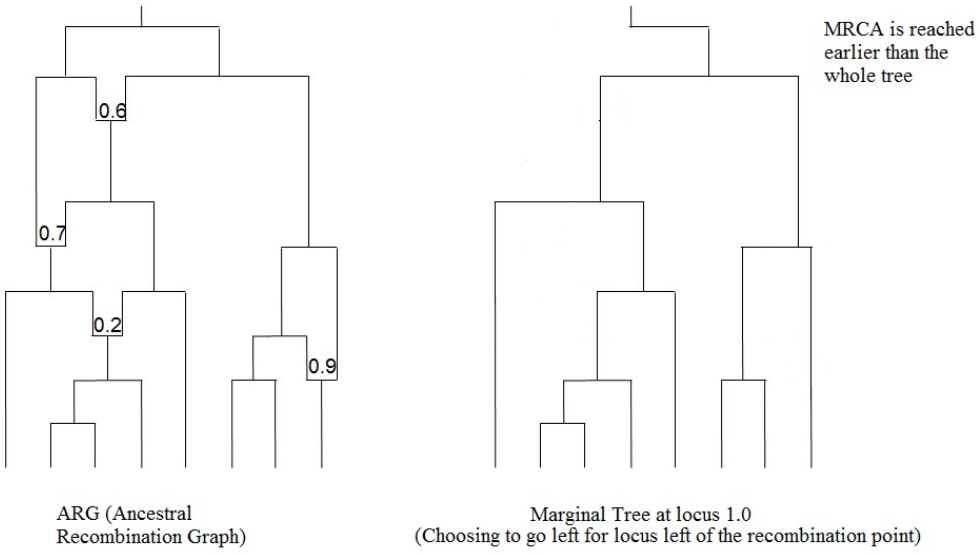

Figure 4: A Coalescent Tree showing Recombination. (Picture source: Simon Myers in SC1 Stochastic Models in Mathematical Genetics course notes, 2016; Marginal tree: Self-made)

Formally the coalescent model is defined as follows. For any natural number n, the coalescent starts with a partition (disjoint subsets whose union form the whole set) of $1, 2, \cdots, n$ and an equivalence relation R such that the equivalence classes of R_0 (at the initial state) consists only of the singleton sets.

Defining the transition rates $q_{\psi\eta} = \lim_{h\downarrow 0} h^{-1}\mathbb{P}(R_{t+h} = \eta | R_t = \psi)$ (where $\eta \neq \psi$ are equivalence relations) are given by $q_{\psi\eta} = 1$ if $\psi \prec \eta$ where $\psi \prec \eta$ means that η is the combining two equivalence classes of ψ. Since there are k equivalence classes in the k^{th} step, the total jump rate of coalescence events is $\binom{k}{2}$.

Every two individuals have a rate of 1 of coalescing and every individual has a rate of ρ of recombining, both independent of any other event. We can model each event as a clock which ticks at exponential rates as modelled above. After each event, all the clocks get reset due to the memoryless property. The jump rates are hence only dependent on the current number of

individuals as it changes the number of competing exponential clocks, hence the actual rate of jumps. This concludes the formulation of the Markov Property of the jump chain.

The effect of mutation is independent of such jumps, happen at rate $\frac{\theta}{2}$. A mutation chooses a certain part of the genetic sequence and changes it to make the genetic sequence different from before the event. The tree resulted from a mutation event will show no difference, and mutation events are essential to inferring about events happening to the tree prior to the observation. When viewing individuals' genetic sequence as an infinite real line (in the infinite sites model), a mutation event corresponds to randomly choosing a branch at random and adding a mark at a random position at the real line. The mark will remain on the individual as well as all future descendants of the individual at that location.

The effect of recombination is slightly different. It happens at rate $\frac{\rho}{2}$ for each individual at any time. At a recombination event, the individual's genetic sequence splits into two, one part coming from a parent and the other part from another parent. In the notions of the coalescent, it corresponds to adding a new member to the set, where the new member forms a singleton set of equivalence class in the next step.

3.2 The Birth-Death Process

The Birth-Death model with events birth, death and recombination has been used to model various biological events. [7] A Birth-Death model is assumed to be the counting process of the number of individuals in the population. Since each individual, at any point in its life, has exponential λ and μ rates of giving birth and dying respectively (memoryless property of the exponential distribution), we can calculate the rates of transition dependent on the current number of individuals. In particular, we can appeal to the definition of Markov Chains to argue that each event transitions is only dependent on the current one. We can appeal to competing exponential clocks to argue that the instantaneous rates of transitions from $n > 0$ individuals to $n + 1$ (birth) and $n - 1$ (death) follow exponential rates $n\lambda$ and $n\mu$ respectively.

We can model the process with leaves and branches of a tree. At time $t = 0$, we start with a single leaf. At a birth event, a leaf splits into two (identical) leaves. At a death event, a leaf dies and is removed from the tree. In the reconstructed birth-death process, if a death event happens before the observation, the dead leaf will not be observed in the data. These two events are the only events which will change the number of leaves. Independently, the following two events may occur. At a mutation event, a leaf is changed

to a different leaf. Visually it can be denoted by marking a leaf, which will, through birth events, duplicate and appear in the observations. At a (reciprocal) recombination event, two leaves are randomly selected. Then the two leaves exchange parts of the leaves. The different events can be illustrated by the following Figure 5. In particular, since every leaf has equal probability in a birth, death or mutation event, the rates of each event happening is $n\lambda, n\mu, n\theta$ respectively (where λ, μ, θ respectively are the rates of happening to one individual). Similarly, any two leaves have the same rate of ρ of experiencing a recombination event, and hence at n individuals, recombination occurs at rate $\binom{n}{2}\rho$.

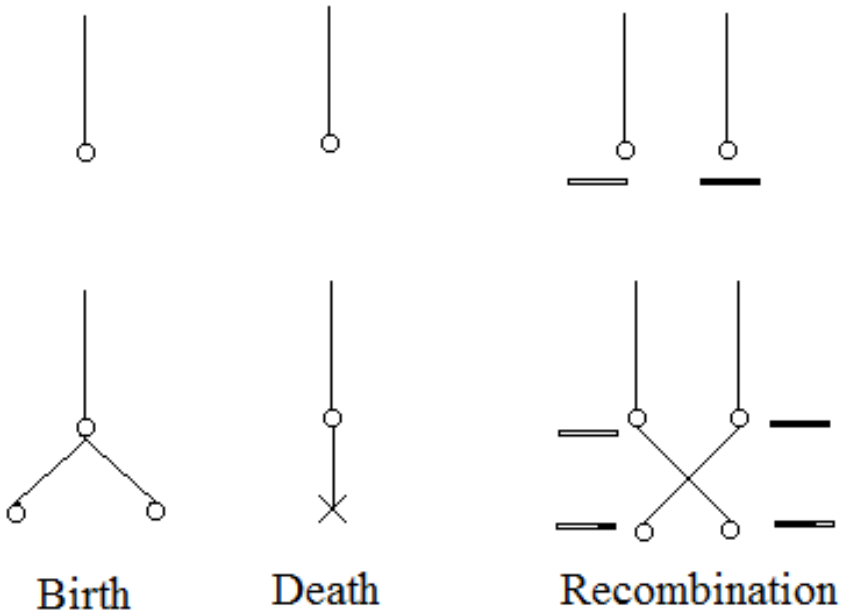

Figure 5: Illustrations of events that happen in the birth-death model.

Formally, in the state-space of $\mathbb{N}^{\geq 0}$ of the birth-death model, the Q-matrix for the Continuous Time Markov Chain is as follows:

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \mu & -(\lambda+\mu) & \lambda & 0 & \cdots & 0 & 0 & 0 \\ 0 & 2\mu & -2(\lambda+\mu) & 2\lambda & \cdots & 0 & 0 & 0 \\ & & \vdots & & \ddots & & & \\ 0 & 0 & \cdots & 0 & n\mu & -n(\lambda+\mu) & n\lambda & 0 \\ & & \vdots & & & & \ddots & \end{bmatrix}$$

Independently mutations happen at rate $n\theta$ and recombinations at $\binom{n}{2}$ when there are n individuals.

With reference to Stadler's 2009 paper [3], 50 Pure Birth-Death trees have been sampled using the Birth-and-Death Sampling-ρ to verify the distribution of coalescent times of two individuals ("leaves on the trees"). The samples as well as results are shown in Appendix B, where the matrices show each tree. A chi-squared test was used to test the goodness-of-fit of the data with Stadler's model. A p-value of 0.17 is recorded, which means we accept that Stadler's model is valid.

However, Stadler's findings were largely only confined to pure birth and death models of specified rates of birth and death. Nonetheless since the calculations of the integrals to obtain the density function would be very complicated for general cases, it is also worth performing sampling and obtain empirical distributions instead of attempting to obtain the actual density.

In order to model real life biological events more accurately, we add features to the birth-death model in a similar fashion to the coalescent model. For instance, we propose that mutations happen to each individual at any point in time with rate θ, which, by a similar argument as above, give the rate of mutation at n individuals as $n\theta$. Recombination happens at rate ρ between any two individuals at any time. Hence, when there are n individuals, the rate of recombination is $\binom{n}{2}\rho$.

In [1], the 'pull of the present' and 'push of the past' events were introduced. They are unexpected events based on the exponential distribution model. This means that near the root of the birth-death tree (the start) the speed of growth in number of individuals is high, due to the fact that we sample the tree based on it having survival until present, which would mean that the first few events would be births before deaths (otherwise the tree would not be observed or sampled). The 'pull of the present' shows a burst in the population size due to the same effect towards the present.

Recombination interpreted forward in time is the aim of the project, and the fact that parents of the recombination event may not reach the time of observation means that it is difficult to trace all of the genetic material, contrasting the Coalescent model, where we trace all the genetic materials of the individuals, including those which contribute only partly (through recombination) backwards in time. This implies that other individuals descended from those are not concerned.

3.3 Effect of recombinations on the tree

In this subsection we explore the probability that recombinations do not change the tree of a phylogenetic tree under both models.

Definition 3.1. A change in the phylogenetic tree refers to whether the marginal trees at different loci are different (in branch lengths or . The different labelling of vertices is defined to be a different tree.

Definition 3.2. A change in the tree topology refers to the event that the partition of the tree vertices (in accordance to the definition of the coalescent model) are different. This would mean that branch lengths are not considered, but the vertices labels are important in distinguishing the differences.

3.3.1 Coalescent Model

For the case of the coalescent model, there are two cases for each recombination that the tree is not changed. This is either that (Case 1) the two immediate ancestors of the individual at the recombination event coalesce with each other, or that (Case 2) one coalesces with another branch immediately (it being the next coalescent event) after the second coalesces with the same branch. The following calculations will allow us to calculate the probability of the event that recombination does not change the tree for the coalescent model.

First we consider the case where there is only one recombination event.

Let $s\rho$ be the recombination rate which is linear with the current population size s and the coalescent rate is $\binom{s}{2}$. Assume that we start with n individuals. We proceed by a counting argument, on the number of events satisfying the descriptions of the cases.

We condition on the event that the recombination happens when there is m individuals. The total number of events satisfying this is $[\binom{n}{2}\ \binom{n-1}{2}\ ...\ \binom{m+1}{2}]\ m\ [\binom{m+1}{2}\ ...\ \binom{2}{2}]$. The [...] show events of coalescent and the m is the recombination event since it can happen to any of the m individuals active at that point.

We consider the two cases where the tree does not change: Note that the two cases are mutually exclusive.

Case 1: The two ancestors from the recombination coalesce with each other. At each stage of the tree, every two lineages have equal probability of coalescing. Hence number of events satisfying case 1 with the coalescent event of interest happening when there are i individuals, is $[\binom{n}{2}\ \binom{n-1}{2}\ ...\ \binom{m+1}{2}]\ m\ [\binom{m-1}{2}\ ...\ \binom{i-1}{2}][\binom{i-1}{2}\ ...\ \binom{2}{2}]$. Where the second [...] comes from the fact that the two branches of interest is not involved in any coalescent event before the population reaches i.

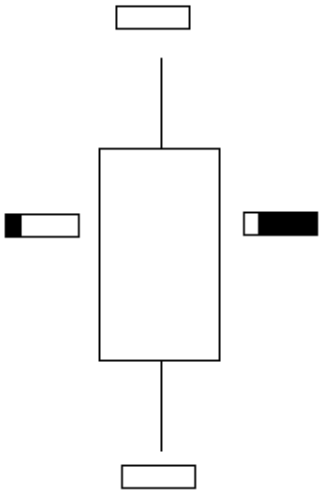

Figure 6: Case 1 in the calculations - tree change cannot be detected in this case.

Case 2: One of the two ancestors from the recombination event another branch followed immediately by the other ancestor coalescing with the same branch. The number of events satisfying this condition with the first coalescent event of interest happening at i individuals is $[\binom{n}{2}\ \binom{n-1}{2}\ ...\ \binom{m+1}{2}]$ m $[\binom{m-1}{2}\ ...\ \binom{i-1}{2}]$ $2(i-2)$ $[\binom{i-2}{2}\ ...\ \binom{2}{2}]$. The factor $2(i-2)$ originates from choosing one of the two branches and one of the remaining $(i-2)$ to coalesce followed by the definitive event of the resulting branch coalescing with the second of the two branches.

Hence summing the two cases for i between 2 and m,

P(recombination does not change the tree when the recombination event happens at m individuals) =

$$\sum_{i=2}^{m} \frac{\binom{n}{2}\binom{n-1}{2}\cdots\binom{m+1}{2}m\binom{m-1}{2}\cdots\binom{i-1}{2}\binom{i-1}{2}\cdots\binom{2}{2}}{+\binom{n}{2}\binom{n-1}{2}\cdots\binom{m+1}{2}m\binom{m-1}{2}\cdots\binom{i-1}{2}2(i-2)\binom{i-2}{2}\cdots\binom{2}{2}}$$

$$\frac{}{\binom{n}{2}\cdots\binom{m+1}{2}m\binom{m+1}{2}\cdots\binom{2}{2}}$$

$$= \sum_{i=2}^{m} \frac{2(i-2)+\binom{i-1}{2}}{\binom{m+1}{2}\binom{m}{2}}$$

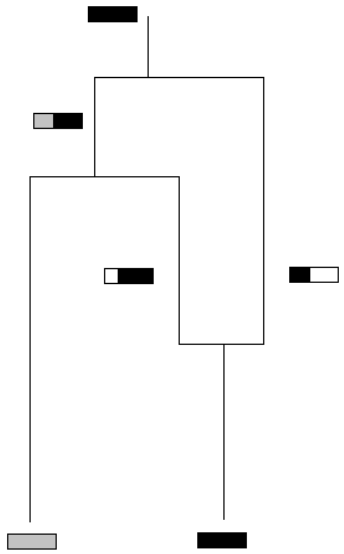

Figure 7: Case 2 in the calculations - tree change can potentially be detected, but the tree topology remains the same

$$= \frac{\frac{m(m+1)(m+2)}{6} - 3m + 2}{\binom{m+1}{2}\binom{m}{2}}$$

P(coalescent as the next event with j individuals) $= \frac{\binom{j}{2}}{\binom{j}{2}+j\rho} = \frac{j-1}{j-1+\rho}$ (by competing exponentials) Similarly, P(recombination as the next event with j individuals) $= \frac{\rho}{j-1+\rho}$.

P(1 recombination happening at m individuals) $= \frac{n-1}{n-1+\rho}\frac{n-2}{n-2+\rho}\cdots\frac{m}{m+\rho}\frac{\rho}{m-1+\rho}\frac{m}{m+\rho}\frac{m-1}{m-1+\rho}\cdots\frac{1}{1+\rho}$

Hence total probability =

$$\sum_{m=2}^{n} \frac{n-1}{n-1+\rho}\frac{n-2}{n-2+\rho}\cdots\frac{m}{m+\rho}\frac{\rho}{m-1+\rho}\frac{m}{m+\rho}\frac{m-1}{m-1+\rho}\cdots\frac{1}{1+\rho}\frac{\frac{m(m+1)(m+2)}{6} - 3m + 2}{\binom{m+1}{2}\binom{m}{2}}$$

We notice that each factor is small. The probability for each m is of order $\mathcal{O}(1/m)$. Hence the total probability is of $\mathcal{O}(1/m^2)$ (leading term $< \frac{2}{3m^2}$). Since $\sum_{m=2}^{\infty} \frac{1}{m^2} = \frac{\pi^2}{6} - 1$, as we take $n \to \infty$, the probability is slightly less than $\frac{1}{2}$ (Assuming that ρ is much smaller than n, or else the probability is even smaller due to the multipliers at the front). The probability of no change in the tree when there are more recombination events is even smaller. Hence in this respect, the birth-death model and coalescent model is slightly different.

In the calculations the probability of one recombination happening was included in the calculation in the fractional terms $\frac{m}{m+\rho}$. If we condition on there being only one recombination, the probability of no change in tree (Case 1) is

$$\sum_{m=2}^{n} \frac{\frac{m(m+1)}{2} - 4m}{\binom{m+1}{2}\binom{m}{2}}$$

and no change in topology (Case 2) is

$$\sum_{m=2}^{n} \frac{\frac{m(m+1)(m+2)}{6} - \frac{3m(m+1)}{4} + m}{\binom{m+1}{2}\binom{m}{2}}$$

where we end up with the same expression as in [8].

Similar calculations are done for when there are 2 recombination events. The expression is very long, but based on the calculations where we partition the possible events into: 1. Where the first recombination does not change the topology and both branches coalesce before the second recombination; 2. Where the branches from first recombination have not coalesced before the second recombination event.

Case 1. simplifies into the convolution between two single recombination events - hence is of order $\mathcal{O}(\text{Probability for one recombination event})^2$. Case 2. is a double sum of terms each is also of order (each term of the summation in the expression for one recombination event)^2, which is less than (Probability for one recombination event)^2 by the Cauchy-Schwarz Inequality. [* The expression is too long to fit into the margins.]

3.3.2 Reconstructed Birth-Death Process

We illustrate difference in tree topology in the birth-death model with a figure. Figure 8 shows a birth-death tree with recombination which shows that the two marginal trees must be different. Since the dead branch and the side branch (the two recombined branches) show different topology prior to the recombination event, it can never be the same, except when the recombination event is the last event before observation, where the topology would be the same up to the labelling of vertices observed.

For the case of a reconstructed Birth-Death process, there are branches that may not be observed, and hence the topology can be the same for both cases. For instance, with the Death of the branch in Figure 8, the left and right tree are indeed the same in topology, though the branch length is different due to the time of birth of the branch being different in the two trees.

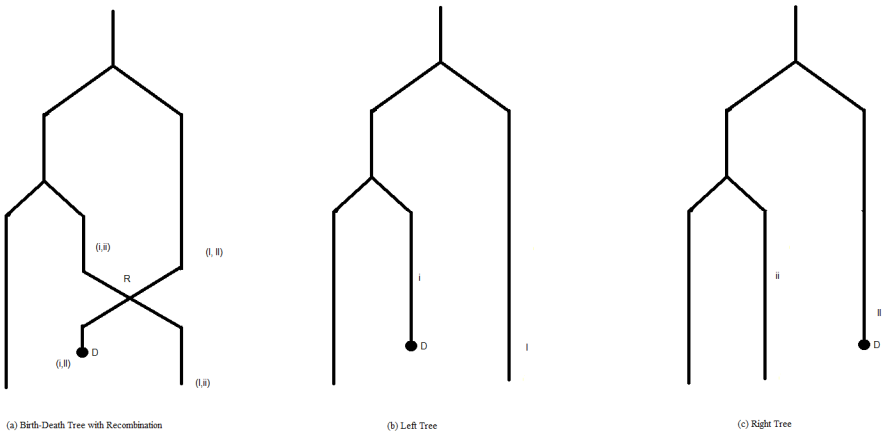

Figure 8: A Birth-Death Tree with Recombination showing a difference in tree topology

In the following we will calculate the probability of the left and right tree being the same in the reconstructed birth-death process conditioning on the process having strictly one recombination event, and that the observation to be very distant in time (so that we can apply convergence results to survival of branches).

We will appeal to the strong Markov Property of the Birth-Death Process and consider the tree before and after the recombination event separately. $\mathbb{P}$(No change in tree topology/tree lengths)=$\mathbb{P}$(No change before the recombination event)$\mathbb{P}$(No change after the recombination event)[Cases 1a,2]+$\mathbb{P}$(Changes before the recombination event)$\mathbb{P}$(Changes are nullified after the event, that is, the extra branches all experience death before observation [Cases 1b,3]) We will also appeal to the memoryless property and independence of branches in the birth-death model to separate out one branch form the tree and independently consider events on the branch. We assume that the time between the recombination event and the observation is ∞ in order to obtain expressions for survival of branches, as calculated later in Theorem 1 (P. 22).

We consider the trees before and after the recombination event. If the tree before the recombination event has to be the same, as with the same argument as the coalescent model, the two cases apply - that the two branches coalesce with each other (case 1) or one of the two branches coalesce with another branch followed immediately by the coalescent with the second (case 2). In the wordings of the birth-death model, a coalescent is a birth, or births and deaths such that at the time of interest (the recombination event), there

is only one surviving branch at the two coalescing branches. The following figure 9 show the cases.

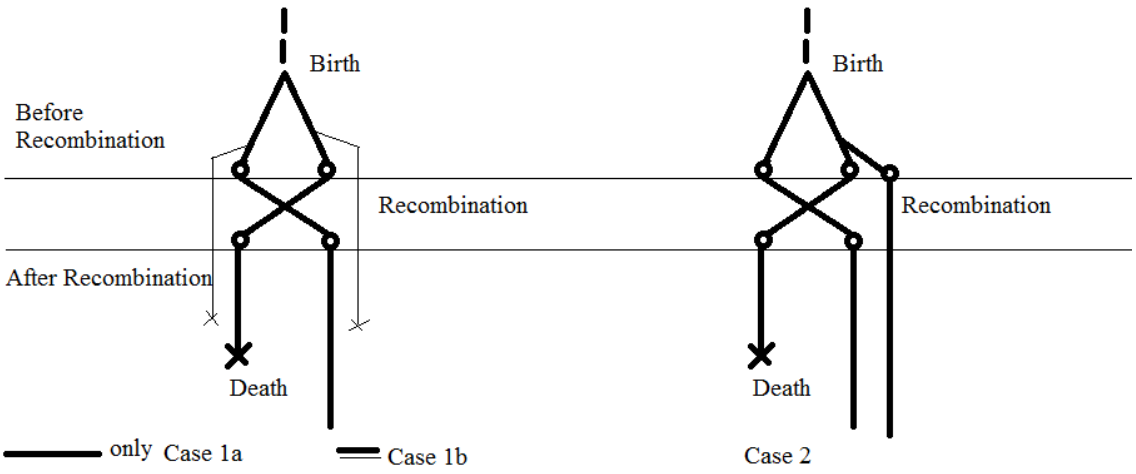

Figure 9: Figure showing cases causing no change in the tree topology or lengths

Case 1a: $\mathbb{P}$(No change before the recombination event)=$\mathbb{P}$(birth observed, within the two branches emerging from the birth, each branch only has one branch surviving into the recombination event). Hence, at any point in the tree, we require that a birth event happens (before a death), which has probability $\frac{\lambda}{\lambda+\mu}$. We then isolate the two branches from this birth event. Then $\mathbb{P}$(one branch only has one branch surviving into the recombination event), the number of births are always larger than or equal to the number of deaths, with equality at the recombination event. This has probability

$$P = \sum_{m=0}^{\infty} \left(\frac{\lambda}{\lambda+\mu}\right)^m \left(\frac{\mu}{\lambda+\mu}\right)^m - \sum_{m=0}^{\infty} \left(\frac{\lambda}{\lambda+\mu}\right)^m \left(\frac{\mu}{\lambda+\mu}\right)^{m+1}$$

$$= \frac{\lambda(\lambda+\mu)}{(\lambda+\mu)^2 - \lambda\mu}$$

The second sum term is the such that the $m+1$ death events happen before all m births. The terms count the births and deaths.

Then we consider $\mathbb{P}$(No change after the recombination event). If both branches from the recombination survive to the present, the event will be detected since the two branches evolve differently. Hence we require that at least one of the two branches not survive until the present observation, which has probability $2\left(\frac{\mu}{\lambda}\right) - \left(\frac{\mu}{\lambda}\right)^2$.

Hence the probability of case $1a$ (no change in tree lengths and topology) is

$$\frac{\lambda}{\lambda+\mu}\Big(\frac{\lambda(\lambda+\mu)}{(\lambda+\mu)^2-\lambda\mu}\Big)^2\left(2\Big(\frac{\mu}{\lambda}\Big)-\Big(\frac{\mu}{\lambda}\Big)^2\right)$$

.

Case $1b$: Extra branches (changes to the topology) all die before observation. Then instead of event P, we have P_n where there are $n+1$ branches surviving into the recombination event (instead of just 1) and then all the n branches die before the observation.

$$P_n = \sum_{m=0}^{\infty}\Big(\frac{\lambda}{(\lambda+\mu}\Big)^{m+n}\Big(\frac{\mu}{\lambda+\mu}\Big)^{m} - \sum_{m=0}^{\infty}\Big(\frac{\lambda}{\lambda+\mu}\Big)^{m+n}\Big(\frac{\mu}{\lambda+\mu}\Big)^{m+n+1}$$

$$= \big(\frac{\lambda}{\lambda+\mu}\big)^{n}\Big(1-\big(\frac{\mu}{\lambda+\mu}\big)^{n+1}\Big)\frac{1}{1-\frac{\lambda\mu}{(\lambda+\mu)^2}}$$

Then for event P_n, we require that after the recombination event, all n branches die out, with probability $\left(\frac{\mu}{\lambda}\right)^n$. At the recombination event, the $n+1$ branches provide $n+1$ choices to select the branch to recombine.

Hence the probability of case $1b$ is (no closed form)

$$\sum_{m,n=0}^{\infty}\frac{\lambda}{\lambda+\mu}\big(\frac{\lambda}{\lambda+\mu}\big)^{m}\Big(1-\big(\frac{\mu}{\lambda+\mu}\big)^{m+1}\Big)\frac{1}{1-\frac{\lambda\mu}{(\lambda+\mu)^2}}$$

$$\big(\frac{\lambda}{\lambda+\mu}\big)^{n}\Big(1-\big(\frac{\mu}{\lambda+\mu}\big)^{n+1}\Big)\frac{1}{1-\frac{\lambda\mu}{(\lambda+\mu)^2}}\big(\frac{\mu}{\lambda}\big)^{m+n}(m+1)(n+1)\left(2\Big(\frac{\mu}{\lambda}\Big)-\Big(\frac{\mu}{\lambda}\Big)^2\right)$$

Case $1b$ is likely to be small because of the high powers of fractional terms. Even if $\mu=\lambda$, the $\frac{\lambda}{\lambda+\mu}$ will tend to 0 very quickly. We require the births to happen rapidly between the birth and the recombination and deaths to happen rapidly after the recombination before the observation, which will have a very small probability compared with case $1a$ and case 2 to follow.

Case 2: We argue with a similar argument to case $1a$. We require that a birth happens followed by $\frac{\lambda}{\lambda+\mu}P$ (the probability that there is a birth event and then births and deaths happen as if the P event).

Then we choose one of the two branches from the second birth for the 2 factor. And then for each of the two recombining branches, we experience event P.

We also require that the branch not chosen to survive until the present (with any number of individuals), such that it is different from case $1b$. This, according to earlier calculations, has probability $1-\frac{\mu}{\lambda}$ (probability of non-extinction of the branch).

And then we also require that one of the two branches die out, which has probability $2\left(\frac{\mu}{\lambda}\right) - \left(\frac{\mu}{\lambda}\right)^2$.

Hence the probability is

$$2\frac{\lambda}{\lambda+\mu}\frac{\lambda}{\lambda+\mu}\frac{\lambda(\lambda+\mu)}{(\lambda+\mu)^2-\lambda\mu}\frac{\lambda(\lambda+\mu)}{(\lambda+\mu)^2-\lambda\mu}\frac{\lambda(\lambda+\mu)}{(\lambda+\mu)^2-\lambda\mu}\left(2\left(\frac{\mu}{\lambda}\right) - \left(\frac{\mu}{\lambda}\right)^2\right)$$

In particular, when we consider case 2, if there are additional branches, it is either in case 2 or in case $1b$ if all branches die out.

In the case of unidirectional recombination, the $\left(2\left(\frac{\mu}{\lambda}\right) - \left(\frac{\mu}{\lambda}\right)^2\right)$ is replaced by only one branch not surviving until the present, that is, $\frac{\mu}{\lambda}$.

There is also another completely exclusive case 3 from cases 1 and 2 above. If the two branches from the recombination dies out before observation, the reconstructed birth-death process will not show the recombination event. Notice that in the cases 1 and 2, we explicitly require that one branch to survive into the present, hence case 3 is completely exclusive of cases 1 and 2. This, from previous calculations, is of probability $\left(\frac{\mu}{\lambda}\right)^2$.

Note that case 3 is actually the most prevalent case especially when μ and λ are similar, as the probability that the recombination event happen to two branches which were the only branches surviving from the MRCA of the branches is very small in the birth-death model, where there recombination event is very likely to happen when there are many individuals.

Part III
Simulations

1 Notes and principles of simulations

In order to simulate the birth-death processes, exponential random variables are simulated, according to the inverse method (through simulating a Uniform[0,1] variable and transform using the inverse of the cumulative distribution function of the distribution). If $U \sim U[0,1]$, $X = \frac{-log(U)}{\lambda}$ follows an $Exp(\lambda)$ distribution. After that, another (independent) Uniform[0,1] variable is simulated to determine whether the event is a birth, death, mutation or recombination. Lastly, a discrete uniform variable is used to determine which individual that the event happens to.

This matrix of such information is used in analysing the tree structures. As we know from the previous part that this matrix uniquely determine the tree, hence we can work with this matrix to get different statistics. When the event is a recombination, two individuals are simulated.

Since they are generated from a discrete uniform variable, the two individuals do not present any ordering (so it is not the case that the first individual number is always larger than the second, or vice versa). Hence without loss of generality, the first individual chosen is assumed to pass on the genes on the first locus to the second individual, and the first individual is assumed to pass on the genes on the second locus onto the first individual. This is because we are considering a two loci model, which means the recombination event must happen at the location between the two loci. This is illustrated by Figure 2.

For the coalescent model, the simulation modelling is given by ms [9] which is a code in C. The library phyclust in R incorporates features from ms. Features from ms allows plotting trees from the coalescent model, incorporating features such as recombination or growing population sizes.

According to [4], a better comparison which conditions on the number of individuals surviving to the present, has a slightly different form. For a fairer comparison between the models, such a scaling on the population size can provide a better convergence and hence comparison between the two models. Hence when performing the two models the scaling is performed. The expected population under a birth-death model is $e^{(\lambda-\mu)t}$ at time t. The adjustment which is conditional on the number of individuals surviving until time t is only significant when λ and μ are similar. Under this choice of parameters, the adjustment is necessary because there is a high non-negligible probability of the population reaching 0. For example, the probability of death before birth at 1 individual is $\frac{\mu}{\lambda+\mu}$ which is close to $\frac{1}{2}$. From calculations regarding Markov Chain jump chains, the probability of reaching 0 individuals is $\frac{\mu}{\lambda}$.

Theorem 1. *For a birth-death process with birth rate λ and death rate μ (with $\mu < \lambda$). Starting with i individuals there is a probability of $(\frac{\mu}{\lambda})^i$ of the population reaching 0 eventually. As 0 is an absorbing state, the chain will not move out of the state.*

Proof. Let p_i be the probability of the chain reaching 0 individuals starting at i individuals.

Conditioning on the first jump,

$$p_i = \frac{\lambda}{\lambda+\mu} p_{i+1} + \frac{\mu}{\lambda+\mu} p_{i-1}$$

Solving the recurrence relation,

$$p_i = A + B(\frac{\mu}{\lambda})^i$$

Since $p_0 = 1$and $p_i \to 0$ as $i \to \infty$, $A = 0$ and $B = 1$. Hence $p_i = (\frac{\mu}{\lambda})^i$. □

If we relax the condition $\mu < \lambda$ we get that the population reaches 0 almost surely if the inequality does not hold.

However, the ms programme is unable to plot recombination graphs due to the lack of such function in the tree-plotting function in the ape library (which is commonly used in phylogenetic analyses). It does output two trees corresponding to the left tree and the right tree.

2 R functions and libraries created

For the purposes of implementing the simulations, different functions are used. In order to efficiently implement the simulations, the simulation tools are built into a library. In particular, since most of the simulations in the coalescent model are done in the 'ms' package, the simulation methods were mostly on the birth-death model. Similar parts regarding the coalescent model would be in 'ms'. The various functions can be found in [9].

2.1 Birth-Death Tree Simulation

The BDn function (Appendix A) simulates a Markov Chain, which has a bijection with the actual birth-death tree based on the formulation. Each line in the simulated matrix correspond to the next event to happen. On the state space of the number of alive individuals, the process is a Discrete-time Markov Chain because every jump depends only on the number of individuals (n), where the transition upwards rate is $\frac{\lambda}{\lambda+\mu+\theta+\frac{n-1}{2}\rho}$ and the transition downwards rate is similarly $\frac{\theta}{\lambda+\mu+\theta+\frac{n-1}{2}\rho}$. The rate for staying at the same state is through recombination or mutation, which accounts for the remaining rate.

From the argument, the jump chain is a Continuous-time Markov Chain. The holding time matrix has jumps up at rate $n\lambda$ and down at rate $n\mu$. These are all incorporated into the simulations.

By the virtue of competing exponentials, events happen in probabilities of the proportions of the jump rates. Hence a realisation of a $Uniform[0, 1]$ variable is used to determine the event. After such, a discrete uniform distribution is simulated to determine which individual(s) of the n (which is alive at this point) that the event happens to.

2.2 Tree-Drawing for Birth-Death Trees

Since the numbering of the branches change after every event (except mutations), tracing branches become complicated. A function findMRCA

attempts to find the location where each branch is branched off (birth). This allows the tracing of the branches.

Another way of approaching this is retrace every step from the start and adding branch lengths chronologically after every event. This is the method used in the MatrixTreeString function. It traces each branch and renames each branch at every event according to the event. During each step, it also marks the death of the branch whenever it branches or dies, since its length will no longer increase after the event. In particular, when a recombination event happens, for the right tree we switch the labels for the two branches (as we only consider the two loci model). By tracing through the whole matrix, the whole tree with its branches and vertices can be constructed. This can be made into Newick Tree format which uniquely determines a tree by its branch lengths and vertex labels. The tree is passed to the ape(within [10]) read.tree and plot commands which are able to plot the trees.

Attempts were made to mark locations where recombinations happen. However due to the plotting algorithm of functions in R, it became difficult. Particularly when we look at a later example in Figure 10 we note that the two trees are merely the renumbering of the vertices. Hence it is impossible to mark the location due to the automatic re-ordering of vertices in the plot.

2.3 Coalescent Tree Simulations

The 'ms' package was applied in R to simulate coalescent trees. 'ms' is written in C and importing it within the R environment required the 'phyclust' library [11]. 'ms' is able to simulate coalescent trees with various features such as outputting the tree structure in a format which can be plotted, population growth rate, recombinations and number of loci. Through outputting as phylogenetic tree formatted data in R, functions in 'ape' and other libraries (phytools, [10] can be applied directly to ease the extraction of the statistics of interest of the tree.

2.4 Time to Coalescent

An important feature investigated with simulation was the Time to Coalescent for branches. Since the left tree and the right tree exhibit different time to coalescent, only the left or the right tree is considered at any time. The findMRCA function is used to generate where the two branches share the same ancestor, in particular, the time where the two branches off from the same individual. These are done through tracing branches throughout the trees and for the output return the location and time where the two branches coincide.

Part IV
Model Comparison

1 Comparison of rates in the two models

In particular in the different events, the rates differ in the forward and backward processes. Especially when we look at the events, birth and death rates are inexistent in the coalescent model. However, an interpretation of the coalescent model has given probabilities of extant edges in accordance to the birth-death rates under the birth-death model (Kendall). On the other hand, coalescent is an event that happens backwards in time, while birth and death are forward in time. Since the coalescent rate is the scaling factor, the interpretation is sensible.

Stadler [4] explained the choice of population growth rate r of the coalescent to be the net birth-death rate. This is due to the expected population of the number of individuals under the coalescent is $N_0 e^{-rt}$ where t goes backwards in time.

However, in terms of actual tree topology, the effect of recombination may produce different tree topologies, as investigated in further sections. Moving backwards in time, a recombination event concerns one individual, while forwards in time, in the birth-death model, it concerns two individuals because exchanging genetic materials require the existence of two individuals. As it is a random process, the rate is quadratic, proportional to $\binom{n}{2}$. However when we scale it according to the population size, the scaling of $\lambda - \mu$ to match the coalescent rate would suffice to produce comparable models.

2 Comparison of Statistics of the two models

In order to compare the statistics of the two models, simulations were done. The statistics that are of interest include the tree topologies as well as the branches of the trees.

2.1 Branch Lengths

In this section birth-death trees and coalescent trees were simulated. Branch lengths refer to the time to coalescence of two randomly selected branches. This is Stadler's Birth-and-Death Sampling-ρ (2008). However, since different parts of the genome of the individuals may coalesce at different times, we propose a new definition of 'time to coalesce'.

Definition 2.1. Time to coalesce for two individuals is the minimal time passed before there exists an individual which, apart from mutations, contains the same (or part of) genetic materials as the two individuals.

Since we obtain different trees when we trace different loci of the genome, we can refer to the time to coalesce of two individuals in a specific loci, or in all loci (which means the most recent common ancestor of the two individuals), which might not be equal to any of the time to coalesce for any loci, especially when more than 1 recombinations occur. This is illustrated in Figure 10. Unfortunately due to the plotting algorithms within R, instead of changing the tree, the labels are changed in the illustration, which might be confusing. Here coalescent events happened between branches 2 and 4 and between branches 1 and 3 before the birth events of the 3 individuals. In the left tree, the coalescent time between vertices 1 and 2 is further back in time than the right tree, and both are not equal to the MRCA of the four individuals.

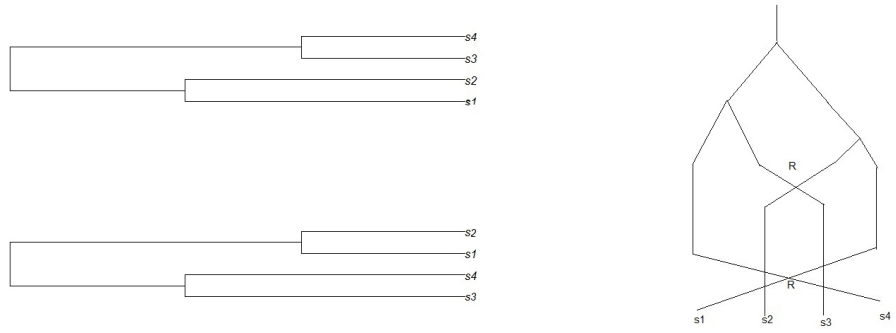

Figure 10: Birth-death tree with recombination which has different coalescent times. Top Left figure is the tree on the first locus (left tree) and the bottom left figure is the tree on the second locus (right tree). The tree on the right part of the figure is the Birth-Death tree associated with these two marginal trees.

For the purposes of comparison, the adjustment to the coalescent model is taken to be the growth rate which is $\lambda - \mu$. This applies throughout the tree, even though Stadler [4] suggests that the expected population growth is also conditional on it being able to grow quickly at the beginning (of the birth-death process) such that the population size never reaches 0. In our simulations we only output trees where the population never dies out.

In the simulations, birth-death processes of various rates were simulated. The corresponding pure birth-death processes without recombination were

used for comparison. In particular, since Stadler [3] has also calculated the exact expressions, that can be used for simulation as well. The corresponding coalescent process with the growth rate equal to $\lambda - \mu$. For comparison, the times recorded for each case were compared using non-parametric test for equal distribution.

The scaling of time the coalescent model allows comparison. While during the assumptions of the coalescent model, the time scaling is taken to be the length of each generation. The actual scaling can be derived as follows, through the comparison between the instantaneous

The following table describes the results. The data and test results are attached in the Appendix. All tables show the Wilcoxon Signed Rank statistic as well as the corresponding p-value. By symmetry we only consider the left or right tree in each case.

2.2 Expectation of branches showing mutations (Expected number of branches of a node)

The second property of recombination in Birth-Death trees is the estimation of the proportion of branches showing a mutation which happened long enough in the past. This uses the CheckBranches function. It essentially traces all branches surviving at the end of observation until the mutation event and determine the number of such branches that trace back to the location of the mutation. A function in the library phytools called getDescendants was used for the coalescent tree case. Both functions would output the number of descendants from a specified node in the tree which can then be used in the calculation for the proportion of branches from the node.

Notably it is the case that some realisations in the birth-death model give that the number of branches of a node that survives until the present observation time is 0 due to all branches from that node dying off before the observation. This is particularly prominent when the death rate is high. In the case that both birth and death rates are high, this happens more often. In order to perform a fair comparison, only the branches which have descendants at the time of observation are included in the comparisons. In the coalescent model however, since deaths are not of interest, this case never happens as at each stage, all individuals contain genetic material of the individuals of observation, which is not the case for the nodes that generate 0 in the birth-death model.

Part V
Importance Sampling

1 General Formulation and the Application to the Coalescent Model

Importance Sampling has been used in inference of distribution of statistics related to a phylogenetic tree. Both Importance Sampling and Markov-Chain Monte-Carlo (MCMC) have been used in the inference of phylogenetic trees. ([3], [12]) Monte Carlo methods are utilised in these situations because the function of the likelihood is difficult to compute. In this section we will outline the formulation of importance sampling and its application to the coalescent model to build a foundation of performing inference on the birth-death model with recombination. Felsenstein [13] and Stadler [14] showed that it is impossible to calculate even for the simple birth-death (without recombination) without further assumptions. Hence the importance sampling is motivated.

An estimator of

$$I = \int_A \phi(x)\pi(x)dx$$

is an importance sampling estimator taking the form

$$\sum_{i=1}^{n} \phi(X_i)w(X_i)$$

where X_i are realisations of proposal distribution and w is the weight function which gives the weighting of each realisation. In particular, the importance sampling is a Monte Carlo method of estimating I, where it can be the mean of $\phi(X)$ or be used to estimate other integrals of the form I.

The importance sampling scheme has to satisfy that the proposal distribution $q(x) > 0$ when the target $\pi(x) > 0$. $w(x) = \frac{\pi(x)}{q(x)}$ is the weight function. In the context of phylogenetic trees, the importance sampling is often used to infer distribution of the data, with the data based on the tree which is dependent on model parameters (for example, the rate of recombination, coalescent and mutation in the case of the coalescent model). Importance sampling estimator is an unbiased and consistent estimator and hence will converge to the real distribution when the above conditions hold, though there is no guarantee with reference to the rate of convergence [15].

In the context of coalescent with recombination, we are interested in the likelihood of observing a set of data, $L(\Phi) = \mathbb{P}(\mathcal{D}; \Phi)$, based on model parameters (rates of recombination and mutation). The likelihood of interest is

$$L(\Phi) = \int_{\mathcal{H},\mathcal{T}} \mathbb{P}(\mathcal{D}|\mathcal{H},\mathcal{T})\mathbb{P}(\mathcal{H},\mathcal{T};\Phi)d\mathcal{H}d\mathcal{T}$$

where $\mathcal{H}$ and $\mathcal{T}$ are the histories of the tree and the time of events of the tree respectively. $\mathcal{H} = (H_0, H_{-1}, \cdots, H_{-n})$ are the data collected at times $(T_0, T_0+T_{-1}, \cdots, T_0+\cdots+T_{-n})$, T_{-m} are the jump times viewed backwards in time. In the importance sampler, the likelihood

$$\hat{L}(\Phi) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{P}\big(\mathcal{D}|\mathcal{H}^{(i)},\mathcal{T}^{(i)}\big)\frac{\mathbb{P}\big(\mathcal{H}^{(i)},\mathcal{T}^{(i)};\Phi\big)}{\mathbb{Q}\big(\mathcal{H}^{(i)},\mathcal{T}^{(i)}|\mathcal{D};\Phi\big)}$$

. [16] shows the construction of the proposal distribution for

$$\mathbb{Q}(\mathcal{H}_{-k}|\mathcal{H}_{-k-1};\Phi)$$

with the event rates through sequential sampling. We shall devise the same for the birth-death model.

2 Birth-Death Model

In comparison with the coalescent model (Appendix of [16]), instead of $\mathbb{P}(H_{-k-1}|H_{-k})$ we can obtain the expressions for $\mathbb{P}(H_{-k}|H_{-k-1})$ from the formulation of the birth-death model. $\mathbb{P}(H_{-k-1}|H_{-k})$ can then be calculated via Bayes' theorem. The notations can also be found in the following Figure 11. The scheme proves too difficult to calculate - as we are going backwards on the birth-death model, and the rates are such that they are no longer independent of the event occurring.

Following the same notation, if we are able to obtain samples with a samples specified in locus A, b samples in locus B, c samples in both loci, with $a+b+c=n$, and a_i be the number of samples with allele i at A and unknown allele at B and b_j with allele j at B and unknown allele at A. Hence $\sum_i a_i = a$ and $\sum_j b_j = b$. Letting $\mathbf{a} = (a_i)$, $\mathbf{b} = (b_j)$ and $\mathbf{c} = (c_{ij})$. Denoting the data $\mathbf{n} = (\mathbf{a}, \mathbf{b}, \mathbf{c})$, and $\mathbf{n} - \mathbf{e}_i^A = (\mathbf{a} - \mathbf{e}_i^A, \mathbf{b}, \mathbf{c})$, it is possible to define the scheme for obtaining the proposal distribution. For total rate $D = \lambda n + \mu n + \theta_A(a+c) + \theta_B(b+c) + \rho\binom{n}{2}$, we have the following table in the next page. Unfortunately the backwards $\mathbb{P}(H_{-k-1}|H_{-k})$ is impossible to calculate due to the rates are no longer independent competing exponential clocks.

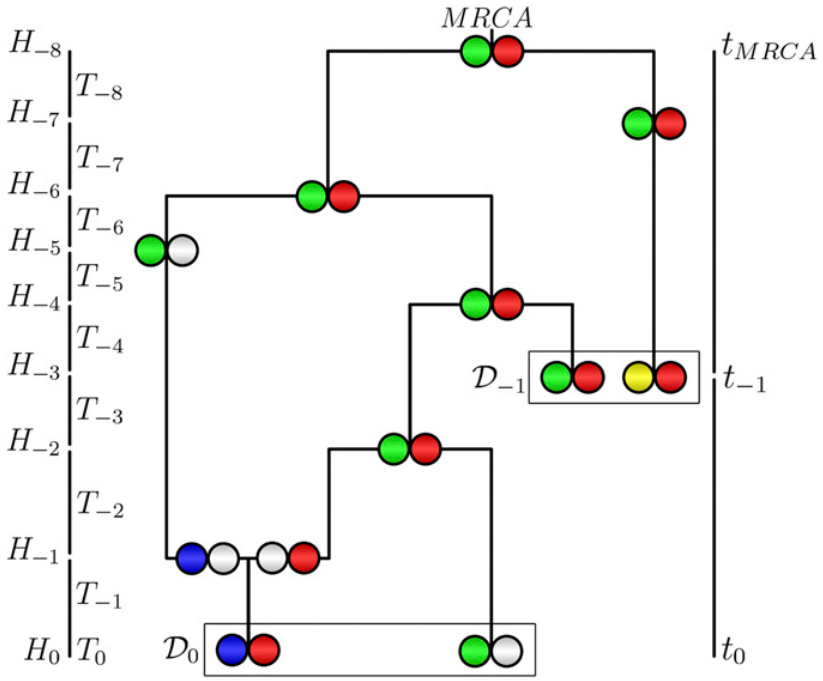

Figure 11: A coalescent tree to illustrate the notations (Picture source: [16])

Instead we proceed with the sequential sampling algorithm forwards in time, starting with one individual until the time of observation, obtaining hence the actual tree of interest.

$$\hat{L}(\Phi) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{P}\big(\mathcal{D}|\mathcal{H}^{(i)}\big)\frac{\mathbb{P}\big(\mathcal{H}^{(i)};\Phi\big)}{\mathbb{Q}\big(\mathcal{H}^{(i)}|\mathcal{D};\Phi\big)}$$

$\mathcal{D}$ would be the genetic sequence data we obtain from the observation. $\mathbb{P}(\mathcal{H})$ would be the product of the probabilities of each event (by independence from the Markov Property). Probability of events is specific - the birth event to a certain unique sequence would have probability $\frac{\lambda}{n(\lambda+\mu+\theta)+\binom{n}{2}\rho}$, whereas a death event of a sequence observed m times at the step would have probability $\frac{m\mu}{n(\lambda+\mu+\theta)+\binom{n}{2}\rho}$. $\mathbb{P}(\mathcal{D}|\mathcal{H})$ will hence be the probability of the mutations giving the intended data. When we work on the infinite sites model, this means that the mutations are at the intended position and is mutated to give the sequences in the observation. For the $\mathbb{Q}$ (proposal distribution), since the sampler will converge whenever the support for $\mathbb{Q}(\mathcal{H})$ is non-zero at the support for $\mathbb{P}(\mathcal{H})$, we can, as Griffiths and Tavare [17] use the uniform distribution of all possible trees in the previous step (in case we sample a death, we can then sample the dead branch). This is not very efficient, but guarantees convergence of the likelihood to $\mathbb{P}(\mathcal{D};\Phi)$ as desired.

Table 1: Forwards Transitions from state $H_{-k-1} = \mathbf{n}$ to H_{-k} in the 2-locus model in the Birth-Death Model

Event	H_{-k-1}	$\mathbb{P}(H_{-k-1}\|H_{-k})$	$\mathbb{P}(H_{-k}\|H_{-k-1})$
Birth of (i,j)	$\mathbf{n}+\mathbf{e}_{ij}^{C}$	-	$\frac{c_{ij}\lambda}{D}\frac{1}{\pi[(i,j)\|\mathbf{n};\Phi]}$
Birth of $(i,*)$	$\mathbf{n}+\mathbf{e}_{i}^{A}$	-	$\frac{a_{i}\lambda}{D}\frac{1}{\pi[(i,*)\|\mathbf{n};\Phi]}$
Birth of $(*,j)$	$\mathbf{n}+\mathbf{e}_{j}^{B}$	-	$\frac{b_{j}\lambda}{D}\frac{1}{\pi[(*,j)\|\mathbf{n};\Phi]}$
Death of (i,j)	$\mathbf{n}-\mathbf{e}_{ij}^{C}$	-	$\frac{c_{ij}\mu}{D}\frac{1}{\pi[(i,j)\|\mathbf{n};\Phi]}$
Death of $(i,*)$	$\mathbf{n}-\mathbf{e}_{i}^{A}$	-	$\frac{a\mu}{D}\frac{1}{\pi[(i,*)\|\mathbf{n};\Phi]}$
Death of $(*,j)$	$\mathbf{n}-\mathbf{e}_{j}^{B}$	-	$\frac{b\mu}{D}\frac{1}{\pi[(*,j)\|\mathbf{n};\Phi]}$
Mutation of $(i,j)\to(k,j)$	$\mathbf{n}-\mathbf{e}_{ij}^{C}+\mathbf{e}_{kj}^{C}$	-	$\frac{\theta_{A}p_{ki}^{A}c_{ij}}{D}\frac{1}{\pi[(i,j)\|\mathbf{n};\Phi]}$
Mutation of $(i,j)\to(i,l)$	$\mathbf{n}-\mathbf{e}_{ij}^{C}+\mathbf{e}_{il}^{C}$	-	$\frac{\theta_{A}p_{lj}^{B}c_{ij}}{D}\frac{1}{\pi[(i,j)\|\mathbf{n};\Phi]}$
Mutation of $(i,*)\to(k,*)$	$\mathbf{n}-\mathbf{e}_{i}^{A}+\mathbf{e}_{k}^{A}$	-	$\frac{\theta_{A}p_{ki}^{A}a_{i}}{D}\frac{1}{\pi[(i,*)\|\mathbf{n};\Phi]}$
Mutation of $(*,j)\to(*,l)$	$\mathbf{n}-\mathbf{e}_{j}^{B}+\mathbf{e}_{l}^{B}$	-	$\frac{\theta_{A}p_{lj}^{B}b_{j}}{D}\frac{1}{\pi[(*,j)\|\mathbf{n};\Phi]}$
Recombination of $(i,*)$ and $(*,j)$	$\mathbf{n}+\mathbf{e}_{ij}^{C}-\mathbf{e}_{i}^{A}-\mathbf{e}_{j}^{B}$	-	$\frac{\rho a_{i}b_{j}}{D}\frac{1}{\pi[(i,*),(*,j)\|\mathbf{n};\Phi]}$

Part VI
Discussion

For the simulations, different birth and death rates (which are used to determine the population growth rate in the coalescent model) as well as recombination rates were simulated. For the simulated data, in the cases where the recombination rate is 0, which correspond to the model without recombination.

1 Similarities between the models

1.1 General tree topology

From the simulations regarding the expected number of branches of a node, no matter which recombination rate we take, the null hypothesis that the two models follow the same distribution with regards to this statistic appear to be true. This is true when comparing between Birth-Death model with recombination and without recombination; coalescent model with recombination and without; comparing both birth-death and coalescent models with recombination; as well as comparing both birth-death and coalescent models without recombination. These suggest that the recombination events do not affect the tree topologies of the marginal trees. This is expected under our assumptions. In particular, this means that the case that many vertices at present trace to one branch early on in the tree is very rare.

Recombination, under the birth-death model, is either extending the branch lengths between the birth or death events for recombination events which do not have an effect on the locus, or increasing the waiting time between birth or death events with a renumbering of the active edges. While tracing the marginal trees, these two both do not affect the distribution of the expected number of branches of a node. Since the observation does not take into account the numbering of the vertices, essentially both cases, whether the recombination event switches genes on the locus, has no effect on the overall tree topology under the birth-death model.

The recombination event does not change the number of branches originating from the node. The only events that can change the number of branches is through birth or death events. With no change in the number, the statistic being same as that for a model without recombination is expected.

Whereas under the coalescent model, a similar argument holds. Even though the rate of occurrence of recombination is smaller than birth-death

process, the existence of recombination has no effect on the topology of the marginal trees, especially the one that we are tracing lineages in. The only way in which the number of branches from a node changes is only through coalescence events. As the expected number of branches of a node only regards the marginal tree at one locus, hence the renumbering does not have any effect.

Hence we can see that from the tree topological point of view, moving forwards and backwards in time does not affect the topology of marginal trees. The tree topology of the two models are similar in this sense.

Table 2: Comparison of branches from a node in the two models

Birth Rate	Death Rate	Population Growth Rate	Recombination Rate	Median (BD)	Median (C)	Rank Test p-value
1	0.5	0.5	0	0.0475	0.0522	0.1338
1	0.5	0.5	0.03	0.0405	0.0438	0.1011
1	0.5	0.5	0.1	0.0941	0.0581	0.0101
1	0.25	0.75	0	0.107	0.109	0.9091
1	0.25	0.75	0.03	0.114	0.0587	0.1967
1	0.25	0.75	0.1	0.104	0.0499	0.04335
2	0.5	1.5	0	0.0553	0.0207	0.4807
2	0.5	1.5	0.03	0.0601	0.0283	0.2203
2	0.5	1.5	0.1	0.0562	0.0379	0.2489

2 Differences between the models

2.1 Time to coalescence

Upon comparison with the times to coalescent for different models, we observe that the two models give different lengths of time to coalescent. Upon scaling time for the coalescent model for fair comparison with the birth-death model (both without recombination), as evidenced from the Wilcoxon test, we can conclude with certainty ($p < 10^{-5}$ in all cases) that the distributions of the times to coalescent are not the same. This is the case for every rate of recombination sampled as well as birth and death rates (and the corresponding population growth rate in the coalescent model).

A difference is expected due to the nature of the models. The rate of recombination is linear for the coalescent model throughout the time while the rate of coalescent decreases quadratically (highest order of the rate is 2). In the coalescent model, recombination happens relatively more frequently compared with coalescent events when the population is small. However when the population size is small, the time to wait before the next coalescent event is longer due to the low total rate of events.

Whereas for the birth-death model, even though the occurrence of recombination is low at first due to the low number of individuals. When the population starts to grow, the rate of recombination grows quadratically while the rate of birth (and death) are linear, which implies that the rate of recombination and hence the rate of all events is increased compared with the model without recombination. This implies that the occurrence of recombination is high, and since recombination only changes the labelling of the branches and does not change the population size under the birth-death model, the rates do not change before and after a recombination event. Time to coalescent is the product of bernoulli random variables with increasing success probability. The probability that the branches coalesce before the population size becomes small is high. Hence the effect of recombination is likely to affect the birth-death model more.

As in the coalescent model, the number of branches increase at the event of recombination, a longer time would be expected for the time to coalesce. However recombination happens more often near the top of the tree where there are fewer individuals. The probability that two vertices reach their MRCA before there are few individuals is high. Hence the effect of recombination is not very high in the coalescent model.

Table 3: Comparison of times to coalescent from two random vertices in the two models (500 trees of each model)

Birth Rate	Death Rate	Population Growth Rate	Recombination Rate	Median (BD)	Median (C)	Rank Test p-value
1	0.5	0.5	0	5.747896	5.909659	0.2365
1	0.5	0.5	0.05	4.593864	5.679024	2.407×10^{-13}
1	0.5	0.5	0.1	4.168371	5.818991	$< 2.2 \times 10^{-16}$ (machine precision)
1	0.25	0.75	0	5.085282	5.194262	0.4631
1	0.25	0.75	0.05	4.060859	5.138997	2.372×10^{-15}
1	0.25	0.75	0.1	3.533799	5.757791	$< 2.2 \times 10^{-16}$ (machine precision)
2	0.5	1.5	0	2.236065	2.326651	0.3273
2	0.5	1.5	0.05	2.130011	2.415385	1.951×10^{-13}
2	0.5	1.5	0.1	2.00869	2.554943	1.737×10^{-15}
2	1	1	0	2.262991	2.412973	0.1698
2	1	1	0.05	2.285671	2.58474	4.706×10^{-9}
2	1	1	0.1	2.25067	2.463243	2.203×10^{-6}

3 Summary of results

In terms of the interpretation of the birth-death model with recombination, the simulations mainly concern the reciprocal type of recombination. Other common types of recombination include unidirectional and non-homologous recombination. Unidirectional recombination is similar to the reciprocal type in the sense that two marginal trees will be formed. For one of the trees, the event corresponds to a birth and a death to another branch; for the other tree, the event does not change anything and the tree before the event is exactly identical to the tree after the event. The statistics that were looked at would be different from that of the trees generated by the reciprocal type of recombination due to the increase of number of branches of one genotype while decreasing that of another genotype, as well as pure birth-death processes, since the number of individuals actually do not change and hence even though the event corresponds to a birth and death simultaneously on one of the trees, it is different from such events in the birth-death models. Even though the differences are different to the reciprocal recombination case, the simulation codes can be easily modified to perform the equivalent simulations.

From the results of the simulations, it is apparent that the branch lengths in the birth-death model with recombination are different from that of coalescent. Since recombination events tend to happen at the end of the branches (near the current time of observation) for the birth-death model, because the rate increases quadratically with the number of individuals while other events increase linearly, it is different from the coalescent, where the rate of recombination is only comparable with the rate of coalescent at the earlier stages of the tree (near the MRCA).

4 Limitations of the Methodology

4.1 Method of Comparison

For the simulations the number of samples of the coalescent model is set at the number of individuals at the end of the birth-death model sampling after 500 events. Even though 500 events is reasonable to estimate the topologies of the tree, it may be a biased value to start the coalescent process, even though that takes care of the sampling bias for the birth-death model for the population size in the coalescent model. 500 is chosen to balance between the computation time and the limit to the statistic of interest. Since each statistic entry requires simulating one birth-death tree (which is a 501x5 matrix) and running through the whole matrix to get the statistic, and a coalescent tree (which requires inputting into tree format to do any analysis, which requires a relatively long runtime under the 'ms' package).

The method of Wilcoxon rank test is useful in testing the distribution of the difference of the two distribution and whether it is symmetric about 0. However, it is not a maximum likelihood test which is uniformly the most powerful test at a specified significant level (Neyman-Pearson Lemma, Neyman and Pearson, 1933). The fact that we are unable to find the exact distribution in closed form would imply that it is possible that our samples may not reflect the actual distribution (by sampling unlikely samples from tails). Hence it is possible to lower this probability through sampling more samples (currently 1000 samples from each model of each value are collected and compared) and performing bootstrapping on the samples.

5 Future Work

Even though the model accounts only for 2 loci, it can be easily adapted to a multi-loci model, which will give more marginal trees corresponding to different loci. However, as far as the statistics are concerned, it is reasonable to assume that the two loci model is enough to model the behaviour. No matter at which location the recombination event happens, if we only look at the two loci at either end of the gene, we obtain the same marginal trees as would expect from the two loci model. Nonetheless, the multi-loci model would be able to further the study in determining the effect of recombination on linkage disequilibrium under the birth-death model assumptions. Linkage disequilibrium studies the correlations between different genes at different locations on the genome. Since recombinations would have smaller effects on genes which are close to each other than far (by the fact that the two loci at either ends of the genome always come from different parents under any reciprocal recombination effect), the linkage disequilibrium can be utilised to describe this effect.

The distributions were estimated and compared using simulations. In order to more accurately show the differences as described, evaluation of the expressions for the statistics would be necessary. Efforts have been made to derive the expressions but as hinted from calculations from Stadler [5], the expressions would involve many complicated functions.

Another feature that has not been included in the simulations is the mutation events. The existence of mutation events is the basis that recombination events are important. Recombinations help spread mutations by introducing the mutated sequences into populations in the reciprocal case, or through recombination birth in the unidirectional case. The mutation events are important biologically as it is the process where diversity is introduced as well as makes recombinations traceable. However, the effects of mutation are the same under both models in terms of the rate. This hence means that

it does not change the topologies. Mutations change the branch lengths to the same extent for both models since it happens uniformly. The effect of mutation can be investigated through simulations. The 'ms' programme has an inbuilt mutation function while the Birth-Death simulation also has the function, though the mutation parameter was set to 0 for the simulations in the previous parts.

References

[1] S. Mike, *Phylogeny: Discrete and Random Processes in Evolution*. 2016.

[2] J. Kingman, "Stochastic processes and their applications," 1982.

[3] T. Stadler, "On incomplete sampling under birthdeath models and connections to the sampling-based coalescent," *Journal of Theoretical Biology*, vol. 261, no. 1, pp. 58–66, 2009.

[4] T. Stadler, T. G. Vaughan, A. Gavryushkin, S. Guindon, D. Khnert, G. E. Leventhal, and A. J. Drummond, "How well can the exponential-growth coalescent approximate constant-rate birth-death population dynamics?," *Proceedings. Biological sciences / The Royal Society*, vol. 282, p. 20150420, May 7, 2015.

[5] T. Gernhard, "The conditioned reconstructed process," *Journal of theoretical biology*, 2008.

[6] F. Wilcoxon, "Individual comparisons by ranking methods," vol. 1, pp. 80–83, 1945.

[7] D. G. Kendall, "On the generalized "birth-and-death" process," *The Annals of Mathematical Statistics*, vol. 19, 1948.

[8] J. Hein, M. H. Schierup, and C. Wiuf, *Gene genealogies, variation and evolution*. Oxford [u.a.]: Oxford Univ. Press, 2006.

[9] R. R. Hudson, "Generating samples under a wright-fisher neutral model of genetic variation," *Bioinformatics*, vol. 18, pp. 337–338, Feb 1, 2002.

[10] "Package 'phytools' in r," 2016.

[11] "Package 'phyclust' in r," 2016.

[12] M. Stephens and P. Donelly, "Inference in molecular popular genetics," *J. R. Stat. Soc. Ser. B*, pp. 605–635, 2000.

[13] J. Felsenstein, *Inferring phylogenies*. Sunderland, Mass: Sinauer, 1. ed. ed., 2004.

[14] T. Stadler, "Sampling through time in birth-death trees," *Journal of theoretical biology*, 2010.

[15] J. A. Bucklew, *Introduction to rare event simulation*. New York, NY [u.a.]: Springer, 2004.

[16] K. Dialdestoro, J. A. Sibbesen, L. Maretty, J. Raghwani, A. Gall, P. Kellam, O. G. Pybus, J. Hein, and P. A. Jenkins, "Coalescent inference using serially sampled, high-throughput sequencing data from intrahost hiv infection," *Genetics*, vol. 202, p. 1449, Apr 1, 2016.

[17] R. C. Griffiths and S. Tavare, "Simulating probability distributions in the coalescent," *Theoretical Population Biology*, vol. 46, no. 2, pp. 131–159, 1994.

A R functions and libraries created

A.1 Birth-Death Model Simulation

```
BDn<-function(n=10,lambda=1,
mu=0.0005,theta=2.6*10^-9,rho=0.002){ #Rates
  A=matrix(0,n+1,5)
  A[1,1]=1
  for(i in 1:n){
    m=A[i,1]
    A[i,2]=-log(runif(1))/(m*lambda+m*mu
    +m*theta+choose(m,2)*rho)
    U=runif(1)
    A[i,3]=(U<m*lambda/(m*lambda+m*mu+m*theta+choose(m,2)*rho))
    +(U<(m*lambda+m*mu)/(m*lambda+m*mu+m*theta+choose(m,2)*rho))+
    (U<(m*lambda+m*mu+m*theta)/(m*lambda+m*mu
    +m*theta+choose(m,2)*rho)) #Choosing which event it is
    A[i,4]=floor(runif(1)*m)+1
    #Which individual it happens to, independent of event
    if(A[i,3]==0){A[i,5]=floor(runif(1)*m)+1 } #Recombination
    if(A[i,3]==0){c=0} #Change in number of individuals is 0
    if(A[i,3]==1){c=0  #Mutation, change in number is 0
    A[i,5]=floor(runif(1)*2)+1}
    if(A[i,3]==2){c=-1} #Death
    if(A[i,3]==3){c=1} #Birth
    A[i+1,1]=A[i,1]+c #Change in number of individuals
    if(A[i+1,1]==0)break #Stop if all individuals die
  }
  m=A[n+1,1]
  A[n+1,2]=-log(runif(1))/(m*lambda+
  m*mu+m*theta+choose(m,2)*rho)
  #Amount of time between last event and observation
  if(any(A[,1]==0)==0){return(A)
  #If the tree does not die out, output
  if(!(any(A[,1]==0)==0)){
  return(A=BDn(n,lambda,mu,theta,rho))}
  #Re-run if the tree dies out
}
```

A.2 Tree Plotting

Here we use the left tree as an example.

```
MatrixTreeStringleft<-function(X){
  x=dim(X)[1]
  A=matrix(0,1,4)
```

```
C=matrix("",1,3)
A[1,1]=1 #branch number
C[1,2]=":" #separation
A[1,2]=0 #time
A[1,3]=1 #live string
A[1,4]=1 #count number
death.count<-0
for(i in 1:x){
  if(X[i,3]==3){
    B<-matrix(0,dim(A)[1]+2,4)
    D<-matrix("",dim(A)[1]+2,3)
    n=which(A[,1]==X[i,4])
    if(!(n==1)){
      B[1:(n-1),]=A[1:(n-1),]
      D[1:(n-1),]=C[1:(n-1),]}
      B[(n+2):(dim(A)[1]+2),2:4]=A[n:dim(A)[1],2:4]
      for(k in (n+2):(dim(A)[1]+2)){if(!(A[k-2,1]==0))
      {B[k,1]=A[k-2,1]+1}}
      D[(n+2):(dim(A)[1]+2),]=C[n:dim(A)[1],]
    for(j in 1:dim(B)[1]){
      if(B[j,3]==1){B[j,2]<-B[j,2]+X[i,2]}
    }
    if(!(C[n,1]=="")){D[n,1]=paste(c(C[n,1],"("),collapse="")}
    if(C[n,1]==""){D[n,1]="("}
    B[n,1]=X[i,4]
    D[n,2]=":"
    B[n,2]=0
    D[n,3]=","
    B[n,3]=1
    B[n,4]=X[i,4]
    B[(n+1),1]=X[i,4]+1
    D[(n+1),2]=":"
    B[(n+1),2]=0
    B[(n+1),3]=1
    B[(n+1),4]=X[i,4]+1
    D[(n+2),1]=""
    B[(n+2),1]=0 #place holder to become ) later
    B[(n+2),3]=0 #branch length will not increase
    A<-B
    C<-D
  }
  if(X[i,3]==2){
    for(j in 1:dim(A)[1]){
      if(A[j,3]==1){A[j,2]<-A[j,2]+X[i,2]}
```

```
      }
      m=which(A[,1]==X[i,4])
      A[m,3]=0
      for(k in 1:dim(A)[1]){
        if(A[k,1]>A[m,1]){A[k,1]<-A[k,1]-1}
      }
      A[m,1]=death.count+1+X[x,1]*10
      death.count<-death.count+1
    }

    if(X[i,3]==1){
      for(j in 1:dim(A)[1]){
        if(A[j,3]==1){A[j,2]<-A[j,2]+X[i,2]}
      }}
    if(X[i,3]==0){
      if(X[i,4]<X[i,5]){
        for(j in 1:dim(A)[1]){
          if(A[j,3]==1){A[j,2]<-A[j,2]+X[i,2]
          c<-A[j,1]
          if(X[i,4]==A[j,1]){c=X[i,5]}
          if(X[i,5]==A[j,1]){c=X[i,4]}
          A[j,1]<-c
          }
        }
      }
    }
  }
  for(i in 1:dim(A)[1]){

    if(A[i,1]==0){A[i,1]=")"}
    if(!(A[i,1]==")")){A[i,1]=paste(c("s",A[i,1]),collapse="")}
  }
  E=matrix(0,dim(A)[1],5)
  E[,1]=C[,1]
  E[,2]=A[,1]
  E[,3]=C[,2]
  E[,4]=A[,2]
  E[,5]=C[,3]
  return(E)
}
```

This is followed by the plotting code.

```
X=BDn(n=20,rho=0.5)
Y=MatrixTreeStringleft(X)
```

```
n=dim(Y)[1]
z=paste(c(t(Y[1:(n-1),]),")",";"),collapse="")
a<-read.tree(text=z)
plot(a)
axis(1)
```

A.3 Branches from a node

Here we use the left tree as an example.

```
CheckBranchesleft<-function(A,level){
  #A=tree, level=event level, count=which individual at that event
  x=dim(A)[1]
  n=A[x,1]
  X=matrix(0,n,2)
  X[,1]=1:n
  for (i in 1:n){
    count<-i
    for (j in (x-1):level){
        if(A[j,3]==3){if(A[j,4]<count-1){count<-count-1}}
        if(A[j,3]==2){if(A[j,4]<count){count<-count+1}}
      if(A[j,3]==0){
        if(A[j,4]<A[j,5]){
          l=count
          m=A[j,4]
          n=A[j,5]
          if(count==m){l=n}
          if(count==n){l=m}
          count=l
        }}
    }
    X[i,2]<-count
  }
  return(X)
}
```

A.4 Comparison of Time to Coalescent of two nodes

Here is a code which finds the time to coalescent of two selected nodes.

```
BDFindMRCAleft<-function(A,n1,n2){
  y=dim(A)[1]
  count1=n1
  count2=n2
  continue=1
```

```
  t=0
  x=y-1
  for(j in (y-1):1){
    smallcount=min(count1,count2)
    small=which(c(count1,count2)==smallcount)
    bigcount=max(count1,count2)
    big=which(c(count1,count2)==bigcount)
    if((smallcount==bigcount|bigcount==smallcount+1)
    && A[j,3]==3 && continue==1) {
      t<-t+A[j,2]
      x<-j
      if(A[j,3]==3){if(A[j,4]<count1-1){count1<-count1-1}}
      if(A[j,3]==3){if(A[j,4]<count2-1){count2<-count2-1}}
      if(A[j,3]==2){if(A[j,4]<count1){count1<-count1+1}}
      if(A[j,3]==2){if(A[j,4]<count2){count2<-count2+1}}
      continue=0
    }
    if(continue==1){
      t<-t+A[j,2]
      x<-j
      if(A[j,3]==3){if(A[j,4]<count1-1){count1<-count1-1}}
      if(A[j,3]==3){if(A[j,4]<count2-1){count2<-count2-1}}
      if(A[j,3]==2){if(A[j,4]<count1){count1<-count1+1}}
      if(A[j,3]==2){if(A[j,4]<count2){count2<-count2+1}}

    }
  }
  return(t)
}
```

This is followed by the comparison codes.

```
C=matrix(0,1000,2)
for (i in 1:1000){
  X=BDn(100,1,0.5,0,0)
  m=X[dim(X)[1],1]
  n1=floor(runif(1)*m)+1
  n2=floor(runif(1)*m)+1
  n=BDFindMRCA(X,n1,n2)
  C[i,1]=n
  Y=ms(nsam=m,opts="-T␣-G␣0.5")
  treel<-read.tree(text=Y[3])
  C[i,2]=dist.nodes(treel)[floor(runif(1)*m+1),floor(runif(1)*m+1)]*m/
}
wilcox.test(C[,1]-C[,2])
```

```
median(C[,1]/C[,2])
median(C[,1])
median(C[,2])
```

A.5 Comparison of proportion of branches surviving to the present originating from a node

This is the proportion of branches from a branch of the left tree.

```
CheckBranchesleft<-function(A,level){
  #A=tree, level=event level, count=which individual at that event
  x=dim(A)[1]
  n=A[x,1]
  X=matrix(0,n,2)
  X[,1]=1:n
  for (i in 1:n){
    count<-i
    for (j in (x-1):level){
        if(A[j,3]==3){if(A[j,4]<count-1){count<-count-1}}
        if(A[j,3]==2){if(A[j,4]<count){count<-count+1}}
      if(A[j,3]==0){
        if(A[j,4]<A[j,5]){
          l=count
          m=A[j,4]
          n=A[j,5]
          if(count==m){l=n}
          if(count==n){l=m}
          count=l
        }}
    }
    X[i,2]<-count
  }
  return(X)
}
```

This is the code for comparison.

```
E=matrix(0,1000,2)
for (i in 1:1000){
  X=BDn(500,2,0.5,0,0.1)
  m=X[dim(X)[1],1]
  n=floor(runif(1)*10)+1
  r=X[n,1]
  t=floor(runif(1)*r)+1
  s=CheckBranchesleft(X,n)
```

```
  E[i,1]=length(which(s[,2]==t))/m
  Y=ms(nsam=m,opts="-T␣-G␣1.5␣-r␣0.1␣2")
  treel<-read.tree(text=Y[3])
  E[i,2]=length(
  which(getDescendants(treel,m+floor(runif(1)*10))<m))/m
}
wilcox.test(E[-which(E[,1]==0),1],E[,2])
median(E[-which(E[,1]==0),1])
median(E[,2])
```

B Comparison of 50 trees compared with Stadler's calculations

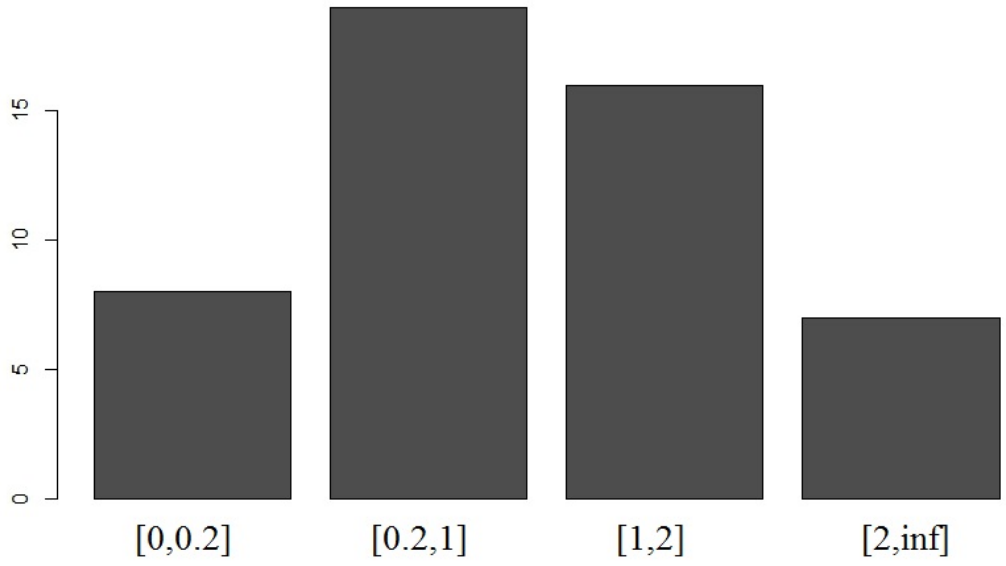

Figure 12: A Bar Plot of the simulated trees with Stadler's Calculations

By the χ^2-test, this has p-value 0.17, hence Stadler's calculations are not rejected in this sample.