A MARKOV MODEL OF THE COALESCENT WITH RECOMBINATION AND POPULATION SUBSTRUCTURE

1. Abstract

Different biological phenomena can leave similar signatures in the genomic data. This project explores how recombination and population substructure can lead to patterns of ancestry across different places in the genome that could appear similar to the signatures of other biological phenomena, such as horizontal gene transfer. In this project, we develop a framework for simulating genomic data under a neutral model with population substructure and recombination. We extend the model of Simonsen and Churchill (1997), a Markov model of the coalescent with recombination between a discrete number of loci, to also include two subpopulations with a symmetric migration rate between them. We describe the state space and transition matrix of this model. Using the transition matrix, we find that the probability of equal time to the most recent common ancestor at each loci decreases with decreasing migration rate and increasing recombination rate. The model described in this paper is a first step towards developing a statistic to discriminate between the genomic signature of horizontal gene transfer and the possible genomic signature of recombination and population substructure.

2. Introduction

Incomplete lineage sorting refers to the neutral process by which ancestral lineages from a population fail to coalesce more recently in time than the ancestral population split with one or more other populations (Degnan and Rosenberg 2009). Incomplete lineage sorting creates the possibility that a lineage, $a1$, from population A, shares a common ancestor with another lineage, $b1$ from population B, more recently than $a1$ shares an ancestor with other lineages from population A. Due to incomplete lineage sorting, when species are closely related, the topology of the sample genealogy (the gene tree), can differ from the topology of the species phylogeny (the species tree) (Hudson 1983, Tajima 1983).

Recombination can lead to variation in the genealogy of a sample across loci. Because gene trees can differ from species trees and gene trees can vary across different loci, some loci in the genome may have genealogies which accord with the overall species tree, while neighboring loci do not.

We will refer to the phenomena in which the local gene tree does not accord with the topology of the overall species tree as *gene-species tree discordance*. With recombination and incomplete lineage sorting, gene-species tree discordance can vary across the genome. For example, the relationship between humans, chimps, and gorillas varies across the genome (Dutheil et al. 2009). The species tree for humans, chimps, and gorillas is considered to be {{human, chimp}, gorilla}. While {{human, chimp}, gorilla} is the topology over most of the genome, at some loci, the topology relating the samples is {{human, gorilla}, chimp} or {{chimp, gorilla}, human}.

Beyond incomplete lineage sorting during descent from a shared ancestral population, gene-species tree discordance can also be evidence for other biological phenomena. Some methods for identifying horizontal gene transfer rely on identifying tracts of the genome in which gene-species discordance is adjacent to other regions of concordance between the gene tree and species tree (Ochman 2001). Modeling different processes which lead to gene-species tree discordance can be utilized to improve inference methods for differentiating between them (Ruths and Nakhleh 2005). In this project we are interested in modeling how variation in gene-species tree discordance across different loci can arise in a neutral model.

2.1. **A coalescent model.** Slatkin and Pollack (2006) introduced a framework for analyzing the probability of gene-species tree discordance in a simple three-species coalescent model with recombination. Slatkin and Pollack (2006) considered the case of two loci and a sample of one individual per species. Their model utilized a spatially discrete model of the coalescent with recombination introduced by Simonsen and Churchill (1997). The probability of gene-species tree discordance at each of the loci depends on the number of lineages present in the ancestral population.

Slatkin and Pollack (2008) described a second model in which ancestral population substructure increases the probability of gene-species tree discordance. They showed how even in a simple model, extreme substructure within the ancestral population can lead to a discordant gene tree topology that is more likely than the concordant topology. De Giorgio and Rosenberg (2016) demonstrated that this discordance due to a structured ancestral population can introduce challenges for several different phylogenetic inference algorithms.

A simple demographic history which could lead to the scenario of increased gene-species tree discordance consists of three modern day populations A, B, and C with the topology $((A, B)C)$. In the shared ancestral population between A and B, which is not ancestral to C, structure persists between A and B. Further back in time, in the ancestral population which includes A, B, and C, population structure continues to persist between A and B, but the ancestors of B and C mix and have no population structure. In this context, Slatkin and Pollack (2008) demonstrated that the discordant gene tree topology $(A(BC))$ can increase in probability above the probability of the concordant topology $((AB)C)$. Figure 1 illustrates this scenario.

In the model of Slatkin and Pollack (2008), the number of lineages in a population at a given time point is represented as a continuous Markov process.

3. Extending the model

In the current project, we are interested in incorporating recombination into the model suggested by Slatkin and Pollack (2008), in which population substructure can increase the probability of gene-species tree discordance. We use a Markov model of the coalescent that explicitly incoporates both recombination with discrete loci and population substructure, to model a neutral process in which the probability of gene-species tree discordance can vary at different loci.

As Slatkin and Pollack (2006) demonstrated, Markov models of genealogies at discrete loci are useful for modeling ancestry in the context of changing demographic history. To explicitly model recombination and migration in our model, we build upon the model described by Simonsen and Churchill (1997).

In the first part of this project, we delineate the state space and transition probabilities of a coalescent model with recombination and population substructure for two loci with both two and three individuals. We have created a program in `python` to generate these state spaces and the associated diagrams that describe them for general sample size, n. We also introduce an analytical formula for the size of the state space for two loci, two subpopulations, and general n samples.

Following the analysis of Simonsen and Churchill (1997), we find the probability of different tree heights between the two loci when there are two lineages present at each locus in the model. With two individuals and two loci, Simonsen and Churchill (1997) demonstrate how a model of recombination between two loci can lead to different coalescence times in the genealogies at each locus. When the demographic features of the population change over time, such as in a three-species model, different timing of coalescent events between lineages associated with each locus could lead to scenarios in which the coalescence events for different loci occur in different regimes of the demographic model, e.g. before or after a population splits into two. This scenario leads to the possibility that the joint probability that two loci have the same topology depends on the recombination level and the ancestral population substructure.

This project is the first step toward the goal of modeling the effect of population substructure on gene-species concordance across loci. The transition matrix and state space described in this project could be used in future projects to expand the Markov model of Slatkin and Pollack (2008) to explicitly incorporate recombination and analyze the probability of different loci falling into different topology classes due to incomplete linkage, such that their coalescent times are not entirely correlated.

4. Markov Model of Coalescent with Recombination and Population Substructure

4.1. **State space and transition matrix for model with recombination but no population substructure.** To delineate the state space and transition matrix of a Markov model of the coalescent with recombination, we build from the model of Simonsen and Churchill (1997). The state space of the process described by Simonsen and Churchill (1997) characterizes the numbers of lineages present at each locus. Similar to the standard coalescent model, the process begins in the present and progresses backwards in time.

Similarly to the standard coalescent model, in the Simonsen and Churchill (1997) model, a coalescence corresponds to an individual producing two offspring forward in time. However, because the model incorporates recombination, it is possible that an individual produces two offspring, but that one or both offspring only inherit one of the two loci from the individual of interest due to recombination. Introducing recombination into the model also introduces the possibilty that the two loci which are ancestral to the present-day individual recombined onto the same haplotype from two different individuals. Figure 2(A) illustrates how, thinking forward in time, individuals can produce offspring that only inherit one of the two loci or that inherit both; these events can be considered coalescences between haplotypes that are ancestral at one or both loci. Figure 2(B) illustrates how, thinking forward in time, loci ancestral to the present day sample can recombine onto the same haplotype from two individuals who only have one locus ancestral to the present-day sample. Coalescence and recombination define the transitions between states in the Markov process.

Using the terminology from Simonsen and Churchill (1997), the states of the Markov process represent the number of lineages present for each locus by a unique tuple (i, j, k). We define i as the number of "left type" lineages, those that are ancestral to the present day sample at the left hand locus but not the right (represented as a circle in Figure 3(A)). Similarly j represents the number of "right type" lineages, those that are ancestral to the present day sample at the right locus, but not the left (represented as a square in Figure 3(B)). Lastly, k represents the number of "double type" lineages, those that are ancestral to the present day sample at both loci (represented as a circle in Figure 3(C)). Note that a haplotype which is ancestral to only one locus has some other allele at the other locus, but we are only interested in tracking the genealogy at loci which are ancestral to the present day sample.

The transitions between the states in the Simonsen and Churchill (1997) model are either coalescence or recombination events. In a recombination event, one double becomes a right type and left type (see Figure 2(B)). The population recombination rate, $R = Nr$, represents the rate at which two lineages which are each ancestral at opposite loci recombine onto the same haplotype, r, scaled by the number of haplotypes in the population, N. In a recombination event, k decreases and both i and j increase. In a coalescent event, a double type can coalesce with a double (k decreases), a left type (i decreases), or right type (j decreases) or a right and left type can also coalesce into a double (i and j decrease, k increases) (see Figure 2(A)). Another type of coalescent event can occur in which two haplotypes which are ancestral to the sample only at one locus coalesce. Simonsen and Churchill (1997) model this possibility as two right types coalescing to one right type (i decreases) or two left types coalescing to one left type (j decreases). The coalescence rates are determined by the number of lineages present of each type. As in the standard coalescent model, Simonsen and Churchill (1997) assumes that the probability of simultaneous events is small enough to ignore. Figure 4 illustrates how the Markov chain describes the genealogy of the process.

Similar to the standard coalesccent model, Simonsen and Churchill (1997) define a starting state with n lineages. In the starting state, for each individual both loci are part of the genealogy so all n lineages are in double types. As lineages coalesce, the total number of lineages decreases at each locus. Recombination does not change the total number of lineages, only changing whether or not both loci are on the same haplotype (in a double type) or on two different haplotypes (a right type and left type). At any given time, the number of lineages, ν_r, which are ancestral to the present day sample at the right locus must be contained in a right type or a double type. Similarly the number of lineages, ν_ℓ, ancestral to the present-day sample at the left locus must be in a left type or a double type. As in the standard coalescent model, the number of lineages decreases from n to

1, so $1 \leq \nu_r \leq n$ and $1 \leq \nu_\ell \leq n$. Therefore the following equations define the state space,

$$1 \leq i + k \leq n \tag{1}$$

$$1 \leq j + k \leq n \tag{2}$$

$$0 \leq i, j, k \leq n \tag{3}$$

Simonsen and Churchill (1997) consider the state space for $n = 2$ fom eqs. (1)–(3). This computation yields 9 total states. We illustrate the state space diagram with transitions in Figure 5.

Simonsen and Churchill (1997) also describe how to extend their model to other numbers of loci and samples but only illustrate the state space and transition matrix explicitly for $n = 2$. For two loci and general n, they show that the number of states is $n(2n^2 + 9n + 1)/6$.

For $n = 3$ the number of states is 23. Using, `python`, we delineate the case of $n = 3$ case explicitly. We generated the states based on the allowable tuples given in inequalities (1)–(3). We generated the transition matrix by considering the transitions- recombination or coalescence- which lead to valid states. We generated a diagram using the software `graphviz`. For $n = 3$, we illustrate the state space diagram with transitions in Figure 6.

4.2. **Extending the model to include population substructure.** To extend this model to include population substructure, we consider a model with two populations with a migration rate, $M = Nm$, where N is the number of haplotypes in each subpopulation and m is the rate of migration for an individual haplotype to migrate to the other subpopulation.

Recombination and coalescent events change the number of lineages of each type present in one population only. Two lineages can only coalesce at a given time if they are currently in the same population. A double type can recombine into a right type and left type, both of which remain in the same population. Any type of lineages (right type, left type, or double type) can migrate. As in the model with only recombination, we assume that the resulting lineages remain in the same population given that coalescence, recombination, and migration are so rare that we can ignore the possibility of two events occuring simultaneously.

From the nine states in the Simonsen and Churchill (1997) model with two samples and two loci, the extended population substructure and recombination Markov chain includes 46 states, as shown in Figure 8. We label the states using the following terminology: $\{iA,\ jA,\ kA\}\{iB,\ jB,\ kB\}$ where $\{iA,\ jA,\ kA\}$ represents the number of lineages of right types, left types, and double types in population A and $\{iB,\ jB,\ kB\}$ represents the number of lineages of different types in population B.

The number of lineages in the model with recombination and no population substructure are constrained by eqs. (1)–(3). In the model with both recombination and population substructure, we can generate all of the states by accounting for all possible population labels on the lineages in each state from the original model with recombination but no population substructure, i.e. the states are constrained by eqs. (1)–(3) for $i = iA + iB$, $j = jA + jB$, and $k = kA + kB$.

Figure 7 illustrates how the Markov chain corresponds to the coalescent process with recombination and migration.

Ignoring topology, extending the Simonsen and Churchill model to include n samples and 2 loci, we used a combinatorial argument to find the number of states in the model given a value of n.

Starting from the number of possible states from n lineages in the model without substructure, we find the number of lattice point solutions to inequalities (1)–(3):

$$\sum_{k=0}^{n}\sum_{j=0}^{n-k}\sum_{i=0}^{n-k} \mathbb{1}\{0 < i+k\}\mathbb{1}\{0 < j+k\} \tag{4}$$

$$\sum_{k=1}^{n}\sum_{j=0}^{n-k}\sum_{i=0}^{n-k} \mathbb{1}\{0 < i+k\}\mathbb{1}\{0 < j+k\} + \sum_{j=1}^{n}\sum_{i=1}^{n} \mathbb{1}\{0 < i\}\mathbb{1}\{0 < j\} \tag{5}$$

$$\sum_{k=1}^{n}\sum_{j=0}^{n-k}\sum_{i=0}^{n-k} \mathbb{1}\{0 < i+k\}\mathbb{1}\{0 < j+k\} + n^2 \tag{6}$$

which simplifies to $n(2n^2+9n+1)/6$ from Simonsen and Churchill (1997).

For any given tuple (i, j, k) in the model with only recombination and no population substructure, we can partition the lineages into two subpopulations. Each of the i left-type lineages can be in population A or population B. Because the lineages are exchangeable, placing the i lineages into two non-exchangeable populations give $i+1$ total states: either 0, 1, 2, …, or i left-type lineages are placed in population A. Similarly, the j right-type lineages generate $j+1$ states, and the k double-type lineages generate $k+1$ states. The number of ways to create this partition is equivalent to counting the ways to split into two groups three series of i, j, and k identical objects, because all lineages of a given type are exchangeable with the other lineages of that type and the placement of any lineage in a population is independent on the placement of other lineages.

We have

$$S(n) = \sum_{k=0}^{n}\sum_{j=0}^{n-k}\sum_{i=0}^{n-k} \mathbb{1}\{0 < i+k\}\mathbb{1}\{0 < j+k\}(i+1)(j+1)(k+1) \tag{7}$$

$$= \Big(\sum_{k=1}^{n}\sum_{j=0}^{n-k}\sum_{i=0}^{n-k}(i+1)(j+1)(k+1)\Big) + \sum_{j=1}^{n}\sum_{j=1}^{n}(i+1)(j+1) \tag{8}$$

$$= \sum_{z=0}^{n}\Bigg((n-z+1)\bigg[\Big(\sum_{y=1}^{z}(y+1)\Big)^2 + \mathbb{1}\{n-z>0\}(1+2\Big(\sum_{y=1}^{z}(y+1)\Big)\bigg]\bigg]\Bigg). \tag{9}$$

We used Faulhaber's formulas to simplify the expression to find the closed form expression is below.

$$S(n) = \frac{n^6}{120} + \frac{n^5}{8} + \frac{3n^4}{4} + \frac{55n^3}{24} + \frac{329n^2}{120} + \frac{n}{12}. \tag{10}$$

For $n = 2$, we have $S(n) = 46$ and for $n = 3$ we have $S(n) = 184$. The program we created in `python` generate these states and transition matrix, given the allowable states, and the allowable transitions between states- coalescence, recombination, and migration. Figure **??** shows the state space and transitions for the case of $n = 3$ model with recombination and migration.

4.3. **Probability of equal tree height.**

Using the transition matrix of the Markov process defined in Sections 4.1 and 4.2, we can observe the properties of the genealogical process with recombination between 2 loci in the case of $n = 2$ and $n = 3$ samples per population, with recombination between the two loci in a model with or without population structure.

In the model with no population substructure and two loci, Griffiths (1981) determined an analytical formula for the probability that the two trees have equal height, i.e. that the time to the most recent common ancestor between the right and left locus are the same. This probability is recapitulated in Simonsen and Churchill (1997) who explored the probability of equal tree height analytically by analyzing long-term behavior of the transition matrix. This analysis enabled them to find the probability that the process never enters the states of the process in which $i + k = 1$ but $j + k > 1$ or, similarly, in which $j + k = 1$ but $i + k > 1$. They found the probability of equal tree height to be

$$\frac{9+R}{9+13R+2R^2}, \tag{11}$$

which decreases as R increases.

In the model with only recombination and no population substructure, if the process enters states $\{101\}$, $\{011\}$, $\{210\}$, $\{120\}$, or $\{110\}$ the tree heights will be unequal. The probability of equal tree height can be determined by finding the probability that the process never enters one of these states.

For the model with recombination and population substructure, each state corresponds to a state from the model with recombination and no population labels. To find the probability of equal tree height for a given recombination rate, R, and migration rate, M, we can consider whether the process ever enters states $\{iA, jA, kA\}\{iB, jB, kB\}$ such that $\{iA + iB, jA, +jB, kA + kB\}$ equals one of $\{101\}, \{011\}, \{210\}, \{120\}$, or $\{110\}$.

In the $n = 2$ case, this results in 24 states in the model with recombination and population substructure which lead to unequal tree heights between the two loci. We estimated the probability of equal tree height by running 10^4 simulations of the Markov process for the model with recombination and population substructure with different recombination rates, R, and migration rates, M. We found that for increasing recombination rate, R, and decreasing migration rate, M, the probability of equal tree heights decreases.

At the start of the process, the two lineages begin in different subpopulations. A small migration rate increases the time in which recombination can break up the two loci onto different haplotypes because coalescences cannot occur unless there are two lineages associated with the same locus in the same population. Once the two loci are broken up, the probability of equal tree height is decreased because the two lineages will behave independently and each must migrate separately to the other population in order for a coalescence to occur. Thus, the probability of unequal tree height increases as the migration rate, M, decreases for a given recombination rate, R.

5. Conclusion

In this project, we have delineated the state space and transition matrix for a two-locus coalescent model with recombination and migration for $n = 2$ and $n = 3$ lineages per population in a two-population model by extending the model of Simonsen and Churchill (1997). For the $n = 2$ case, by simulating the Markov process from its transition matrix we estimated the probability of unequal tree height between the two loci.

We found that the size of the state space increases polynomially with n, in particular following n^6. The probability of equal tree height decreases with decreasing M and increasing R. The Markov framework described in this chapter could be used to extend the model described in Slatkin and Pollack (2008) to better understanding how neutral processes could cause the probabilities of different topologies at incompletely linked loci to differ. As described in Section 4.3, population substructure increases the tree height and increases the probability that the tree heights between loci will differ compared to a model with no population substructure.

For a demographic scenario with population substructure which changes over time, such as in the Slatkin and Pollack (2008) model, the probabilities determining migration– and as a consequence, coalescence– may vary between two loci if the process at one locus is complete (i.e. one ancestor) while the other is still ongoing. In the context of the demographic schema described by Slatkin and Pollack (2008), population substructure also increases the probability of gene-species tree discordance. One use of our model would be to explore how an increased probability of unequal tree heights, due to the presence of population substructure, and an increased probability of gene-species tree discordance, due to the presence of population substructure, could increase the probability of gene-species tree discordance at one locus and gene-species tree concordance at the other locus.

Modeling gene-species tree discordance in this way lays the groundwork for developing a 3-point statistic to estimate the parameters necessary to produce gene-species tree discordance in short tracts in the genome flanked by gene-species tree concordance on either side. Here, we only explore the 2 loci model with migration, but an extension to include 3 loci, could model the probability that across three loci in the genome one left most loci has a concordant topology, the middle loci has a discordant topology, and the right most loci has a concordant topology– a genomic signature similar to that of horizontal gene transfer.

References

De Giorgio, M. and N. A. Rosenberg (2016). Consistency and inconsistency of consensus methods for inferring species trees from gene trees in the presence of ancestral population structure. *Theoretical Population Biology 110*, 12–24.

Degnan, J. H. and N. A. Rosenberg (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution 6*, 332–340.

Dutheil, J. Y., G. Ganapathy, A. Hobolth, T. Mailund, M. K. Uyenoyama, and M. H. Schierup (2009). Ancestral population genomics: The coalescent hdden markov model approach. *Genetics 183*, 259–274.

Griffiths, R. C. (1981). Neutral two-locus multiple allele models with recombination. *Theoretical Population Biology 19*(169-186).

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology 23*, 183–201.

Ochman, H. (2001). Lateral and oblique gene transfer. *Current opinion in genetics and development 11*, 616–619.

Ruths, D. and L. Nakhleh (2005). Recombination and phylogeny: effects and detection. *International journal of bioinformatics research and applications 1*, 1–11.

Simonsen, K. L. and G. A. Churchill (1997). A markov chain model of coalescence with recombination. *Theoretical Population Biology 52*, 43–59.

Slatkin, M. and J. L. Pollack (2006). The concordance of gene trees and species trees with two linked loci. *Genetics 172*, 1979–1984.

Slatkin, M. and J. L. Pollack (2008). Subdivision in an ancestral species creates asymmetry in gene trees. *Molecular Biology and Evolution 10*, 2241–2246.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics 105*, 437–460.

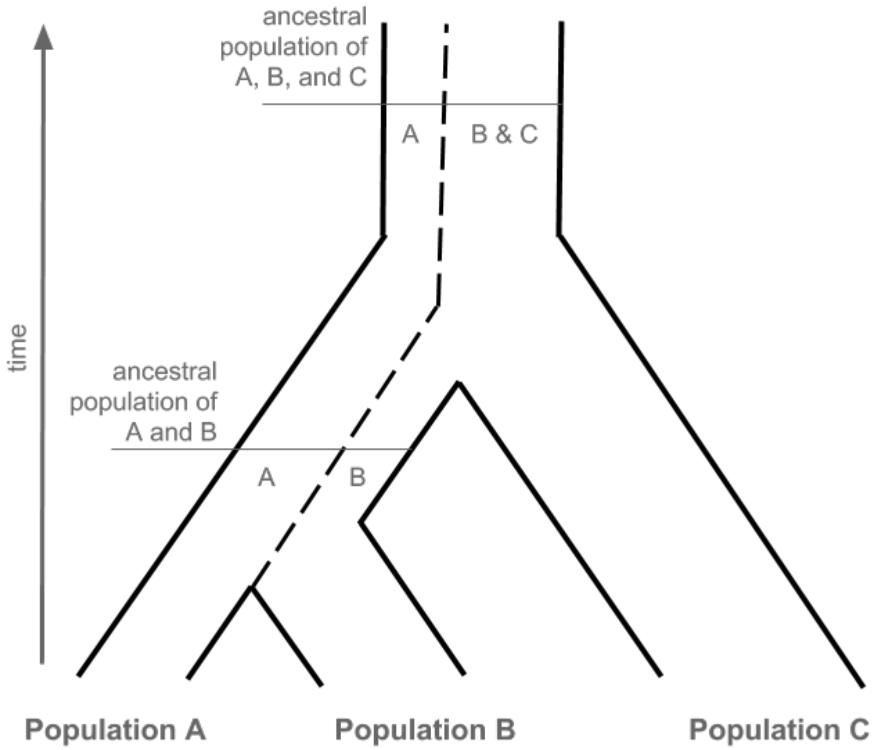
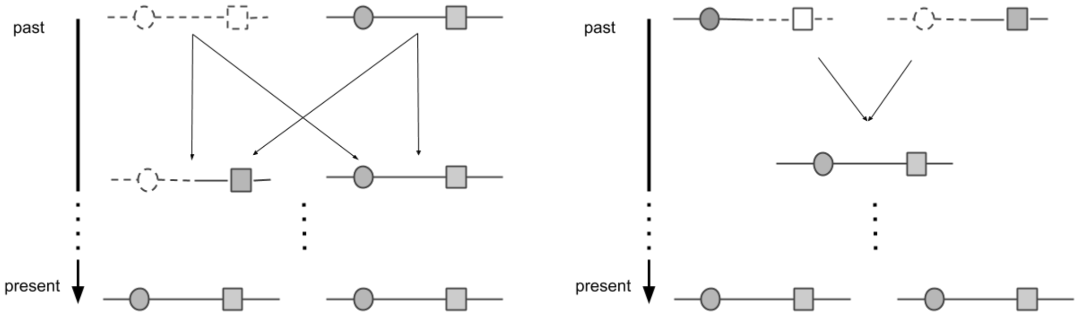

ancestral population of A, B, and C
A
B & C
time
ancestral population of A and B
A
B
Population A
Population B
Population C

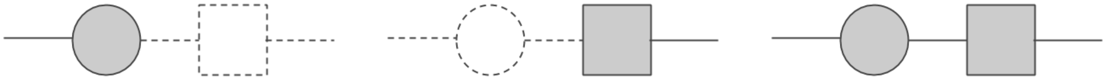
FIGURE 1. A simple demographic history between three populations in which population substructure leads to an increase in the probability of gene-species tree discordance compared to a model without substructure (Slatkin and Pollack 2008).



(A) Coalescence

(B) Recombination

Figure 2. Coalescence and recombination for pairs of loci. Consider the light gray shapes as the loci that are ancestral to the present-day sample, and consider empty shapes as loci that are not ancestral. The circle and square represent two loci of interest. In a model with recombination, looking backwards in time, the loci that are ancestral to the present-day sample can coalesce onto the same haplotype or can recombine away from each other. (A) An individual can produce offspring who inherit both loci or only one locus of interest. (B) Lineages ancestral to the present-day sample at different loci can recombine onto the same haplotype.



(A) Left type

(B) Right type

(C) Double type

Figure 3. Ancestral loci in a two locus model. The states of the Markov chain are determined by the unique tuple denoting the number of “right type”, “left type”, and “double type” locus pairs. For an individual, the gray shapes represent loci ancestral to the present-day sample.

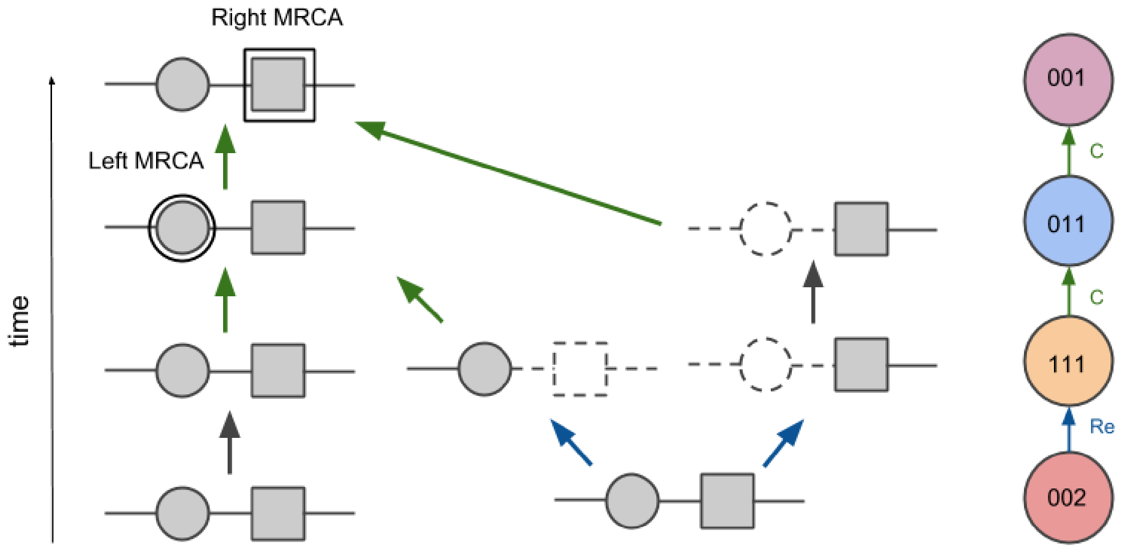


FIGURE 4. An illustrative example of a process that traces the lineages of two loci for two individuals. Recombination and coalescence can change the number of lineages of each type in each population. In this case, the tree associated with left loci is shorter than the tree associated with the right loci.

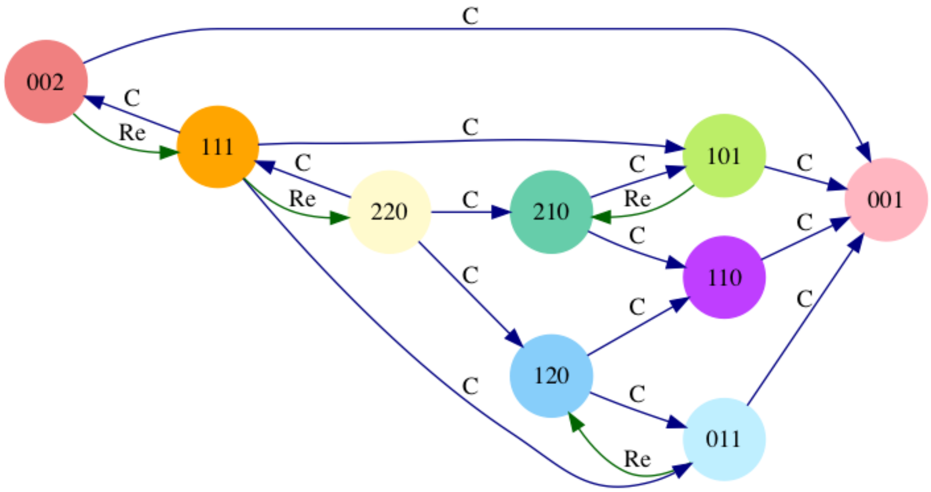


FIGURE 5. State space with transitions for $n = 2$ with recombination but no population substructure (recreation of Figure 3 from Simonsen and Churchill (1997)) Each triplet ijk gives the number of left, right, and double lineages in a state respectively. C indicates a coalescent event to transition between states and Re represents a recombination event to transition between states. The process begins in state 002 and proceeds to coalescence in state 001

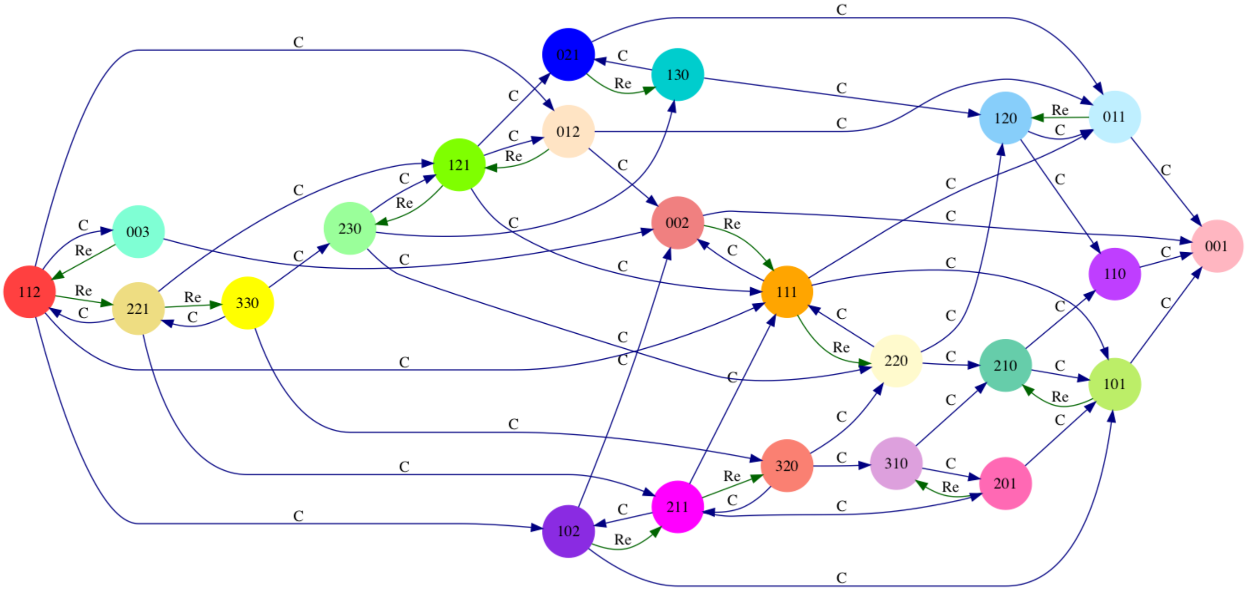


FIGURE 6. The model of Simonsen and Churchill (1997) with $n = 3$. The model has recombination but no population substructure. The 9 states of the $n = 2$ case also appear in the model with $n = 3$. The colors of the states are consistent between this figure and Figure 5, to illustrate how the $n = 2$ chain is embedded in the state space of $n = 3$.

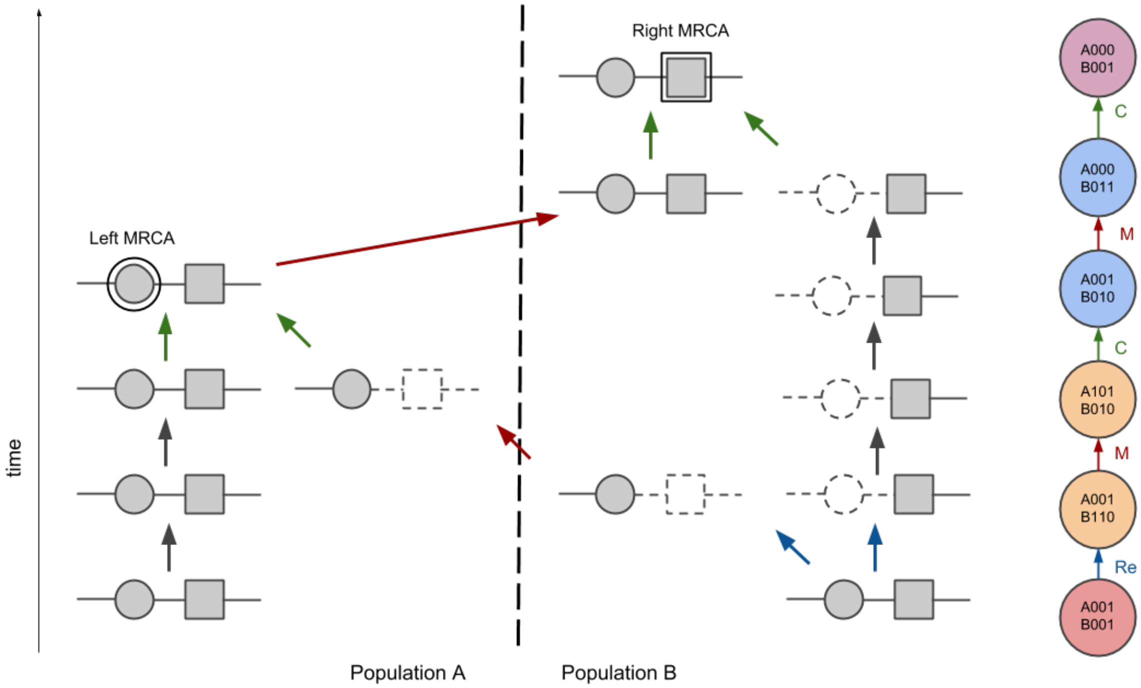


Figure 7. An illustrative example of a process that traces the lineages of two loci for two individuals in two populations. Recombination, migration, and coalescence can change the number of each lineage type in each population. In this case, the tree associated with left locus is shorter than the tree associated with the right locus.

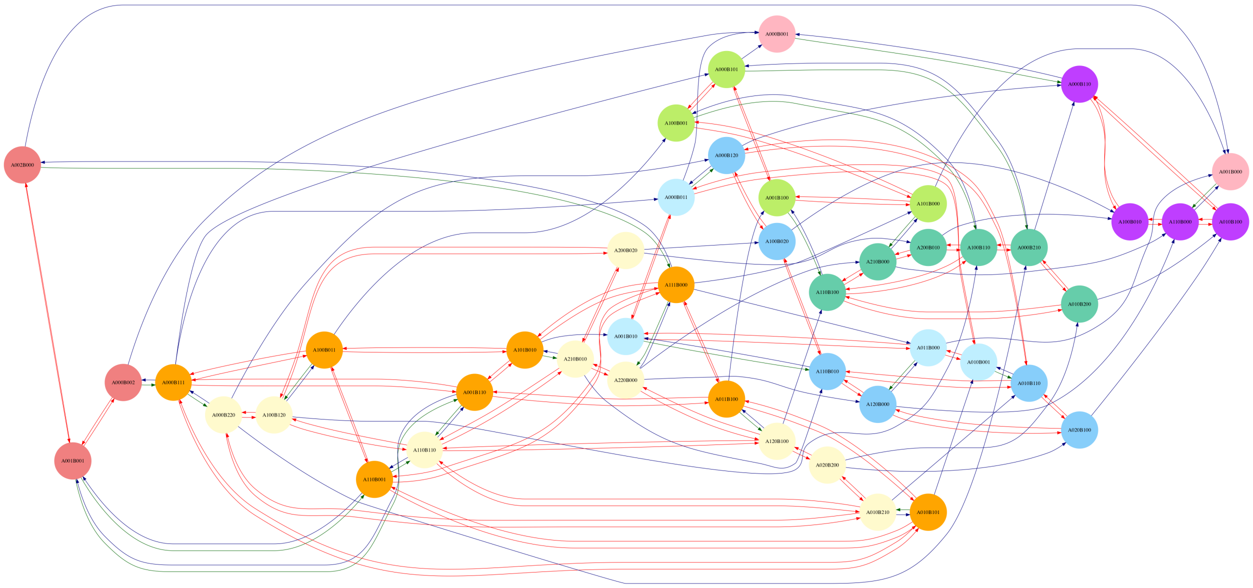


FIGURE 8. Full state space and transitions for $n = 2$, model with recombination and migration. The number of states is 46. Coalescence events are represented by blue lines. Recombination events are represented by green lines. Migration events are represents by red lines. The colors of states corresponds to the states in the recombination only model if the population labels are removed.

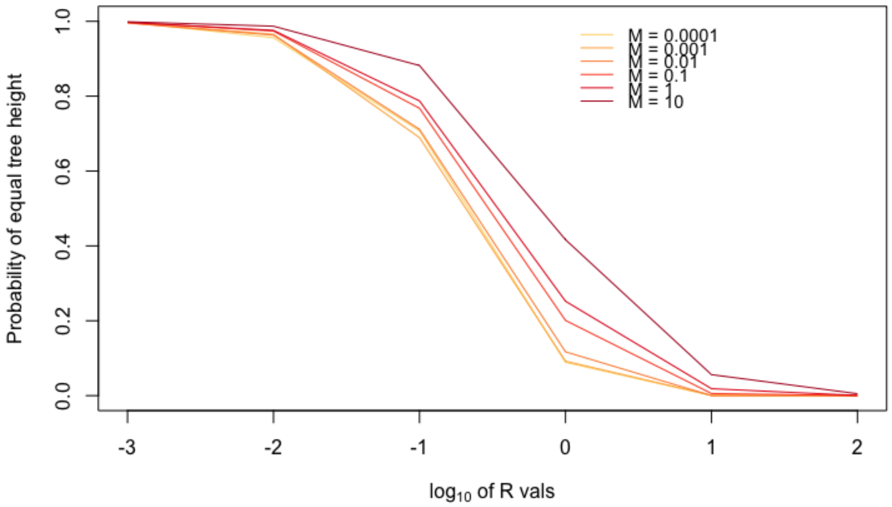


FIGURE 9. Probability of equal tree heights for different migration rates, $2Nm = M$ and recombination rates, $2Nr = R$. estimated from 10^4 simulations of the Markov process