Expected patterns of local ancestry in a hybrid zone

Joel Smith[1*], Bret Payseur[2], John Novembre[1,3]

1 Department of Ecology and Evolution, University of Chicago, Chicago, IL

2 Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI

3 Department of Human Genetics, University of Chicago, Chicago, IL

*joelsmith@uchicago.edu

1 Abstract

The initial drivers of reproductive isolation between species are poorly characterized. In cases where partial reproductive isolation exists, genomic patterns of variation in hybrid zones may provide clues about the barriers to gene flow which arose first during the early stages of speciation. Purifying selection against incompatible substitutions that reduce hybrid fitness has the potential to distort local patterns of ancestry relative to background patterns across the genome. The magnitude and qualitative properties of this pattern are dependent on several factors including migration history and the relative fitnesses for different combinations of incompatible alleles. We present a model which may account for these factors and highlight the potential for its use in verifying the action of natural selection on candidate loci implicated in reducing hybrid fitness.

2 Introduction

A large fraction of research aiming to describe the process of speciation involves mapping genetic variants responsible for reproductive isolation. Despite its difficulty, this task has nevertheless been carried out for a number of cases in which the link between a reproductive isolating mechanism mapped in a laboratory setting and its effect on an individual's fitness in nature is demonstrated [Schluter, 2009]. However, in many of these cases, reproductive isolation is already complete such that the initial cause of speciation cannot be attributed to any one locus or set of loci due to a lack of information regarding the order in which these isolating barriers arose [Turelli et al., 2014]. Hybrid zones present a convenient situation where reproductive isolation is incomplete. In these cases, the mechanisms of reproductive isolation are both fewer and more recently derived. Relative to scenarios with complete reproductive isolation, systems with ongoing hybridization may provide a more narrow set of candidate loci to consider as the initial drivers of speciation.

The next task would be to describe the mechanism by which the incompatible substitutions were fixed. Functional annotations for the implicated loci can yield some clues about the ecological context or genetic causes that resulted in these substitutions. A rigorously tested explanation would require that field experiments be carried out to establish their effect on fitness in nature [Schemske, 2000, Schemske and Bradshaw, 1999]. However, patterns of genomic variation can provide a complementary source of evidence for the action of natural selection on genetic variants which are relevant to a phenotype of interest [Tiffin and Ross-Ibarra, 2014]. The robustness of any given metric or model for the signature of natural selection depends on well-conceived theory that describes both the conditions under which the signature is detectable as well as any non-selective processes that can explain the pattern. This observational approach has been a driver of both theoretical and empirical research which aims to implicate loci responsible for genetic incompatibilities that decrease fitness among hybrids in nature [Barton, 1979, Barton and Hewitt, 1985, Endler, 1973, White, 1968].

Hybrid zones are thought to present a useful situation where the interaction between gene flow and natural selection can leave identifiable patterns associated with genetic incompatibilities in genomic data [Harrison and Larson, 2016, Payseur, 2010, Payseur and Rieseberg, 2016]. Historically, most work on this problem has relied on using differences in allele frequencies across the hybrid zone while ignoring patterns of linkage disequilibrium among neighboring sites [Barton and Hewitt, 1985]. More recently, increased access to sequencing technology has prompted the use of methods which can infer local ancestry across the genomes of admixed individuals [Gompert and Buerkle, 2013]. In this regard, population genetic inference has made a significant shift toward developing models which leverage this information for a variety of purposes. Several models aim to infer the migration history between genetically distinct populations using the length of ancestry tract lengths among admixed individuals [Gravel, 2012, Harris and Nielsen, 2013, Hellenthal et al., 2014, Liang and Nielsen, 2014, Loh et al., 2013, Patterson et al., 2012, Pool and Nielsen, 2009, Price et al., 2009, Sedghifar et al., 2015]. As the primary intention of these approaches has been to focus on populations within a species, there is a lack of work which aims to describe the effect of genetic incompatibilities which commonly arise between species after a prolonged period of geographic isolation.

Theory with formal treatment of genetic incompatibilities and ancestry tracts has been slow to accumulate, in large part due to the large parameter space of both migration histories and genetic architectures that may contribute to reduced fitness in hybrid individuals. As a result, forward simulations of whole chromosomes under differing migration and selection regimes have been used to describe some general patterns [Gompert et al., 2012, Hvala et al., 2018, Lindtke and Buerkle, 2015, Schumer and Brandvain, 2016]. In a few of these cases, the primary goal is to describe the conditions

which may account for the heterogeneous patterns of genomic differentiation which have been widely observed across hybrid zones [Harrison and Larson, 2016]. For example, Gompert et al. [2012] focus on describing differences in both the number of contributing loci and the mechanism of their effect through either underdominance at single loci or two-locus epistasis. They also introduce a formalized approach to identify outlier loci responsible for reduced hybrid fitness using allele frequency clines across the genome. Lindtke and Buerkle [2015] pay particular attention to two-locus models of genetic incompatibilities and compare the relative efficiency with which different kinds of epistatic interactions can maintain genomic differentiation in a hybrid zone under both high and low migration.

In an effort to make use of ancestry tract lengths rather than allele frequencies at individual loci, Sedghifar et al. [2015] derive a null expectation for the length of ancestry tracts in a geographic context where distance from the contact zone of two genetically distinct populations is explicitly modeled. They then provide a likelihood function which they use to infer the age of the contact zone, or time at which admixture between the populations began. Sedghifar et al. [2016] extends this spatially-explicit framework further to model the mean ancestry tract length which is contiguous with an under-dominant locus.

Another approach that uses local ancestry inference to identify genetic incompatibilities relies on computing correlations in ancestry among pairs of loci in a hybrid zone [Schumer et al., 2014]. Schumer and Brandvain [2016] use simulation to demonstrate how selection against incompatible alleles at two loci can lead to a positive correlation in species ancestry at those loci. They find good power to identify these associations for genetic architectures that feature ubiquitous selection (see Figure 1d). The intuition for this pattern is that genotypes with the same ancestry at both loci are the only genotypes with high fitness, such that an over-representation of ancestry at those loci relative to background levels of linkage disequilibrium (LD) should lead to an identifiable signal. For genetic architectures that only feature strong selection against derived allele combinations, they find much less power to identify significant pairs.

The variety of approaches and data available to study this problem have prompted a few questions of where to proceed next. We first describe a few of the well-studied genetic architectures for two-locus genetic incompatibilities as well as others that have received less attention but which have also been identified in nature. We then present a model to compute the expected distribution of ancestry tract lengths around incompatibility loci.

2.1 Two-Locus Genetic Incompatibilities

The two-locus fitness matrix provides a useful representation of different genetic architectures which might contribute to genetic incompatibility between species (Figure 1 and Table 1). In addition to theoretical arguments and simulations, much of our current understanding for how relevant any of these genetic architectures might be in nature has been driven by genetic dissection of reproductive barrier phenotypes in the lab [White et al., 2011]. There are a number of empirical examples in a variety of species which have hinted at the potential importance of meiotic drive and neutral (or nearly neutral) causes for the fixation of incompatible substitutions [Maheshwari and Barbash, 2011, Presgraves, 2010, Sweigart and Willis, 2012]. While the precise combination of evolutionary forces which are responsible for incompatibility formation remain unknown, the evolution of incompatibilities in hybrid populations can be reasonably approximated with simple epistatis [Schumer et al., 2014].

		bb ($\mathbf{B_1B_1}$)	**Bb ($\mathbf{B_1B_2}$)**	**BB ($\mathbf{B_2B_2}$)**
aa	($\mathbf{A_1A_1}$)	1	$1 - s_a h_a$	$1 - s_a$
Aa	($\mathbf{A_1A_2}$)	$1 - s_e h_1$	$1 - s_e h_0$	$1 - s_a h_a$
AA	($\mathbf{A_2A_2}$)	$1 - s_e$	$1 - s_e h_1$	1

Table 1. Genotype fitnesses for the DMI and symmetric incompatibility models. The first pairs of bold letters are DMI model genotypes and the genotypes in parentheses indicate the symmetric model. s_a and s_e denote the selection coefficient against the ancestral and incompatible alleles, respectively. h_a, h_0 and h_1 denote the dominance effects of ancestral, double-heterozygotes and single-heterozygotes, respectively.

The most well-known model is described in Dobzhansky [1937] in which alleles fix at two interacting loci among populations that are geographically isolated. The top row in Figure 1 shows a range of possible fitness matrices that might result from this scenario, also known as the Dobzhansky-Muller incompatibility model (DMI). If we denote the ancestral genotype as **aaBB** in all of these cases, then the derived genotypes before coming into secondary contact are **aabb** and **AABB**. We chose these example matrices to emphasize the diversity of fitness configurations that might result from this model. The fitness matrix in Figure 1a is an example where the the derived substitutions were fixed by positive selection, such that the ancestral genotype suffers a fitness cost. Figures 1a and 1b are examples where the derived alleles interact dominantly; whereas in Figure 1c, derived alleles interact recessively.

Lindtke and Buerkle [2015] draw attention to a different model of genetic incompatibility where allele substitutions occur at two loci in both populations leading to a symmetric pattern of fitnesses

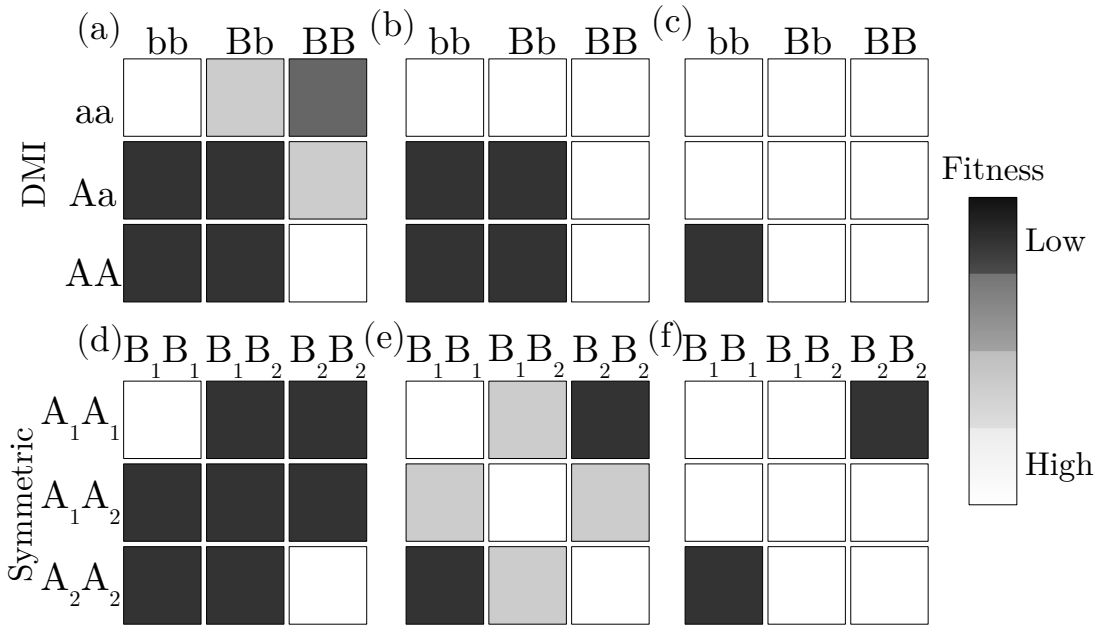

Fig 1. Two-locus fitness matrices for six models of genetic incompatibility. Each matrix includes the fitnesses of all possible two-locus genotypes where each locus is biallelic. Shaded boxes represent genotypes with a fitness cost that varies positively with the amount of shading. The top row of matrices are variations of the DMI model with the **aaBB** genotype representing the ancestral state and the bottom row shows variations of a symmetric incompatibility model. For both rows, the dominance effect of derived substitutions decreases from left to right.

between the two derived genotypes $\mathbf{A_1A_1B_1B_1}$ and $\mathbf{A_2A_2B_2B_2}$ (Figure 1d, 1e, 1f). Their results suggest that this mechanism could provide a better explanation for the observed patterns of genetic differentiation that occur at extended genomic distances between species that hybridize [Harrison and Larson, 2016]. Regulatory interactions between a transcription factor encoded at one locus and the corresponding binding site at a second locus would be one scenario consistent with this model. Seehausen et al. [2014] note that this model could also be common in meiotic drive scenarios where a substitution that promotes biased transmission of a selfish genetic element at one locus is counteracted by a substitution at a second locus which restores unbiased inheritance. The bottom row in Figure 1 shows a range of possible fitness matrices under this model, where the left-most matrix results from dominant substitutions which interfere between haplotypes, and the right most matrix results from recessive substitutions. Simulated data in Figure 2 (using the software dfuse from Lindtke and Buerkle [2015]) illustrates the effect of the DMI model in Figure 1b where selection against derived alleles leads

to a bias towards the ancestral genotype (**aaBB**) of recombined ancestries.

In the following section, we first review an approach taken by Gravel [2012] to model the distribution of ancestry tract lengths across the genomes of an admixed population. We then describe the framework for our own extension to this approach which aims to model the distribution of ancestry tract lengths that are contiguous with a locus undergoing epistatic interactions according to any of the incompatibility scenarios outlined above.

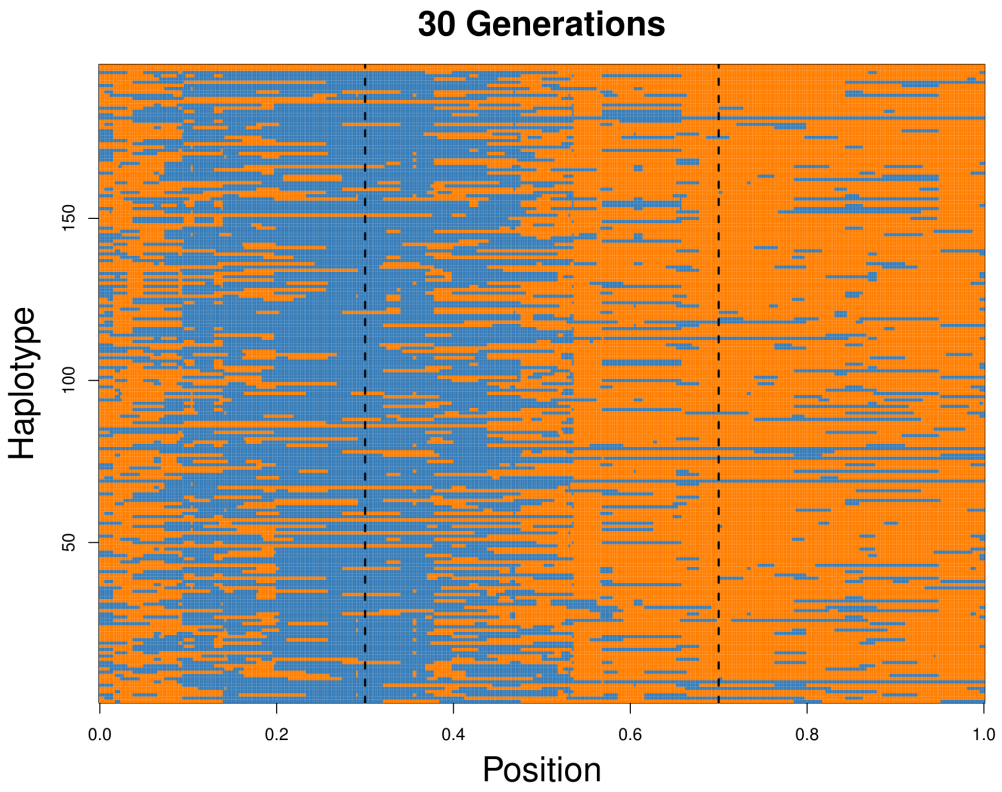

Fig 2. Haplotype data simulated using the software dfuse with the fitness matrix in Figure 1b. The forward-in-time simulation begins with two infinite source populations contributing equal fractions of ancestry (0.5) to a target population of 100 individuals 30 generations in the past. Each generation to the present follows a Wright-Fisher model, whereby both source populations contribute a fraction of individuals m to the target population. In this case $m = 0.1$. Recombination occurs uniformly along the chromosome at rate 1 crossover per chromosome per generation. After recombination, individuals are removed from the population according to a specified fitness matrix. The parameter values defined in Table 1 take the following values: $s_a = 0$, $s_e = 0.9$, $h_1 = 1$, $h_0 = 1$, and $h_a = 0$. The incompatibility loci are indicated by the vertical dotted lines.

3 Model Description

3.1 Tract Length Distributions Under Neutral Admixture

Gravel [2012] defines a Markov chain along a chromosome with transition rates between both an ancestry state variable, p, and the time, t, at which ancestry p arrives in a hybrid population. Consider the demography of a sample up to the first hybridization event T generations ago, where each generation is labeled $s \in \{0, 1, 2, ..., T-1\}$. Let $m_p(t)$ denote the fraction of individuals in the target population replaced by individuals from source population p at time t. $m(t)$ is the total fraction of individuals in the target population replaced by migrants in generation t where $\sum_p m_p(t) \leq 1$. Moving along a chromosome from any point, the probability of encountering state (p, t) after a recombination event that occured at generation τ is

$$P(p, t|\tau) = m_p(t) \prod_{t'=\tau+1}^{t-1} (1 - m(t')). \tag{1}$$

τ is uniformly distributed on $(1, t-1)$, so the discrete transition probabilites can be expressed as

$$R(p, t \to p', t') = \sum_{\tau=1}^{min(t,t')-1} \frac{P(p', t'|\tau)}{(t-1)} \tag{2}$$

To get the continuous transition rate, one can multiply the discrete transition rate by the continuous overall transition rate $t-1$. This follows from the fact that a recombination event occurs at each generation such that probability of observing an ancestry junction depends on the number of generations since admixture:

$$Q(p, t \to p', t') = m_{p'}(t') \sum_{\tau=1}^{min(t,t')-1} \prod_{s=\tau+1}^{t'-1} (1 - m(s)). \tag{3}$$

Using Q, one can compute the tract length distribution for a given ancestry. Q is first uniformized to adjust self-transition probabilities such that the total transition rate from each state is equal to the rate of the state with the highest transition rate, Q_0 [Stewart, 1994]. One can then compute the distribution of the number of steps spent in a particular ancestry, $\{b_n\}_{n=1,...,\Lambda}$, up to a cutoff Λ, where $\sum_{i=1}^{\Lambda} b_i \approx 1$. $\{b_n\}_{n=1,...,\Lambda}$ is computed by multiplying the state vector with the transition matrix for Λ iterations while recording the amount of probability absorbed by the non-p ancestries at each step. The Erlang distribution models the length of a trajectory, l, with k steps as:

$$\mathbb{E}_{k,Q_0}(l) = \frac{Q_0^k l^{k-1} e^{-Q_0 l}}{(k-1)!} \tag{4}$$

This leads to the tract length distribution:

$$\phi(l) = \sum_{k=1}^{\Lambda+1} b_k \, \mathbb{E}_{k,Q_0}(l) \tag{5}$$

3.2 A Locus-Specific Tract Length Distribution With Selection

Equation 5 describes the length of tracts in a way that is not locus specific. We are interested in how the effects of purifying selection against alleles at two loci under negative selection, according to the incompatibility models described above, may skew the tract length distribution. More specifically, we want to model the distribution of ancestry tracts lengths that are contiguous with a negatively selected allele on a chromosome. In this case, the probability of observing a transition, or recombination event, depends on its recombination distance from the incompatibility loci of interest.

We define the number of basepairs between loci A and B to be $v + w = L$, where v is the number of basepairs from the A locus to the vth position and w is the number of basepairs from position $v+1$ to L (Figure 3). We extend the transition matrix Q in equation (3) such that each value of v denotes a new Q_v by multiplying each transition rate by the probability, Ψ_v^τ, that an ancestry junction which arises at time τ at position v survives to the present:

$$Q_v(p, t \to p', t') = m_{p'}(t') \sum_{\tau=1}^{min(t,t')-1} \Psi_v^\tau \prod_{s=\tau+1}^{t'-1} (1 - m(s)). \tag{6}$$

Equation 6 is computed as a function of the sequence of genotypic backgrounds the junction encounters each generation to the present. Using a two-allele model, let **A** and **a** refer to alternative alleles at the locus of interest, and alleles **B** and **b** refer to the second locus located at some distance away from the **A** locus. We can define a state space, S, of two-locus genotypes in which the junction can exist:

$$S = \begin{bmatrix} \mathbf{AB}|ab \\ \mathbf{AB}|Ab \\ \mathbf{AB}|aB \\ \mathbf{AB}|AB \\ \mathbf{Ab}|ab \\ \mathbf{Ab}|Ab \\ \mathbf{Ab}|aB \\ \mathbf{Ab}|AB \\ \mathbf{aB}|ab \\ \mathbf{aB}|Ab \\ \mathbf{aB}|aB \\ \mathbf{aB}|AB \\ \mathbf{ab}|ab \\ \mathbf{ab}|Ab \\ \mathbf{ab}|aB \\ \mathbf{ab}|AB \\ \epsilon \end{bmatrix}$$

where the bold pair of alleles refers to the chromosome on which the junction resides. In cases where the interacting loci are on different chromosomes, the bold alleles refer to the genomic complement from which the junction is inherited.

Let $\mathbf{P}_v^{t,t-1}$ be a symmetric 17×17 transition matrix among the states in S from time t to $t-1$ for a junction at the vth position. The transition probabilities in $\mathbf{P}_v^{t,t-1}$ depend on the fitness of genotypes carrying the junction, ω, the recombination rate between the interacting loci, r, and the frequency of possible gametes with which to pair in the hybrid population at time $t-1$: x_1^{t-1}, x_2^{t-1}, x_3^{t-1}, x_4^{t-1}. Let x_1, x_2, x_3, x_4 refer to the frequencies of gametes **AB**, **Ab**, **aB** and **ab**, respectively. Gamete frequencies are computed numerically by simulation [Gavrilets 1997, Appendix A.5]. Let ω_i denote the marginal fitness of gamete i where ω_1, ω_2, ω_3, ω_4 refer to gametes **AB**, **Ab**, **aB** and **ab**, respectively. Let ω_{ij} refer to the fitness of an individual with gametes i and j. Figure 3 provides some intuition for how the following transition probabilities in $\mathbf{P}_v^{t,t-1}$ are computed.

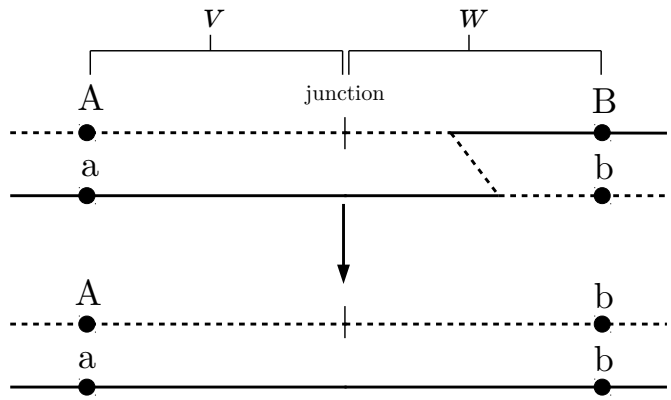

Fig 3. A visual description of the transition probability $\mathbf{P}_{1,5}^{t,t-1}$. For the first state in S, $\mathbf{AB}|ab$, the transition probability to state $\mathbf{Ab}|ab$, is a product of the probability that the bold haplotype ($\mathbf{AB}$) is chosen (0.5), a recombination event occurs between the junction and locus B, $r\frac{w}{v+w}$, the recombined gamete gets paired with gamete x_4 at time $t-1$, and the individual with genotype $\mathbf{Ab}|ab$ survives, ω_{14}.

$$\mathbf{P}_{1,j}^{t,t-1} = \begin{cases} .5\omega_{14}(1-r)x_4^{t-1} & \text{if } j=1; \\ .5\omega_{14}(1-r)x_2^{t-1} & \text{if } j=2; \\ .5\omega_{14}(1-r)x_3^{t-1} & \text{if } j=3; \\ .5\omega_{14}(1-r)x_1^{t-1} & \text{if } j=4; \\ .5\omega_{14}r\frac{w}{v+w}x_4^{t-1} & \text{if } j=5; \\ .5\omega_{14}r\frac{w}{v+w}x_2^{t-1} & \text{if } j=6; \\ .5\omega_{14}r\frac{w}{v+w}x_3^{t-1} & \text{if } j=7; \\ .5\omega_{14}r\frac{w}{v+w}x_1^{t-1} & \text{if } j=8 \\ .5\omega_{14}r\frac{v}{v+w}x_4^{t-1} & \text{if } j=9; \\ .5\omega_{14}r\frac{v}{v+w}x_2^{t-1} & \text{if } j=10; \\ .5\omega_{14}r\frac{v}{v+w}x_3^{t-1} & \text{if } j=11; \\ .5\omega_{14}r\frac{v}{v+w}x_1^{t-1} & \text{if } j=12 \\ 0 & \text{if } j=13; \\ 0 & \text{if } j=14; \\ 0 & \text{if } j=15; \\ 0 & \text{if } j=16 \\ 1-\sum_j^{16}\mathbf{P}_{1,j}^{t,t-1} & \text{if } j=17 \end{cases} \tag{7}$$

$$\mathbf{P}_{2,j}^{t,t-1} = \begin{cases} .5\omega_{12}(((1-r)x_4^{t-1}) + (r\frac{v}{v+w}x_4^{t-1})) & \text{if } j=1; \\ .5\omega_{12}(((1-r)x_2^{t-1}) + (r\frac{v}{v+w}x_2^{t-1})) & \text{if } j=2; \\ .5\omega_{12}(((1-r)x_3^{t-1}) + (r\frac{v}{v+w}x_3^{t-1})) & \text{if } j=3; \\ .5\omega_{12}(((1-r)x_1^{t-1}) + (r\frac{v}{v+w}x_1^{t-1})) & \text{if } j=4; \\ .5\omega_{12}r\frac{w}{v+w}x_4^{t-1} & \text{if } j=5; \\ .5\omega_{12}r\frac{w}{v+w}x_2^{t-1} & \text{if } j=6; \\ .5\omega_{12}r\frac{w}{v+w}x_3^{t-1} & \text{if } j=7; \\ .5\omega_{12}r\frac{w}{v+w}x_1^{t-1} & \text{if } j=8 \\ 0 & \text{if } j=9; \\ 0 & \text{if } j=10; \\ 0 & \text{if } j=11; \\ 0 & \text{if } j=12 \\ 0 & \text{if } j=13; \\ 0 & \text{if } j=14; \\ 0 & \text{if } j=15; \\ 0 & \text{if } j=16 \\ 1-\sum_j^{16}\mathbf{P}_{2,j}^{t,t-1} & \text{if } j=17 \end{cases} \tag{8}$$

$$\mathbf{P}_{3,j}^{t,t-1} = \begin{cases} .5\omega_{13}(((1-r)x_4^{t-1}) + (r\frac{w}{v+w}x_4^{t-1})) & \text{if } j=1; \\ .5\omega_{13}(((1-r)x_2^{t-1}) + (r\frac{w}{v+w}x_2^{t-1})) & \text{if } j=2; \\ .5\omega_{13}(((1-r)x_3^{t-1}) + (r\frac{w}{v+w}x_3^{t-1})) & \text{if } j=3; \\ .5\omega_{13}(((1-r)x_1^{t-1}) + (r\frac{w}{v+w}x_1^{t-1})) & \text{if } j=4; \\ 0 & \text{if } j=5; \\ 0 & \text{if } j=6; \\ 0 & \text{if } j=7; \\ 0 & \text{if } j=8; \\ .5\omega_{13}r\frac{v}{v+w}x_4^{t-1} & \text{if } j=9; \\ .5\omega_{13}r\frac{v}{v+w}x_2^{t-1} & \text{if } j=10; \\ .5\omega_{13}r\frac{v}{v+w}x_3^{t-1} & \text{if } j=11; \\ .5\omega_{13}r\frac{v}{v+w}x_1^{t-1} & \text{if } j=12 \\ 0 & \text{if } j=13; \\ 0 & \text{if } j=14; \\ 0 & \text{if } j=15; \\ 0 & \text{if } j=16 \\ 1-\sum_j^{16}\mathbf{P}_{3,j}^{t,t-1} & \text{if } j=17 \end{cases} \tag{9}$$

$$\mathbf{P}_{4,j}^{t,t-1} = \begin{cases} .5\omega_{11}(((1-r)x_4^{t-1}) + (rx_4^{t-1})) & \text{if } j = 1; \\ .5\omega_{11}(((1-r)x_2^{t-1}) + (rx_2^{t-1})) & \text{if } j = 2; \\ .5\omega_{11}(((1-r)x_3^{t-1}) + (rx_3^{t-1})) & \text{if } j = 3; \\ .5\omega_{11}(((1-r)x_1^{t-1}) + (rx_1^{t-1})) & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{4,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{10}$$

$$\mathbf{P}_{5,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ .5\omega_{24}(((1-r)x_4^{t-1}) + (r\frac{w}{v+w}x_4^{t-1})) & \text{if } j = 5; \\ .5\omega_{24}(((1-r)x_2^{t-1}) + (r\frac{w}{v+w}x_2^{t-1})) & \text{if } j = 6; \\ .5\omega_{24}(((1-r)x_3^{t-1}) + (r\frac{w}{v+w}x_3^{t-1})) & \text{if } j = 7; \\ .5\omega_{24}(((1-r)x_1^{t-1}) + (r\frac{w}{v+w}x_1^{t-1})) & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ .5\omega_{24}r\frac{v}{v+w}x_4^{t-1} & \text{if } j = 13; \\ .5\omega_{24}r\frac{v}{v+w}x_2^{t-1} & \text{if } j = 14; \\ .5\omega_{24}r\frac{v}{v+w}x_3^{t-1} & \text{if } j = 15; \\ .5\omega_{24}r\frac{v}{v+w}x_1^{t-1} & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{5,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{11}$$

$$\mathbf{P}_{6,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ .5\omega_{22}(((1-r)x_4^{t-1}) + (rx_4^{t-1})) & \text{if } j = 5; \\ .5\omega_{22}(((1-r)x_2^{t-1}) + (rx_2^{t-1})) & \text{if } j = 6; \\ .5\omega_{22}(((1-r)x_3^{t-1}) + (rx_3^{t-1})) & \text{if } j = 7; \\ .5\omega_{22}(((1-r)x_1^{t-1}) + (rx_1^{t-1})) & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{6,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{12}$$

$$\mathbf{P}_{7,j}^{t,t-1} = \begin{cases} .5\omega_{23} r \frac{w}{v+w} x_4^{t-1} & \text{if } j = 1; \\ .5\omega_{23} r \frac{w}{v+w} x_2^{t-1} & \text{if } j = 2; \\ .5\omega_{23} r \frac{w}{v+w} x_3^{t-1} & \text{if } j = 3; \\ .5\omega_{23} r \frac{w}{v+w} x_1^{t-1} & \text{if } j = 4; \\ .5\omega_{23}(1-r)x_4^{t-1} & \text{if } j = 5; \\ .5\omega_{23}(1-r)x_2^{t-1} & \text{if } j = 6; \\ .5\omega_{23}(1-r)x_3^{t-1} & \text{if } j = 7; \\ .5\omega_{23}(1-r)x_1^{t-1} & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ .5\omega_{23} r \frac{v}{v+w} x_4^{t-1} & \text{if } j = 13; \\ .5\omega_{23} r \frac{v}{v+w} x_2^{t-1} & \text{if } j = 14; \\ .5\omega_{23} r \frac{v}{v+w} x_3^{t-1} & \text{if } j = 15; \\ .5\omega_{23} r \frac{v}{v+w} x_1^{t-1} & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{7,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{13}$$

$$\mathbf{P}_{8,j}^{t,t-1} = \begin{cases} .5\omega_{12} r \frac{w}{v+w} x_4^{t-1} & \text{if } j = 1; \\ .5\omega_{12} r \frac{w}{v+w} x_2^{t-1} & \text{if } j = 2; \\ .5\omega_{12} r \frac{w}{v+w} x_3^{t-1} & \text{if } j = 3; \\ .5\omega_{12} r \frac{w}{v+w} x_1^{t-1} & \text{if } j = 4; \\ .5\omega_{12}(((1-r)x_4^{t-1}) + (r \frac{v}{v+w} x_4^{t-1})) & \text{if } j = 5; \\ .5\omega_{12}(((1-r)x_2^{t-1}) + (r \frac{v}{v+w} x_2^{t-1})) & \text{if } j = 6; \\ .5\omega_{12}(((1-r)x_3^{t-1}) + (r \frac{v}{v+w} x_3^{t-1})) & \text{if } j = 7; \\ .5\omega_{12}(((1-r)x_1^{t-1}) + (r \frac{v}{v+w} x_1^{t-1})) & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{8,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{14}$$

$$\mathbf{P}_{9,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ .5\omega_{34}(((1-r)x_4^{t-1}) + (r \frac{v}{v+w} x_4^{t-1})) & \text{if } j = 9; \\ .5\omega_{34}(((1-r)x_2^{t-1}) + (r \frac{v}{v+w} x_2^{t-1})) & \text{if } j = 10; \\ .5\omega_{34}(((1-r)x_3^{t-1}) + (r \frac{v}{v+w} x_3^{t-1})) & \text{if } j = 11; \\ .5\omega_{34}(((1-r)x_1^{t-1}) + (r \frac{v}{v+w} x_1^{t-1})) & \text{if } j = 12; \\ .5\omega_{34} r \frac{w}{v+w} x_4^{t-1} & \text{if } j = 13; \\ .5\omega_{34} r \frac{w}{v+w} x_2^{t-1} & \text{if } j = 14; \\ .5\omega_{34} r \frac{w}{v+w} x_3^{t-1} & \text{if } j = 15; \\ .5\omega_{34} r \frac{w}{v+w} x_1^{t-1} & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{9,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{15}$$

$$\mathbf{P}_{10,j}^{t,t-1} = \begin{cases} .5\omega_{23} r \frac{v}{v+w} x_4^{t-1} & \text{if } j = 1; \\ .5\omega_{23} r \frac{v}{v+w} x_2^{t-1} & \text{if } j = 2; \\ .5\omega_{23} r \frac{v}{v+w} x_3^{t-1} & \text{if } j = 3; \\ .5\omega_{23} r \frac{v}{v+w} x_1^{t-1} & \text{if } j = 4 \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ .5\omega_{23}(1-r) x_4^{t-1} & \text{if } j = 9; \\ .5\omega_{23}(1-r) x_2^{t-1} & \text{if } j = 10; \\ .5\omega_{23}(1-r) x_3^{t-1} & \text{if } j = 11; \\ .5\omega_{23}(1-r) x_1^{t-1} & \text{if } j = 12; \\ .5\omega_{23} r \frac{w}{v+w} x_4^{t-1} & \text{if } j = 13; \\ .5\omega_{23} r \frac{w}{v+w} x_2^{t-1} & \text{if } j = 14; \\ .5\omega_{23} r \frac{w}{v+w} x_3^{t-1} & \text{if } j = 15; \\ .5\omega_{23} r \frac{w}{v+w} x_1^{t-1} & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{10,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{16}$$

$$\mathbf{P}_{11,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ .5\omega_{33}(((1-r)x_4^{t-1}) + (r x_4^{t-1})) & \text{if } j = 9; \\ .5\omega_{33}(((1-r)x_2^{t-1}) + (r x_2^{t-1})) & \text{if } j = 10; \\ .5\omega_{33}(((1-r)x_3^{t-1}) + (r x_3^{t-1})) & \text{if } j = 11; \\ .5\omega_{33}(((1-r)x_1^{t-1}) + (r x_1^{t-1})) & \text{if } j = 12; \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{11,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{17}$$

$$\mathbf{P}_{12,j}^{t,t-1} = \begin{cases} .5\omega_{13} r \frac{v}{v+w} x_4^{t-1} & \text{if } j = 1; \\ .5\omega_{13} r \frac{v}{v+w} x_2^{t-1} & \text{if } j = 2; \\ .5\omega_{13} r \frac{v}{v+w} x_3^{t-1} & \text{if } j = 3; \\ .5\omega_{13} r \frac{v}{v+w} x_1^{t-1} & \text{if } j = 4 \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ .5\omega_{13}(((1-r)x_4^{t-1}) + (r\frac{w}{v+w}x_4^{t-1})) & \text{if } j = 9; \\ .5\omega_{13}(((1-r)x_2^{t-1}) + (r\frac{w}{v+w}x_2^{t-1})) & \text{if } j = 10; \\ .5\omega_{13}(((1-r)x_3^{t-1}) + (r\frac{w}{v+w}x_3^{t-1})) & \text{if } j = 11; \\ .5\omega_{13}(((1-r)x_1^{t-1}) + (r\frac{w}{v+w}x_1^{t-1})) & \text{if } j = 12; \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{12,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{18}$$

$$\mathbf{P}_{13,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ .5\omega_{44}(((1-r)x_4^{t-1}) + (rx_4^{t-1})) & \text{if } j = 13; \\ .5\omega_{44}(((1-r)x_2^{t-1}) + (rx_2^{t-1})) & \text{if } j = 14; \\ .5\omega_{44}(((1-r)x_3^{t-1}) + (rx_3^{t-1})) & \text{if } j = 15; \\ .5\omega_{44}(((1-r)x_1^{t-1}) + (rx_1^{t-1})) & \text{if } j = 16; \\ 1 - \sum_j^{16} \mathbf{P}_{13,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{19}$$

$$\mathbf{P}_{14,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ .5\omega_{24} r \frac{v}{v+w} x_4^{t-1} & \text{if } j = 5; \\ .5\omega_{24} r \frac{v}{v+w} x_2^{t-1} & \text{if } j = 6; \\ .5\omega_{24} r \frac{v}{v+w} x_3^{t-1} & \text{if } j = 7; \\ .5\omega_{24} r \frac{v}{v+w} x_1^{t-1} & \text{if } j = 8 \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ .5\omega_{24}(((1-r)x_4^{t-1}) + (r \frac{w}{v+w} x_4^{t-1})) & \text{if } j = 13; \\ .5\omega_{24}(((1-r)x_2^{t-1}) + (r \frac{w}{v+w} x_2^{t-1})) & \text{if } j = 14; \\ .5\omega_{24}(((1-r)x_3^{t-1}) + (r \frac{w}{v+w} x_3^{t-1})) & \text{if } j = 15; \\ .5\omega_{24}(((1-r)x_1^{t-1}) + (r \frac{w}{v+w} x_1^{t-1})) & \text{if } j = 16; \\ 1 - \sum_j^{16} \mathbf{P}_{14,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{20}$$

$$\mathbf{P}_{15,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ .5\omega_{34} r \frac{w}{v+w} x_4^{t-1} & \text{if } j = 9; \\ .5\omega_{34} r \frac{w}{v+w} x_2^{t-1} & \text{if } j = 10; \\ .5\omega_{34} r \frac{w}{v+w} x_3^{t-1} & \text{if } j = 11; \\ .5\omega_{34} r \frac{w}{v+w} x_1^{t-1} & \text{if } j = 12 \\ .5\omega_{34}(((1-r)x_4^{t-1}) + (r \frac{v}{v+w} x_4^{t-1})) & \text{if } j = 13; \\ .5\omega_{34}(((1-r)x_2^{t-1}) + (r \frac{v}{v+w} x_2^{t-1})) & \text{if } j = 14; \\ .5\omega_{34}(((1-r)x_3^{t-1}) + (r \frac{v}{v+w} x_3^{t-1})) & \text{if } j = 15; \\ .5\omega_{34}(((1-r)x_1^{t-1}) + (r \frac{v}{v+w} x_1^{t-1})) & \text{if } j = 16; \\ 1 - \sum_j^{16} \mathbf{P}_{15,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{21}$$

$$\mathbf{P}_{16,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ .5\omega_{44} r \frac{v}{v+w} x_4^{t-1} & \text{if } j = 9; \\ .5\omega_{44} r \frac{v}{v+w} x_2^{t-1} & \text{if } j = 10; \\ .5\omega_{44} r \frac{v}{v+w} x_3^{t-1} & \text{if } j = 11; \\ .5\omega_{44} r \frac{v}{v+w} x_1^{t-1} & \text{if } j = 12 \\ .5\omega_{44} r \frac{w}{v+w} x_4^{t-1} & \text{if } j = 9; \\ .5\omega_{44} r \frac{w}{v+w} x_2^{t-1} & \text{if } j = 10; \\ .5\omega_{44} r \frac{w}{v+w} x_3^{t-1} & \text{if } j = 11; \\ .5\omega_{44} r \frac{w}{v+w} x_1^{t-1} & \text{if } j = 12 \\ .5\omega_{44}(1-r) x_4^{t-1} & \text{if } j = 13; \\ .5\omega_{44}(1-r) x_2^{t-1} & \text{if } j = 14; \\ .5\omega_{44}(1-r) x_3^{t-1} & \text{if } j = 15; \\ .5\omega_{44}(1-r) x_1^{t-1} & \text{if } j = 16; \\ 1 - \sum_j^{16} \mathbf{P}_{16,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{22}$$

$$\mathbf{P}_{17,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16; \\ 1 & \text{if } j = 17 \end{cases} \tag{23}$$

We can define the initial probabilities, π_0^τ, of a junction in each state when it occurs at a particular time

τ. These probabilities will vary depending on the ancestry of interest for the tract length distribution. Conditional on a recombination event occurring between the two loci, the probability that the junction occurs at any particular position is uniform ($1/L$). If the ancestry of interest is that of the A allele, then

$$\pi_0^\tau = \begin{bmatrix} 2p_A^\tau p_B^\tau x_1^\tau \\ 2p_A^\tau p_B^\tau x_2^\tau \\ 2p_A^\tau p_B^\tau x_3^\tau \\ 2p_A^\tau p_B^\tau x_4^\tau \\ 2p_A^\tau p_b^\tau x_1^\tau \\ 2p_A^\tau p_b^\tau x_2^\tau \\ 2p_A^\tau p_b^\tau x_3^\tau \\ 2p_A^\tau p_b^\tau x_4^\tau \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

where p_A^τ, p_a^τ, p_B^τ, p_b^τ are the allele frequencies at time τ. The probability that the junction resides among each of the states after its origination at time τ to the present is

$$\pi_v^\tau = \pi_0^\tau \prod_{t=0}^{\tau} \mathbf{P}_v^{t,t-1}. \tag{24}$$

After defining the vector $\eta = [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1]$, the survival probability of the junction is

$$\Psi_v^\tau = 1 - \pi_v^\tau \eta. \tag{25}$$

The transition matrix Q_v can now be computed using Equation 6 for all values of v where $v \in \{1...L\}$. In contrast to the transition matrix Q defined in Equation 3, the set of transition matrices Q_v are inhomogeneous over positions v. As a result, the uniformization technique outlined in Stewart

[1994] does not apply. However, Andreychenko [2010] describes an approach to uniformize a time-inhomogeneous Markov chain which relies on partitioning the transition matrix into time-dependent and time-independent components. Whereas the time-homogeneous case relies on uniformizing by the constant transition rate of the state with the largest value, the time-inhomogeneous case relies on using the average rate of the state with the largest transition rate value. As before, the distribution for the number of steps in a trajectory, $\{b_n\}_{n=1,\ldots,\Lambda}$, can be computed and used with Equations 4 and 5 to calculate the tract length distribution.

4 Discussion

The model presented above describes an approach which may prove useful in verifying the role of purifying selection against incompatible alleles in a hybrid zone. If shown to be robust under a reasonable set of demographic scenarios and genetic architectures for incompatibility, this model would provide an additional tool for testing the effects of selection on candidate loci which have been identified by QTL mapping of hybrid sterility or inviability traits [White et al., 2011]. This model could also be used to develop an independent test of loci identified by steep clines in allele frequency across a hybrid zone relative to the genomic background [Gompert et al., 2012]. We would like to emphasize the novelty and potential power of considering a genetic architecture of reproductive isolation for which there is strong empirical support (The Dobzhansky-Muller model) in the context of ancestry.

While any formal statements regarding the expected tract length distribution would require a full implementation of this model, we can make a few intuitive statements which follow from previous theory and simulation [Hvala et al., 2018, Lindtke and Buerkle, 2015]. In particular, Hvala et al. [2018] show that the number and density of ancestry junctions scales negatively with selection strength at incompatibility loci. This signature was further influenced by the genetic distance of junctions from the loci, the form of selection and dominance.

There are several challenges that remain before computing expected tract length distributions and performing inference on parameters of interest. In particular, computing Q_v for a large set of positions may be difficult considering the repeated summation over products in Equation 6, the matrix multiplication required both for Equation 24, and computing $\{b_n\}_{n=1,\ldots,\Lambda}$. While Gravel [2012] intended to model admixture events which occurred relatively recently, many hybrid zones of interest are likely to have formed more than 100 generations ago, which produces more computational burden given that the state space of Q_v is $2Tv$. However, it is likely that differences in the junction survival probability, Ψ_v^τ, beyond some value of τ become negligible. The simplified two-locus, two-allele model

that we consider is another effort to reduce the parameter space of genotype fitnesses that might result from higher-order epistasis of 3 or more loci.

Because Ψ_v^τ is dependent on hybrid zone gamete frequencies in a linear stepping-stone model, deviation from this simplifying assumption will most likely affect the results. The linear stepping-stone model which we borrow from Gavrilets [1997] can be generalized to any number of demes between the two infinite source populations. By implementing our model with this population structure, one could compute tract length distributions as a function of distance from the hybrid zone in a similar spirit to the more geography-explicit approach of Sedghifar et al. [2015, 2016]. The assumption of large population size is another assumption which could also be relaxed (see Appendix 1 in Gravel [2012]), as this will likely influence the rate at which ancestry junctions are fixed or lost from the population.

Aside from the challenges of model misspecification, performing inference will be particularly difficult considering the computational burden of computing the tract length distribution for a set of migration rates and fitness matrix parameters. Gravel [2012] uses a maximum-likelihood scheme to identify the set of parameters that best describe the magnitude and timing of migration events from a source into a target population. Given that our primary interest is to infer the effects of purifying selection, it may be more efficient to treat the migration history as a latent variable to be marginalized over using Markov chain Monte Carlo.

Despite these challenges, our framework for computing statistical properties of haplotypes in a hybrid zone represents one of only a few recent efforts which aim to exploit the combination of whole-genome sequencing and dense genotyping approaches that have emerged for non-model systems. In particular, this model is the only example that we know of for deriving locus-specific haplotype patterns under epistatis. Given the complexity of this problem, an alternative option may be to use simulation-based classification in a machine learning framework [Chan et al., 2018, Schrider and Kern, 2018, Sheehan and Song, 2016]. Rather than focusing on any one summary statistic, several summary statistics with potential relevance to purifying selection against genetic incompatibilities could be used simultaneously. Alternatively, Chan et al. [2018] describe another machine learning approach which could instead use genotype data directly.

Regardless of the methods used to identify genomic patterns of purifying selection against incompatibility loci, this effort represents one facet of the many lines of evidence necessary to identify and describe the causes of reproductive isolation between species.

Acknowledgements

We thank Megan Frayer and John Hvala at The University of Wisconsin, Madison for their helpful discussions and insight. We also thank Yaniv Brandvain for additional advice and encouragement. Members of the Novembre, Stephens, and He labs provided useful feedback at an early stage. This work was funded by NSF grant DEB-1353737 to Bret Payseur, DEB-1353737 to Bret Payseur and John Novembre as well as NSF Graduate Research Fellowship and National Institute Of General Medical Sciences of the National Institutes of Health under award numbers DGE-1144082 and T32GM007197 to Joel Smith.

References

Aleksandr Andreychenko. *Uniformization for time-inhomogeneous Markov population models*. PhD thesis, Saarland University, 2010.

NH Barton. The dynamics of hybrid zones. *Heredity*, 43(3):341, 1979.

Nicholas H Barton and Godfrey M Hewitt. Analysis of hybrid zones. *Annual Review of Ecology and systematics*, 16(1):113–148, 1985.

Jeffrey Chan, Valerio Perrone, Jeffrey P Spence, Paul A Jenkins, Sara Mathieson, and Yun S Song. A likelihood-free inference framework for population genetic data using exchangeable neural networks. *bioRxiv*, 2018. doi: 10.1101/267211. URL `https://www.biorxiv.org/content/early/2018/02/18/267211`.

Theodosius Dobzhansky. *Genetics and the Origin of Species*. Columbia University Press, 1937.

John A Endler. Gene flow and population differentiation: studies of clines suggest that differentiation along environmental gradients may be independent of gene flow. *Science*, 179(4070):243–250, 1973.

Sergey Gavrilets. Hybrid zones with dobzhansky-type epistatic selection. *Evolution*, 51(4):1027–1035, 1997.

Zachariah Gompert and C Alex Buerkle. Analyses of genetic ancestry enable key insights for molecular ecology. *Molecular Ecology*, 22(21):5278–5294, 2013.

Zachariah Gompert, Thomas L Parchman, and C Alex Buerkle. Genomics of isolation in hybrids. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 367(1587):439–450, 2012.

Simon Gravel. Population genetics models of local ancestry. *Genetics*, 191(2):607–619, 2012.

Kelley Harris and Rasmus Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, 9(6):e1003521, 2013.

Richard G Harrison and Erica L Larson. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. *Molecular Ecology*, 25(11):2454–2466, 2016.

Garrett Hellenthal, George BJ Busby, Gavin Band, James F Wilson, Cristian Capelli, Daniel Falush, and Simon Myers. A genetic atlas of human admixture history. *Science*, 343(6172):747–751, 2014.

John A Hvala, Megan E Frayer, and Bret A Payseur. Signatures of hybridization and speciation in genomic patterns of ancestry. *Evolution*, 2018.

Mason Liang and Rasmus Nielsen. The lengths of admixture tracts. *Genetics*, 197(3):953–967, 2014.

Dorothea Lindtke and C Alex Buerkle. The genetic architecture of hybrid incompatibilities and their effect on barriers to introgression in secondary contact. *Evolution*, 69(8):1987–2004, 2015.

Po-Ru Loh, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K Pickrell, David Reich, and Bonnie Berger. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 193(4):1233–1254, 2013.

Shamoni Maheshwari and Daniel A Barbash. The genetics of hybrid incompatibilities. *Annual Review of Genetics*, 45:331–355, 2011.

Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

Bret A Payseur. Using differential introgression in hybrid zones to identify genomic regions involved in speciation. *Molecular Ecology Resources*, 10(5):806–820, 2010.

Bret A Payseur and Loren H Rieseberg. A genomic perspective on hybridization and speciation. *Molecular Ecology*, 25(11):2337–2360, 2016.

John E Pool and Rasmus Nielsen. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181(2):711–719, 2009.

Daven C Presgraves. The molecular evolutionary basis of species formation. *Nature Reviews Genetics*, 11(3):175, 2010.

Alkes L Price, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6):e1000519, 2009.

Douglas W Schemske. Understanding the origin of species. *Evolution*, 54(3):1069–1073, 2000.

Douglas W Schemske and HD Bradshaw. Pollinator preference and the evolution of floral traits in monkeyflowers (Mimulus). *Proceedings of the National Academy of Sciences*, 96(21):11910–11915, 1999.

Dolph Schluter. Evidence for ecological speciation and its alternative. *Science*, 323(5915):737–741, 2009.

Daniel R Schrider and Andrew D Kern. Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics*, 2018.

Molly Schumer and Yaniv Brandvain. Determining epistatic selection in admixed populations. *Molecular Ecology*, 25(11):2577–2591, 2016.

Molly Schumer, Rongfeng Cui, Daniel L Powell, Rebecca Dresner, Gil G Rosenthal, and Peter Andolfatto. High-resolution mapping reveals hundreds of genetic incompatibilities in hybridizing fish species. *Elife*, 3, 2014.

Alisa Sedghifar, Yaniv Brandvain, Peter Ralph, and Graham Coop. The spatial mixing of genomes in secondary contact zones. *Genetics*, 201(1):243–261, 2015.

Alisa Sedghifar, Yaniv Brandvain, and Peter Ralph. Beyond clines: lineages and haplotype blocks in hybrid zones. *Molecular Ecology*, 25(11):2559–2576, 2016.

Ole Seehausen, Roger K Butlin, Irene Keller, Catherine E Wagner, Janette W Boughman, Paul A Hohenlohe, Catherine L Peichel, Glenn-Peter Saetre, Claudia Bank, Åke Brännström, et al. Genomics and the origin of species. *Nature Reviews Genetics*, 15(3):176, 2014.

Sara Sheehan and Yun S Song. Deep learning for population genetic inference. *PLoS Computational Biology*, 12(3):e1004845, 2016.

William J Stewart. *Introduction to the numerical solution of Markov chains*. Princeton University Press, 1994.

Andrea L Sweigart and John H Willis. Molecular evolution and genetics of postzygotic reproductive isolation in plants. *F1000 Biology Reports*, 4, 2012.

Peter Tiffin and Jeffrey Ross-Ibarra. Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, 29(12):673–680, 2014.

Michael Turelli, Jeremy R Lipkowitz, and Yaniv Brandvain. On the Coyne and Orr-igin of species: effects of intrinsic postzygotic isolation, ecological differentiation, X chromosome size, and sympatry on Drosophila speciation. *Evolution*, 68(4):1176–1187, 2014.

Michael A White, Brian Steffy, Tim Wiltshire, and Bret A Payseur. Genetic dissection of a key reproductive barrier between nascent species of house mice. *Genetics*, 189(1):289–304, 2011.

MJD White. Models of speciation. *Science*, 159(3819):1065–1070, 1968.