THE UNIVERSITY OF CHICAGO

LEVERAGING HAPLOTYPE-BASED INFERENCE TO DESCRIBE ADAPTATION AND SPECIATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECOLOGY AND EVOLUTION

BY
JOEL HAVILAND SMITH

CHICAGO, ILLINOIS
JUNE 2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am incredibly grateful for the support and mentoring of my PhD advisor, John Novembre. His unending enthusiasm and patience has been an important asset toward the completion of my dissertation, and I feel lucky to have received the attention and care that he has offered for my development as a scientist. Likewise, Dick Hudson has been a constant source of advice and friendship from the beginning of my PhD. Together, with Matthew Stephens and Marty Kreitman, my thesis committee has instilled a keen sense of scientific rigor into my research perspective, and I feel privileged to consider them my role-models. I would also like to acknowledge my undergraduate mentors and collaborators: David Begun, Graham Coop, Dena Grossenbacher, Charles Langley, Jonah Piovia-Scott, Maureen Stanton, David Still, Eric Bishop-von Wettberg and Neal Williams. Their influence and encouragement was an important spark for my early intellect and scientific trajectory.

My labmates, friends, and climbing partners have also been a constant source of support, personal growth and fun distractions. Their catches have always kept me standing on my feet, and I am thankful to have such an awesome group of people to enjoy my time with. And, of course, I am grateful for all of the love and support my family has offered over the years–even from a long distance away. This work is ultimately a product of all those who influenced me; and in that respect, I owe everything to the care and guidance of my parents.

ABSTRACT

Forward progress in empirical population genetics is closely tied to the development of theory which can accommodate and keep pace with the production of genetic data. In recent years, the ability to survey genetic variation at increasingly greater resolution, across the genomes of a variety of species, has prompted new approaches to use this data for population genetic inference. While many models have historically relied on assuming independence among genetic variants in a sample of chromosomes, there are now a variety of methods which can use the non-independence among variants as a source of information. In particular, the unique combination and co-inheritance of variants on a chromosome can be used to define "haplotypes" of linked genetic variation associated with specific populations, individuals, or variants from which they are descended. The work presented here is a contribution to this class of population genetic models which describes: (1) a method to estimate the timing of adaptation for a beneficial allele, including several applications to recent human evolution, (2) an application of the same method to infer the timing of introgression for coat color alleles in North American wolves and high-altitude adaptation in Tibetans, (3) a model to infer the action of purifying selection against genetic incompatibilities in a hybrid zone, and (4) a reanalysis of genomic data from *Heliconius* butterflies which confirms the role of hybridization in transferring mimicry phenotypes between species.

CHAPTER 1

INTRODUCTION

The transmission of chromosomes from parent to offspring leads to correlated inheritance of linked genetic variation, also known as linkage disequilibrium (LD) [Geiringer, 1944]. Recombination, demographic history, and natural selection are known to play a key role in determining patterns of LD at both genome-wide and locus-specific scales, and there are a variety of metrics that can be used to quantify the direction and magnitude of LD with the goal of learning about these processes [Hudson, 2001, Pritchard and Przeworski, 2001]. Understanding the scale of LD is also useful in the design and analysis of trait mapping studies [Risch and Merikangas, 1996, Wall and Pritchard, 2003], and because recombination is the primary mechanism by which LD decays, the development of models which can use patterns of LD to infer fine-scale rates of recombination has been an active area of research for the past two decades [Clark et al., 2010, Stumpf and McVean, 2003].

Early studies which sought to describe LD at the genome-wide scale in human samples discovered an unexpected pattern that was inconsistent with a model of uniform recombination rates across the genome: long stretches of genetic variation in high LD, separated by short regions of low LD and high recombination [Gabriel et al., 2002, Wall and Pritchard, 2003]. Mapping the genetic determinants of these localized regions of low LD, or recombination "hotspots", has since become its own research endeavor with several mechanisms having been described for the maintenance and turnover of these hotspots across multiple species [Auton et al., 2013, Baker et al., 2017, Coop and Przeworski, 2007, Myers et al., 2010, Singhal et al., 2015, Stevison

et al., 2015].

Another consequence of these extended regions of high LD is the maintenance of co-inherited genetic variants as non-recombining blocks of ancestry or "haplotypes" [Gabriel et al., 2002, Wall and Pritchard, 2003]. It was initially thought that this haplotype structure would be a benefit for genome wide association studies (GWAS) which aimed to map genetic variants underlying complex traits and disease. While high levels of LD meant that fewer loci would need to be genotyped in order to tag a causative allele of interest, this also meant that larger sample sizes would be required to localize their position more precisely–especially for alleles at low frequency [Visscher et al., 2012].

One benefit of non-recombining loci is the ease with which a genealogy can be established within a sample. As a result, there has been much effort put forth to survey haplotype diversity among mitochondria and Y chromosomes in human samples around the world in order to reconstruct historical migrations of people with respect to these loci [Cann et al., 1987]. However, recent theoretical and computational advances have allowed the use of autosomal sequence data by inferring ancestral haplotype blocks which are identical by descent (IBD) [Donnelly, 1983]. Ralph and Coop [2013] use the length and number of IBD blocks shared between modern populations to gain insight into their ancestral relationships.

To illustrate this with an example, we can assume that the chromosomes inherited from a particular parent will have some unique combination of genetic variants which can reliably label them as being descended from that parent. We can refer to this combination of variants as a haplotype. In each subsequent generation, there is some

probability that this parent's haplotype will not be inherited. This could happen because an individual carrying this chromosome does not leave any offspring, or the other parent's chromosome is transmitted instead. For those haplotype tracts which are passed forward, recombination events will break this ancestry into smaller pieces flanked by unrelated ancestry that is not descended from the same parent, assuming there is no inbreeding. In addition to decreasing in length, the frequency of these haplotype tracts will change as a result of some combination of random and non-random sampling. Examples of non-random sampling would include frequency increases and decreases according to a selective benefit or detriment conferred by variation carried on a particular haplotype.

While in this example we have defined IBD with respect to a particular parental chromosome, we can instead define IBD with respect to the common ancestral haplotype of a particular locus or a reference panel of chromosomes from a particular population. Assuming there is no effect of selection, the length distribution for a sample of IBD haplotypes is proportional to the local recombination rate, its frequency in the population, and the amount of time since common ancestry. The flexibility with which IBD can be defined has proven useful for inferring time dependent changes in population size and migration history among admixed populations [Gravel, 2012, Harris and Nielsen, 2013, Li and Durbin, 2011, Palamara et al., 2012, Pool and Nielsen, 2009, Sedghifar et al., 2015].

The variety of modeling approaches used in these methods reflect the different approximations and assumptions that are appropriate for inferring particular parameters of interest. Despite these differences, most of these haplotype-based approaches

make use of a Markov assumption to capture the correlation in ancestry along a chromosome. This most often takes the form of a Hidden Markov Model (HMM) which is defined on the basis of an ancestry state space, a likelihood for the parameter of interest conditional on each state (emission probabilities), and a prior for the probability of a nucleotide position being in a particular state given the state of the previous position (transition probabilities). In addition to inferring changes in local ancestry along a chromosome, HMMs can be specified and used to infer a variety of parameters which are relevant to the local recombination rate, historical population size, and migration history [Li and Durbin, 2011, Li and Stephens, 2003, Price et al., 2009].

The most widely used approach relies on a conditional sampling distribution (CSD) first developed and implemented by Li and Stephens [2003] in what are now known as "haplotype copying models". This framework models local ancestry on chromosome as an imperfect mosaic of chromosomes in a reference panel [see Model Description in Section 2.3]. The observed chromosome copies ancestry from one of the haplotypes in the reference panel and switches ancestry to other reference haplotypes at a rate proportional to the population scaled recombination rate. New mutations are also modeled according to a specified miscopying rate which is proportional to the mutation rate. The original application of this framework was focused on identifying recombination hotspots, which would correspond to regions with a high haplotype switching rate relative to the background switching rate. In the following chapter, we take a similar approach to estimate the timing of adaptation for a beneficial allele.

Alternatively, subsequent extensions to the Li and Stephens [2003] haplotype

copying model have used the CSD assumption to directly infer the local ancestry state variable along chromosomes. This is typically done in the context of an admixed sample derived from 2 or more source populations [Falush et al., 2003, Lawson et al., 2012, Price et al., 2009]. For these applications, the inferred length distribution of local ancestry blocks can then be used to describe population structure among samples as well as the migration history that resulted in the observed pattern of admixture [Gravel, 2012, Gravel et al., 2013, Kidd et al., 2012, Leslie et al., 2015, Moreno-Estrada et al., 2013]. In Chapter 4, we adopt the same approach for modeling inferred local ancestry blocks in a hybrid zone between two species or source populations.

While demography is expected to affect the length distribution of haplotypes across the genome, natural selection acts at a locus-specifc scale, and theory describing the effect of positive selection on haplotype lengths has been widely used to identify loci which have undergone recent adaptation [Durrett and Schweinsberg, 2004, Kaplan et al., 1989, Smith and Haigh, 1974, Vitti et al., 2013]. There are, however, fewer statistical methods which can use this signature for parameter inference rather than a tool for identifying outlier loci under selection. More specifically, estimates for the age and strength of selection on beneficial mutations have primarily been based on heuristic approaches which underestimate the true variance around mean point estimates. Alternatively, simulation based approaches which use low dimensional summaries of the data yield less informative estimates with large confidence intervals. In Chapter 2 we review these shortcomings in more detail, and provide a new likelihood-based method which fills this gap to provide a more accurate estimate

for the timing of adaptation on a beneficial allele. In addition to reexamining several previously studied loci under recent positive selection in humans, in Chapter 3 we provide the first estimate for the timing of selection on a haplotype which confers high altitude adaptation in Tibetans, and which is derived from Denisovan archaic hominins. We also provide a novel application to population-specific estimates on the timing of coat color adaptation in North American Wolves.

In contrast to the abundance of methods and theory for signatures of positive selection, there has been much less work devoted to understanding the haplotype patterns associated with negative selection. However, unpublished work by D. Ortega-Del Vecchyo et al. aims to use theory from Maruyama [1974] in combination with haplotype lengths to identify the strength of selection against deleterious mutations. Even more rare are theoretical treatments of the expected signatures of purifying selection against incompatible allele combinations at two or more loci. Negative epistatic interactions between alleles which have fixed in different populations, and which contribute to reproductive isolation, have received particular attention due to widespread interest in understanding the formation of species [Coyne and Orr, 2004]. In Chapter 4, we review the types of two-locus epistatic interactions which are suspected to be most common in generating reproductive isolation, and derive a model for the distribution of haplotype lengths around such loci when they come into contact in a hybrid zone. We conclude in Chapter 5 with a brief test which examines the role of hybridization in transferring alleles which underlie mimicry phenotypes among *Heliconius* butterfly species.

Advances in haplotype-based inference reflect a trend towards using all of the

available information which can be used to understand the demographic and adaptive history of a sample. By modeling the transmission of haplotypes under a variety of scenarios, population genetics inference will come closer to describing a more complete and biologically real picture of heredity. Rather than filter genetic datasets for independent unlinked loci, a model for the non-independence of these loci, in addition to the distribution of their frequencies, will provide a more rich source of information for parameter inference. While the models and applications presented here represent a positive affirmation of this scientific agenda, there is much more work to be done for relaxing particular assumptions and expanding these models for greater flexibility. Future directions in haplotype-based inference will likely benefit from the incorporation of both geographic information and ancient DNA samples. Creative approaches to incorporate spatial and temporal variables into expected haploytpe patterns will be a fruitful way forward; however, the compromise between model complexity and computational feasibility will remain a significant challenge.

CHAPTER 2

ESTIMATING TIME TO THE COMMON ANCESTOR FOR A BENEFICIAL ALLELE

Joel Smith[1], Graham Coop[2], Matthew Stephens[3,4], John Novembre[1,3]

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

2 Department of Evolution and Ecology, University of California, Davis, CA

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

4 Department of Statistics, University of Chicago, Chicago, IL

2.1 Abstract

The haplotypes of a beneficial allele carry information about its history that can shed light on its age and the putative cause for its increase in frequency. Specifically, the signature of an allele's age is contained in the pattern of variation that mutation and recombination impose on its haplotypic background. We provide a method to exploit this pattern and infer the time to the common ancestor of a positively selected allele following a rapid increase in frequency. We do so using a hidden Markov model which leverages the length distribution of the shared ancestral haplotype, the accumulation of derived mutations on the ancestral background, and the surrounding background haplotype diversity. Using simulations, we demonstrate how the inclusion of information from both mutation and recombination events increases accuracy relative to approaches that only consider a single type of event. We also show the behavior of the estimator in cases where data do not conform to model assumptions, and pro-

vide some diagnostics for assessing and improving inference. Using the method, we analyze population-specific patterns in the 1000 Genomes Project data to estimate the timing of adaptation for several variants which show evidence of recent selection and functional relevance to diet, skin pigmentation, and morphology in humans.

2.2 Introduction

A complete understanding of adaptation depends on a description of the genetic mechanisms and selective history that underly heritable traits [Radwan and Babik, 2012]. Once a genetic variant underlying a putatively adaptive trait has been identified, several questions remain: What is the molecular mechanism by which the variant affects organismal traits and fitness [Dalziel et al., 2009]?; what is the selective mechanism responsible for allelic differences in fitness?; did the variant arise by mutation more than once [Elmer and Meyer, 2011]?; when did each unique instance of the variant arise and spread [Slatkin and Rannala, 2000]? Addressing these questions for numerous case studies of beneficial variants across multiple species will be necessary to gain insight into general properties of adaptation [Stinchcombe and Hoekstra, 2008].

Here, our focus is on the the last of the questions given above; that is, when did a mutation arise and spread? Understanding these dates can give indirect evidence regarding the selective pressure that may underlie the adaptation. This is especially useful in cases where it is logistically infeasible to assess fitness consequences of a variant in the field directly [Barrett and Hoekstra, 2011]. In humans, for example, dispersal across the globe has resulted in the occupation of a wide variety of habitats,

and in several cases, selection in response to specific ecological pressures appears to have taken place. There are well-documented cases of loci showing evidence of recent selection in addition to being functionally relevant to known phenotypes of interest [Jeong and Di Rienzo, 2014]. Nakagome et al. [2016] specify time intervals defined by the human dispersal out-of-Africa and the spread of agriculture to show the relative concordance among allele ages for several loci associated with autoimmune protection and risk, skin pigmentation, hair and eye color, and lactase persistence.

When a putative variant is identified as the selected site, the non-random association of surrounding variants on a chromosome can be used to understand its history. This combination of surrounding variants is called a haplotype, and the non-random association between any pair of variants is called linkage disequilibrium (LD). Due to recombination, LD between the focal mutation and its initial background of surrounding variants follows a per-generation rate of decay. New mutations also occur on this haplotype at an average rate per generation. The focal mutation's frequency follows a trajectory determined by the stochastic outcome of survival, mating success and offspring number. If the allele's selective benefit increases its frequency at a rate faster than the rate at which LD decays, the resulting signature is one of high LD and a reduction of polymorphism near the selected mutation [Smith and Haigh, 1974]. Many methods to exploit this pattern have been developed in an effort to identify loci under recent positive selection (reviewed in Nielsen [2005]). A parallel effort has focused on quantifying specific properties of the signature to infer the age of the selected allele.

The most commonly used methods to estimate allele age rely on summary statis-

tics. These approaches can be further classified as either heuristic or model-based methods. Heuristic approximations rely on a point estimate of the mean length of the selected haplotype (using the decay of homozygosity around the selected locus), or a count of derived mutations within an arbitrary cutoff distance from the selected site [Coop et al., 2008, Hudson, 2007, Meligkotsidou and Fearnhead, 2005, Tang et al., 2002, Thomson et al., 2000]. These approaches ignore uncertainty in the extent of the selected haplotype on each chromosome, which can lead to inflated confidence in the point estimates.

Alternative model-based approaches that also use summary statistics employ an Approximate Bayesian Computation (ABC) framework. These methods use an explicit model for simulation to identify a distribution of ages that are consistent with the observed data [Beaumont et al., 2002, Beleza et al., 2013b, Nakagome et al., 2016, Ormond et al., 2016, Peter et al., 2012, Pritchard et al., 1999, Przeworski, 2003, Tavaré et al., 1997, Tishkoff et al., 2007, Voight et al., 2006]. This provides a measure of uncertainty induced by the randomness of recombination, mutation, and genealogical history and produces an approximate posterior distribution on allele age. Despite these advantages, ABC approaches suffer from an inability to capture all relevant features of the sample due to their reliance on summary statistics.

As full-sequencing data become more readily available, defining the summary statistics which capture the complex LD among sites and the subtle differences between haplotypes will be increasingly challenging. For this reason, efficiently computable likelihood functions that leverage the full sequence data, rather than low dimensional summaries of the data, are increasingly favorable.

Several approaches attempt to compute the full likelihood of the data using an importance sampling framework [Chen and Slatkin, 2013, Coop and Griffiths, 2004, Slatkin, 2001, 2008]. Conditioning on the current frequency of the selected allele, frequency trajectories and genealogies are simulated and given weight proportional to the probability of their occurrence under a population genetic model. While these approaches aim to account for uncertainty in the allele's frequency trajectory and genealogy, they remain computationally infeasible for large samples or do not consider recombination across numerous loci.

In a related problem, early likelihood-based methods for disease mapping have modeled recombination around the ancestral haplotype, providing information for the time to the common ancestor (TMRCA) rather than time of mutation [McPeek and Strahs, 1999, Morris et al., 2000, 2002, Rannala and Reeve, 2001, 2003]. These models allowed for the treatment of unknown genealogies and background haplotype diversity before access to large data sets made computation at the genome-wide scale too costly. Inference is performed under Markov chain Monte Carlo (MCMC) to sample over the unknown genealogy while ignoring LD on the background haplotypes, or approximating it using a first-order Markov chain. In a similar spirit, Chen et al. (2015) revisit this class of models to estimate the strength of selection and time of mutation for an allele under positive selection using a hidden Markov model.

Hidden Markov models have become a routine tool for inference in population genetics. The Markov assumption allows for fast computation and has proven an effective approximation for inferring the population-scaled recombination rate, the demographic history of population size changes, and the timing and magnitude of

admixture events among genetically distinct populations [Hinch et al., 2011, Li and Durbin, 2011, Li and Stephens, 2003, Price et al., 2009, Wegmann et al., 2011]. The approach taken by Chen et al. [2015] is a special case of two hidden states—the ancestral and background haplotypes. The ancestral haplotype represents the linked background that the focal allele arose on, while the background haplotypes represent some combination of alleles that recombine with the ancestral haplotype during its increase in frequency. Chen et al. [2015] compute maximum-likelihood estimates for the length of the ancestral haplotype on each chromosome carrying the selected allele. Inference for the time of mutation is performed on these fixed estimates assuming they are known. The authors condition the probability of an ancestry switch event on a logistic frequency trajectory for the selected allele and assume independence among haplotypes leading to the common ancestor. The likelihood for background haplotypes is approximated using a first-order Markov chain to account for non-independence among linked sites.

Here, we present a Hidden Markov model that leverages both the length of the ancestral haplotype on each chromosome as well as derived mutations that have accumulated on the ancestral haplotype. Our method implements an MCMC which samples over the unknown ancestral haplotype to generate a sample of the posterior distribution for the TMRCA. Our emission probabilities account for the LD structure among background haplotypes using the Li and Stephens [2003] haplotype copying model and a reference panel of haplotypes without the selected allele (Figure 2.1b). In contrast to the first order Markov chain employed by Chen et al. (2015), the Li and Stephens [2003] model provides an approximation to the coalescent with

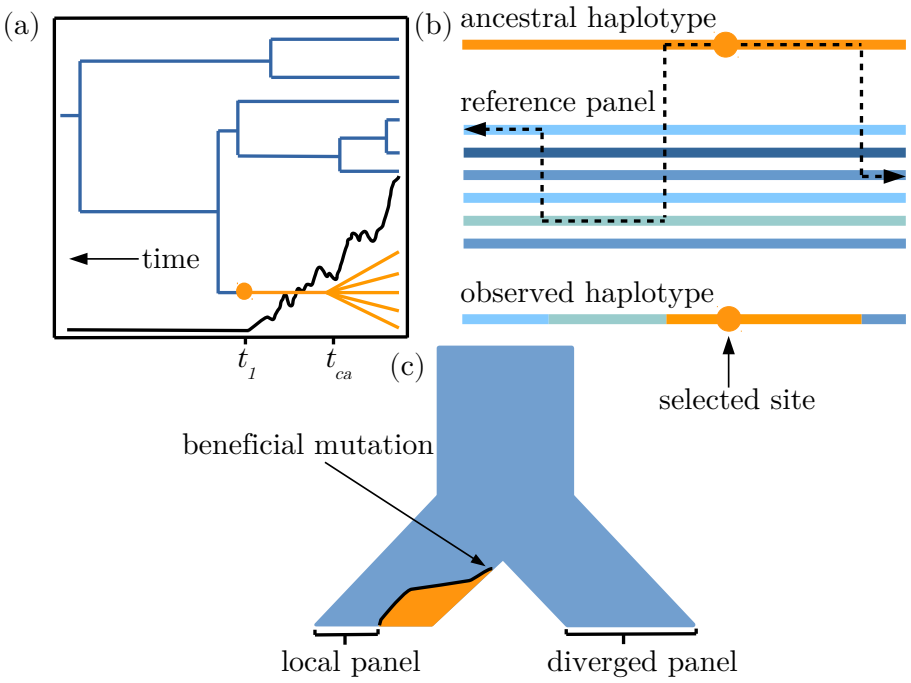

Figure 2.1: Visual descriptions of the model. a) An idealized illustration of the effect of a selectively favored mutation's frequency trajectory (black line) on the shape of a genealogy at the selected locus. The orange lineages are chromosomes with the selected allele. The blue lineages indicate chromosomes that do not have the selected allele. Note the distinction between the time to the common ancestor of chromosomes with the selected allele, t_{ca}, and the time at which the mutation arose, t_1. b) The copying model follows the ancestral haplotype (orange) moving away from the selected site until recombination events within the reference panel lead to a mosaic of non-selected haplotypes surrounding the ancestral haplotype. c) A demographic history with two choices for the reference panel: local and diverged. After the ancestral population at the top of the figure splits into two sister populations, a beneficial mutation arises and begins increasing in frequency. The orange and blue colors indicate frequency of the selected and non-selected alleles, respectively.

recombination by modeling a focal haplotype as an imperfect mosaic of haplotypes in the reference panel.

While Chen et al. (2015) use a mutation parameter in their HMM, the count

of derived mutations on the background haplotype does not directly influence their estimation of time since mutation. The probability of observing a mutation on the selected haplotypes of beneficial allele carriers depends on two parameters: the per generation mutation rate and the time to the common ancestor (TMRCA). The Chen et al model uses a compound parameter for these such that the observed mutations do not directly inform their estimates of timing. In our model we separately include the TMRCA and mutation rate as parameters and thus incorporate information from mutations directly into our inference of the TMRCA.

Our approach also differs in that we do not presume to know the true extent of the ancestral haplotype, and instead treat it as a latent variable to be marginalized over. This allows our estimation of the TMRCA to reflect uncertainty in the precise switch point off of the ancestral haplotype, which in many cases will be difficult to distinguish from the background haplotypes. Another significant difference is that our model does not make assumptions about the frequency trajectory apart from that a sufficiently hard sweep occurred to incur a star-shaped genealogy. Below, we use simulations to show the sensitivity of our model to these simplified assumptions for varying strengths of selection, final allele frequencies, and sampling regimes for the choice of reference panel. An R package is available to implement this method on github (https://github.com/joelhsmth/startmrca).

2.3 Model Description

In general, the TMRCA for a sample of haplotypes carrying the advantageous allele (hereafter referred to as t_{ca}) will be more recent than the time of mutation [Kaplan

et al., 1989]. We aim to estimate t_{ca} in the case where a selectively advantageous mutation occurred in an ancestor of our sample t_1 generations ago (Figure 2.1a). Viewed backwards in time, the selected variant decreases in frequency at a rate proportional to the selection strength. During a rapid drop in allele frequency, the coalescent rate among haplotypes carrying the selected variant is amplified. The same effect would be observed for population growth from a small initial size forward in time [Hudson, 1990, Slatkin and Hudson, 1991]. As a result, the genealogy of a sample having undergone selection and/or population growth becomes more "star-shaped". This offers some convenience, as it becomes more appropriate to invoke an assumption of independence among lineages when selection is strong. We would like to emphasize that this assumption necessarily implies that the beneficial allele has a single ancestral haplotype that has increased in frequency. This is in contrast to a scenario in which the beneficial allele has been present in the population for some time prior to selection. For that case, multiple ancestral haplotypes would increase in frequency simultaneously resulting in a genealogy that is not star-shaped.

We assume no crossover interference between recombination events within a haplotype, and therefore treat each side flanking the focal allele separately. We define one side of the selected site, within a window of some predetermined length, to have L segregating sites, such that an individual's sequence will be indexed from site $s = \{1, ..., L\}$, where $s = 1$ refers to the selected site (a notation reference is provided in Table 2.1). To simplify notation, this description will be written for a window on one side flanking the selected site. Note that the opposing side of the selected site is modeled in an identical fashion after redefining L.

n	Number of haplotypes with the selected allele
m	Number of haplotypes without the selected allele
L	Number of SNPs flanking the selected site (one side considered at a time)
X	$n \times L$ matrix of haplotypes with the selected allele
H	$m \times L$ matrix of haplotypes without the selected allele
X_{ij}	Allele in haplotype i at SNP j, where $i \in \{1, ..., n\}$, and $j \in \{1, ..., L\}$
H_{zj}	Allele in haplotype z at SNP j, where $z \in \{1, ..., m\}$, and $j \in \{1, ..., L\}$
A_j	Allele at site j on the ancestral haplotype
Z_{ij}	The reference panel haplotype from which X_i copies at site j
t_{ca}	Time to the most recent common ancestor (TMRCA)
W_i	The location of the first recombination event off of the ancestral haplotype
r	Recombination rate per basepair per generation
μ	Mutation rate per basepair per generation
θ	Haplotype miscopying rate, or population-scaled mutation rate ($4N\mu$)
ρ	Haplotype switching rate, or population-scaled recombination rate ($4Nr$)
d_w	Physical distance of site w from the selected site, where $w \in \{1, ..., L\}$
c_j	Number of basepairs between sites j and $j + 1$
α_{iw}	Likelihood of haplotype i for sites $1, ..., w$
β_{iw}	Likelihood of haplotype i for SNPs $(w + 1), ..., L$

Table 2.1: Notation used to describe the model.

Let X denote an $n \times L$ data matrix for a sample of n chromosomes with the selected variant. X_{ij} is the observed allelic type in chromosome i at variant site j, and is assumed to be biallelic where $X_{ij} \in \{1, 0\}$. Let H denote an $m \times L$ matrix comprising m chromosomes that do not have the selected variant where $H_{ij} \in \{1, 0\}$. Let A denote the ancestral haplotype as a vector of length L where A_j is the allelic type on the ancestral selected haplotype at segregating site j and $A_j \in \{1, 0\}$. We assume independence among lineages leading to the most recent common ancestor of the selected haplotype. This is equivalent to assuming a star-shaped genealogy

which, as noted above, is a reasonable assumption for sites linked to a favorable variant under strong selection. We can then write the likelihood as

$$\Pr\left(X \mid t_{ca}, A, H\right) = \prod_{i}^{n} \Pr\left(X_i \mid t_{ca}, A, H\right). \tag{2.1}$$

In each individual haplotype, X_i, we assume the ancestral haplotype extends from the selected allele until a recombination event switches ancestry to a different genetic background. Let $W = w$ indicate that the location of the first recombination event occurs between sites w and $w + 1$, where $W \in \{1, ..., L\}$ ($w = L$ indicates no recombination up to site L). We can then condition the probability of the data on the interval where the first recombination event occurs and sum over all possible intervals to express the likelihood as

$$\Pr\left(X_i \mid t_{ca}, A, H\right) = \sum_{w=1}^{L} \Pr\left(X_i \mid t_{ca}, A, H, W_i = w\right) \Pr\left(W_i = w \mid t_{ca}\right). \tag{2.2}$$

Assuming haplotype lengths are independent and identically distributed draws from an exponential distribution, the transition probabilities for a recombination event off of the ancestral haplotype are

$$\Pr\left(W_i = w \mid t_{ca}\right) = \begin{cases} e^{-rt_{ca}d_w}(1 - e^{-rt_{ca}(d_{w+1}-d_w)}) & \text{if } w = \{1, ..., (L-1)\}; \\ e^{-rt_{ca}d_L} & \text{if } w = L \end{cases} \quad labeleq3 \tag{2.3}$$

where d_w is the distance, in base pairs, of site w from the selected site and r is the

local recombination rate per base pair, per generation. The data for each individual, X_i, can be divided into two parts: one indicating the portion of an individual's sequence residing on the ancestral haplotype (before recombining between sites w and $w+1$), $X_{i(j\leq w)}$, and that portion residing off of the ancestral haplotype after a recombination event, $X_{i(j>w)}$. We denote a separate likelihood for each portion:

$$\alpha_{iw} = \Pr(X_{i(j\leq w)} \mid t_{ca}, A, W_i = w) \tag{2.4}$$

$$\beta_{iw} = \Pr(X_{i(j>w)} \mid H_{(j>w)}, W_i = w) \tag{2.5}$$

Because the focal allele is on the selected haplotype, $\alpha_{i1} = 1$. Conversely, we assume a recombination event occurs at some point beyond locus L such that $\beta_{iL} = 1$. We assume the waiting time to mutation at each site on the ancestral haplotype is exponentially distributed with no reverse mutations and express the likelihood as

$$\alpha_{iw} = \Pr(X_{i(j\leq w)} \mid t_{ca}, A, W_i = w) = e^{-t_{ca}\mu(d_w - w)} \prod_{j=2}^{w} \Pr(X_{ij} = a \mid t_{ca}, A) \tag{2.6}$$

$$\Pr\left(X_{ij} = a \mid t_{ca}, A\right) = \begin{cases} e^{-t_{ca}\mu} & \text{if } a = A_j; \\ 1 - e^{-t_{ca}\mu} & \text{if } a \neq A_j \end{cases} \tag{2.7}$$

The term, $e^{-t_{ca}\mu(d_w - w)}$, on the right side of Equation 2.6 captures the lack of mutation at invariant sites between each segregating site. Assuming $t_{ca}\mu$ is small, Equation 2.6 is equivalent to assuming a Poisson number of mutations (with mean

$t_{ca}\mu$) occurring on the ancestral haplotype.

For β_{iw}, the probability of observing a particular sequence after recombining off of the ancestral haplotype is dependent on standing variation in background haplotype diversity. The Li and Stephens [2003] haplotype copying model allows for fast computation of an approximation to the probability of observing a sample of chromosomes related by a genealogy with recombination. Given a sample of m haplotypes, $H \in \{h_1, ..., h_m\}$, a population scaled recombination rate ρ and mutation rate θ, an observed sequence of alleles is modeled as an imperfect copy of any one haplotype in the reference panel at each SNP. Let Z_{ij} denote the reference panel haplotype which X_i copies at the jth SNP, and c_j denote the number of base pairs between SNPs j and $j+1$. Z_{ij} follows a Markov process with transition probabilities

$$\Pr\left(Z_{i(j+1)} = z' \mid Z_{ij} = z\right) = \begin{cases} e^{-\rho_j c_j/m} + (1 - e^{-\rho_j c_j/m})(1/m) & \text{if } z' = z; \\ (1 - e^{-\rho_j c_j/m})(1/m) & \text{if } z' \neq z. \end{cases} \tag{2.8}$$

To include mutation, the probability that the sampled haplotype matches a haplotype in the reference panel is $m/(m+\theta)$, and the probability of a mismatch (or mutation event) is $\theta/(m+\theta)$. Letting a refer to an allele where $a \in \{1, 0\}$, the matching and mismatching probabilities are

$$\Pr\left(X_{ij} = a \mid Z_{ij} = z, h_1, ..., h_m\right) = \begin{cases} m/(m+\theta) + (1/2)(\theta/(m+\theta)) & \text{if } h_{z,j} = a; \\ (1/2)(\theta/(m+\theta)) & \text{if } h_{z,j} \neq a. \end{cases} \tag{2.9}$$

Equation 2.5 requires a sum over the probabilities of all possible values of Z_j using Equations 2.8 and 2.9. This is computed using the forward algorithm as described in Rabiner [1989] and Appendix A of Li and Stephens [2003]. It should be noted that this formulation does not model the observation of an invariant site among the background haplotypes. We tried an approach to model these sites, but saw no improvement in model performance (see Appendix A.2 and Table S.8).

The complete likelihood for our problem can then be expressed as:

$$\Pr(X \mid t_{ca}, A, H) = \prod_{i=1}^{n} \sum_{w=1}^{L} \alpha_{iw}\beta_{iw} \Pr\left(W_i = w \mid t_{ca}, A\right). \tag{2.10}$$

This computation is on the order $2Lnm^2$, and in practice for $m = 20$, $n = 100$ and $L = 4000$ takes approximately 3.027 seconds to compute on an Intel® Core™ i7-4750HQ CPU at 2.00GHz×8 with 15.6 GiB RAM.

2.4 Inference

Performing inference on t_{ca} requires addressing the latent variables w and A in the model. Marginalizing over possible values of w is a natural summation per haplotype that is linear in L as shown above. For A, the number of possible values is large (2^L), and so we employ a Metropolis–Hastings algorithm to jointly sample

the posterior of A and t_{ca}, and then we take marginal samples of t_{ca} for inference. We assign a uniform prior density for both A and t_{ca}, such that $\Pr(A) = 1/2^L$ and $\Pr(t_{ca}) = 1/(t_{max} - t_{min})$ where t_{max} and t_{min} are user-specified maximum and minimum values for t_{ca}. Proposed MCMC updates of the ancestral haplotype, A', are generated by randomly selecting a site in A and flipping to the alternative allele. For t_{ca}, proposed values are generated by adding a normally distributed random variable centered at 0: $t'_{ca} = t_{ca} + N(0, \sigma^2)$. To start the Metropolis–Hastings algorithm, an initial value of t_{ca} is uniformly drawn from a user-specified range of values (10 to 2000 in the applications here). To initialize the ancestral haplotype to a reasonable value, we use a heuristic algorithm which exploits the characteristic decrease in variation near a selected site (see Appendix A.1).

For each haplotype in the sample of beneficial allele carriers, the Li and Stephens [2003] model uses a haplotype miscopying rate θ, and switching rate ρ, to compute a likelihood term for loci following the recombination event off of the ancestral haplotype. For our analyses, we set $\rho = 4.4 \times 10^{-4}$ using our simulated values of $r = 1.1 \times 10^{-8}$ per bp per generation and $N = 10000$, where $\rho = 4Nr$. Following Li and Stephens [2003] we fix $\theta = (\sum_{m=1}^{n} 1/m)^{-1}$; as derived from the expected number of mutation events on a genealogy relating n chromosomes at a particular site. We found no discernible effects on estimate accuracy when specifying different values of ρ (Figure S.1 in Appendix A.3).

2.5 Results

Because our model requires a sample (or "panel") of reference haplotypes without the selected allele, we tested our method for cases in which the reference panel is chosen from the local population in which the selected allele is found, as well as cases where the panel is from a diverged population where the selected haplotype is absent (Figure 2.1). Regardless of scenario, the estimates are on average within a factor of 2 of the true value, and often much closer. When using a local reference panel, point estimates of t_{ca} increasingly underestimate the true value (TMRCA) as selection becomes weaker and the final allele frequency increases (Figure 2.2). Put differently, the age of older TMRCAs tend to be underestimated with local reference panels. Using the mean posteriors as point estimates, mean values of $\log_2$(estimate/true value) range from -0.62 to -0.14. Simulations using a diverged population for the reference panel removed the bias, though only in cases where the divergence time was not large. For a reference panel diverged by $0.5N$ generations, mean $\log_2$(estimate/true value) values range from -0.21 to -0.18. As the reference panel becomes too far diverged from the selected population, estimates become older than the true value (0.36 to 0.94 $\log_2$(estimate/true value)). In these cases, the HMM is unlikely to infer a close match between background haplotypes in the sample and the reference panel, leading to many more mismatches being inferred as mutation events on the ancestral haplotype and an older estimate of t_{ca}.

The bottom panel of Figure 2.2 shows the effect of selection strength and final allele frequency on the size of the 95% credible interval around point estimates normalized by the true TMRCA for each simulated data set. Before normalizing,

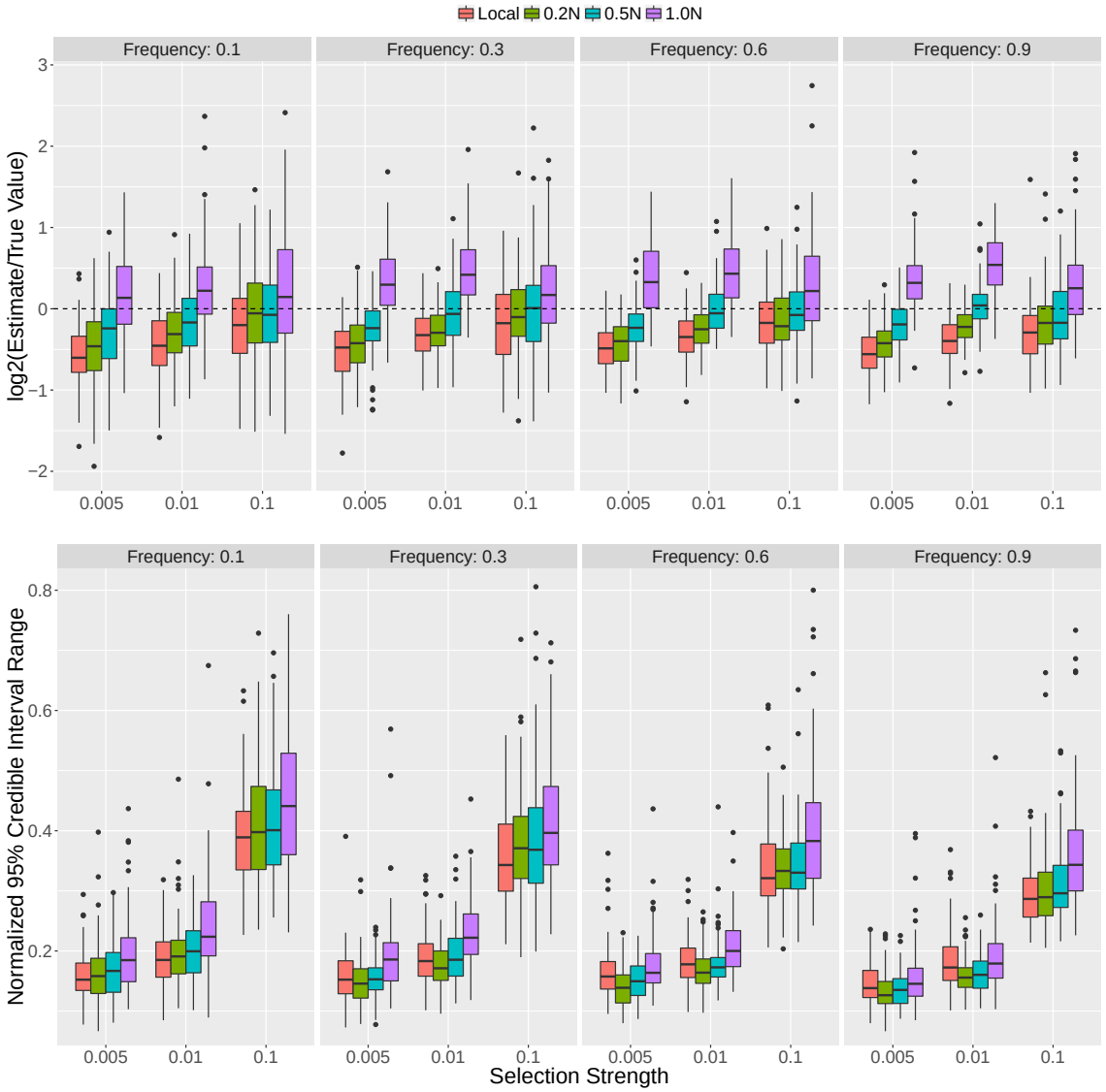

Figure 2.2: Accuracy of TMRCA point estimates and 95% credible interval ranges from posteriors inferred from simulated data under different strengths of selection, final allele frequencies and choice of reference panel. Credible interval range sizes are in units of generations and are normalized by the true TMRCA for each simulated data set. See Materials and Methods below for simulation details.

credible interval sizes using a local reference panel range from 73 to 213 generations for $2Ns = 100$, versus 18 to 22 generations when $2Ns = 2000$. Using local and diverged reference panels, we found a minimal effect of the sample size on point estimates (Figures S.2, S.3 in Appendix A.3). As expected, larger sample sizes for the carrier panel improve estimate accuracy. However, higher allele frequencies and weak selection are likely to induce more uncertainty due to the ancestral haplotype tracts recombining within the sample. We find this effect more pronounced with large sample sizes for the reference panel. We speculate that a large sample of reference haplotypes leads the focal selected haplotype to have increased probability copying from the reference panel leading to a shorter selected haplotype and slight over-estimate of the TMRCA.

We also performed simulations under varying degrees of mutation and recombination rate misspecification (Figure S.4 in Appendix A.3). In most cases, mean values of $\log_2$(estimate/true value) stay within an order of magnitude of 0. As expected, when both the mutation and recombination rate are misspecified, we find the most discrepancy. To assess the convergence properties of the MCMC, five replicate chains were run for each of 20 simulated data sets produced under three $2Ns$ values (100, 200 and 2000) for frequency trajectories ending at 0.1 (Figure S.5 in Appendix A.3). While care is always warranted with MCMC approaches, we find in practice that convergence among our replicate chains is attained relatively quickly ($\approx$ 3000 iterations for simulated data and 3000 - 9000 iterations for applied cases; see Figure S.9 in Appendix A.3).

We compared the performance of our estimator with three other model-based

approaches for allele age estimation by matching the simulation scheme performed by Chen et al. [2015] (Table S.7 in Appendix A.3). Our method shows improvement in accuracy (RMSE) and/or lower bias for simulations with lower frequencies of the beneficial allele (40%) regardless of the reference panel used. In cases where the final beneficial allele frequency is higher (80%), our method's accuracy remains as good or better than the other methods when using a diverged reference panel, with a two-orders-of-magnitude improvement of bias under strong selection ($s = 0.05$). Estimates when using a local reference panel and a high final beneficial allele frequency remain comparable to the other methods for strong selection, but tend to have more bias and decreased accuracy as selection strength decreases.

Assuming a star-genealogy among beneficial allele carriers may result in underestimating the variance for the posterior distribution when there is non-independence in our sample. To measure this affect, we computed TMRCA estimates on 100 bootstrap replicates for 4 simulated datasets under 2 selection strengths and 2 final allele frequencies (Table S.9 in Appendix A.4). We find close agreement between the 95% posterior credible intervals of the original data and the 95% confidence intervals computed on the bootstrap estimates for a selection strength of 0.1. for both final allele frequencies of 0.4 and 0.8. As expected, older TMRCAs are likely to violate the star-genealogy assumption, and in these cases we find that estimates from our original data are more narrow than the bootstrap confidence intervals.

2.5.1 Recombination Versus Mutation as a Source of Information

We compared our model-based inference with simpler estimates of the TMRCA using the number of derived mutations on the ancestral haplotype, and the mean length of the ancestral haplotype. In addition to quantifying the improvement our method has over these calculations, this also serves as an ad-hoc way to understand how the relative weight of information from mutation and recombination affects the performance of our method. One can model the haplotype lengths as independent and exponentially distributed to derive a recombination-based estimator, $\hat{t}_r$, as

$$\hat{t}_r = \frac{1}{\bar{w}_o r} \tag{2.11}$$

where r is the recombination rate and $\bar{w}_o$ is the observed mean ancestral haplotype length. To leverage the count of derived mutations on the ancestral haplotype, we use the Thomson et al. [2000] estimator. In a sample of n haplotypes with the selected allele, a mutation-based estimator, $\hat{t}_m$, can be calculated as

$$\hat{t}_m = \frac{1}{n} \sum_{i}^{n} \frac{y_i}{w_i \mu} \tag{2.12}$$

where y_i is the number of derived mutations on the ith haplotype which has length w_i basepairs. See Hudson (2007) for a derivation of the estimate for the variance of the Thomson estimator.

When using derived mutations, uncertainty in both the ancestral haplotype sequence and the length of the ancestral haplotype on each chromosome (w_i) can lead to poor estimation. To improve inference, researchers typically define a restricted

"non-recombining" region that may reliably contain derived mutations on the ancestral haplotype. This has two disadvantages: (1) There is more information available in the data which cannot be used because excess caution is necessary to prevent over-counting of derived mutations; and (2) There may still be unobserved recombination events in this restricted locus. To minimize the use of heuristics for a derived mutation approach, we used our model to find maximum-likelihood estimates of the ancestral haplotype breakpoints using Equation 2.2 in the model description. We also used the mean posterior estimate of the ancestral haplotype from our model to identify derived mutations. To calculate a recombination-based estimator of the TMRCA, we calculated $\bar{w}_o$ using the same maximum-likelihood estimates of the ancestral haplotype lengths inferred for the mutation estimator.

When using a local reference panel, the simple mutation estimator $\hat{t}_m$ consistently under-estimates the true TMRCA. The recombination estimate, however, remains accurate (Figure S.6 in Appendix A.3). We suspect this to be a result of poor estimation of the ancestral haplotype and violation of the star-genealogy assumption. In cases where selection is weaker and the genealogy is not star-shaped, derived mutations occurring early in the genealogy will be over-represented and incorrectly inferred to be the ancestral allele. In this way, high frequency derived alleles will not be counted. As predicted, increasing selection strength improves mutation estimator accuracy. The recombination estimator appears robust to this effect as long as selection is not too strong. For very strong selection, and young TMRCA values, maximum likelihood estimates of the haplotype lengths become constrained by the size of the locus studied. For example, in simulations with a selection strength of 0.05

and frequency of 0.1, the mean TMRCA is around 100 generations. Using equation 14 above, the mean length of the ancestral haplotype for a TMRCA of 100 generations is 2Mb, which is twice as large as the window size we use to make computation for our simulations feasible. Using a larger window around the selected locus would ameliorate this effect.

When using a diverged reference panel we find an opposite effect. In this case, the count of derived mutations result in an over-estimate and the haplotype lengths yield an under-estimate. We suspect this to be driven by poor matching between the reference panel and the background haplotypes among beneficial allele carriers. The low probability of matching between the reference and background haplotypes means that the lengths of the ancestral haplotype are inferred to extend further than their true lengths. This also leads to an overestimate for the mutation estimator because differences between the ancestral and background haplotypes are incorrectly inferred as derived mutations on the ancestral haplotype.

2.5.2 Application to 1000 Genomes Data

We applied our method to five variants previously identified as targets of recent selection in various human populations (Figure 2.3). Using phased data from the 1000 Genomes Project, we focused on variants that are not completely fixed in any one population so that we could use a local reference panel. The Li and Stephens [2003] haplotype copying model is appropriate in cases where ancestry switches occur among chromosomes within a single population, so we excluded populations in the Americas for which high levels of admixture are known to exist.

While the simulation results described above provide some intuition for the effects of selection strength, final allele frequency and choice of reference panel, we also performed simulations using the demographic history inferred by Tennessen et al. [2012] to explore the effects of non-equilibrium demographic history on our estimation accuracy (Figure S.10 in Appendix A.3). We find subtle differences in accuracy between the two demographic histories, where the non-equilibrium histories lead to negligible differences in mean values for $\log_2$(estimate/true value) and larger credible interval ranges.

ADH1B

A derived allele at high frequency among East Asians at the ADH1B gene (rs3811801) has been shown to be functionally relevant for alcohol metabolism [Eng et al., 2007, Osier et al., 2002]. Previous age estimates are consistent with the timing of rice domestication and fermentation approximately 10,000 years ago [Li et al., 2007, Peng et al., 2010a, Peter et al., 2012]. However, a more recent estimate by Peter et al. (2012) pushes this time back several thousand years to 12,876 (2,204 - 49,764) years ago. Our results are consistent with an older timing of selection, as our CHB sample (Han Chinese in Beijing, China) TMRCA estimate is 15,377 (13,763 - 17,281) years. Replicate chains of the MCMC are generally consistent, with the oldest estimates in the CHB sample showing the most variation among resampled datasets and the youngest estimate of 10,841 (9,720 - 12,147) in the KHV sample showing the least. When using a fine-scale recombination map, all of the ADH1B TMRCAs are inferred to be slightly older (Figure S.7 in Appendix).

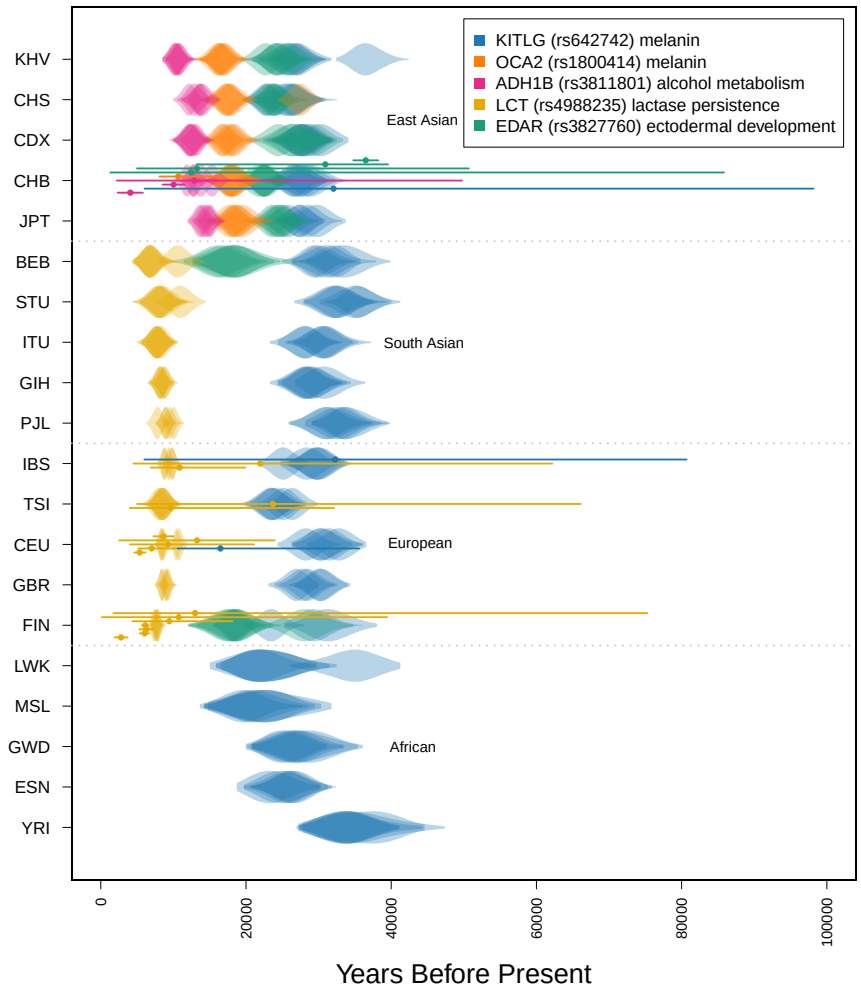

Figure 2.3: Comparison of TMRCA estimates with previous results. Violin plots of posterior distributions for the complete set of estimated TMRCA values for the 5 variants indicated in the legend scaled to a generation time of 29 years. Each row indicates a population sample from the 1000 Genomes Project panel. Replicate MCMCs are plotted with transparency. Points and lines overlaying the violins are previous point estimates and 95% confidence intervals for each of the variants indicated by a color and rs number in the legend (see Tables S.3, S.4, S.5, and S.6 in Appendix A.3). The population sample abbreviations are defined in text.

EDAR

Population genomic studies have repeatedly identified the gene EDAR to be under recent selection in East Asians [Akey et al., 2004, Voight et al., 2006, Williamson et al., 2005] with a particular site (rs3827760) showing strong evidence for being the putative target. Functional assays and allele specific expression differences at this position show phenotypic associations to a variety of phenotypes including hair thickness and dental morphology [Bryk et al., 2008, Fujimoto et al., 2008, Kamberov et al., 2013, Kimura et al., 2009].

Our estimate of 22,192 (19,683 - 25,736) years for the EDAR allele in the CHB sample is older than ABC-based estimates of 12,458 (1,314 - 85,835) and 13,224 (4,899 - 50,692) years made by Bryk et al. [2008] and Peter et al. [2012], respectively. Kamberov et al. [2013] use spatially explicit ABC and maximum likelihood approaches to compute older estimates of 30,925 (13,175 - 39,575) and 36,490 (34,775 - 38,204). We included all populations for which the variant is present including the FIN and BEB samples where it exists at low frequency. Our results for the youngest TMRCAs are found in these two low frequency populations, where the estimate in FIN is 17,386 (13,887 - 20,794) and the estimate in BEB is 18,370 (14,325 - 22,872). Among East Asian populations, the oldest and youngest TMRCA estimates are found in the KHV sample (25,683; 23,169 - 28,380) and CHB sample (22,192; 19,683 - 25,736).

LCT

Arguably the best studied signature of selection in humans is for an allele at the LCT gene (rs4988235) which confers lactase persistence into adulthood–a trait unique

among mammals and which is thought to be a result of cattle domestication and the incorporation of milk into the adult diets of several human populations [Bersaglieri et al., 2004, Enattah et al., 2002, Tishkoff et al., 2007]. There are multiple alleles that show association with lactase persistence [Tishkoff et al., 2007]. We focused on estimating the age of the T-13910 allele, primarily found at high frequency among Northern Europeans, but which is also found in South Asian populations. In addition to association with the lactase persistence phenotype, this allele has been functionally verified by *in vitro* experiments [Kuokkanen et al., 2006, Olds and Sibley, 2003, Troelsen et al., 2003].

Mathieson et al. [2015a] use ancient DNA collected from 83 human samples to get a better understanding of the frequency trajectory for several adaptive alleles spanning a time scale of 8,000 years. For the LCT persistence allele (rs4988235), they find a sharp increase in frequency in the past 4,000 years ago. While this is more recent than previous estimates, an earlier TMRCA or time of mutation is still compatible with this scenario.

Our estimates using European and South Asian samples fall between the range from 5000 to 10,000 years ago, which is broadly consistent with age estimates from modern data. The credible intervals for estimates in all of the samples have substantial overlap which makes any ranking on the basis of point estimates difficult. We infer the PJL (Punjabi from Lahore, Pakistan) sample to have the oldest TMRCA estimate of 9,514 (8,596 - 10,383) years. Itan et al. [2009] use spatial modeling to infer the geographic spread of lactase allele from northern to southern Europe. Consistent with their results, the youngest estimate among European populations

is found in the IBS sample at 9,341 (8,688 - 9,989) years. Among all samples, the youngest estimate was found in BEB at 6,869 (5,143 - 8809).

KITLG and OCA2

The genetic basis and natural history of human skin pigmentation is a well studied system with several alleles of major effect showing signatures consistent with being targets of recent selection [Beleza et al., 2013b, Eaton et al., 2015, Jablonski and Chaplin, 2012, Wilde et al., 2014]. We focused on an allele found at high frequency world-wide among non-African populations at the KITLG locus (rs642742) which shows significant effects on skin pigmentation differences between Europeans and Africans [Miller et al., 2007]; although more recent work fails to find any contribution of KITLG toward variation in skin pigmentation in a Cape Verde African-European admixed population [Beleza et al., 2013a]. We also estimated the TMRCA for a melanin-associated allele at the OCA2 locus (rs1800414) which is only found among East Asian populations at high frequency [Edwards et al., 2010].

For the KITLG variant, our estimates among different populations vary from 18,000 to 34,000 years ago, with the oldest age being in the YRI (Yoruba in Ibadan, Nigeria) sample (33,948; 28,861 - 39,099). The youngest TMRCA is found in FIN at 18,733 years (16,675 - 20,816). The next two youngest estimates are also found in Africa with the TMRCA in the MSL (Mende in Sierra Leone) sample being 22,340 (15,723 - 28,950) years old, and that for LWK (Luhya in Webuye, Kenya) being 22,784 (17,922 - 2,8012) years old, suggesting a more complex history than a model of a simple allele frequency increase outside of Africa due to pigmentation related

selection pressures. Previous point estimates using rejection sampling approaches on a Portuguese sample (32,277; 6,003 - 80,683) and East Asian sample (32,045; 6,032 - 98,165) are again most consistent with our own results on the IBS (29,731; 26,170 - 32,813) and CHB samples (26,773; 24,297 - 30,141) [Beleza et al., 2013b, Chen et al., 2015]. Among East Asians, the oldest and youngest estimates are again found in the JPT (28,637; 24,297 - 30,141) and KHV (24,544; 21,643 - 27,193) samples, respectively. The TMRCA for OCA2 alleles in the JPT (18,599; 16,110 - 20,786) and KHV (16370; 14,439 - 18,102) samples are also the oldest and youngest, respectively.

2.6 Discussion

Our method improves estimation for the timing of selection on a beneficial allele using a tractable model of haplotype evolution. This approach leverages detailed information in the data while remaining amenable to large sample sizes. Using both carriers and non-carriers of the allele, we can more effectively account for uncertainty in the extent of the ancestral haplotype and derived mutations. We show the performance of our method using simulations of different selection strengths, beneficial allele frequencies and choices of reference panel. By applying our method to five variants previously identified as targets of selection in human populations, we provide a comparison among population-specific TMRCAs. This gives a more detailed account of the order in which populations may have acquired the variant and/or experienced selection for the trait it underlies.

In that regard, it is hypothesized that local selection pressures and a cultural shift toward agrarian societies has induced adaptive responses among human populations

around the globe. The data associated with some variants seem to indicate more recent selective events than others. Our results for variants associated with dietary traits at the LCT and ADH1B genes both imply relatively recent TMRCAs ($< 15,000$ ya), consistent with hypotheses that selection on these mutations results from recent changes in human diet following the spread of agriculture [Peng et al., 2010a, Simoons, 1970]. In contrast, the inferred TMRCAs for EDAR, KITLG and OCA2 imply older adaptive events which may have coincided more closely with the habitation of new environments or other cultural changes.

Several hypotheses have been suggested to describe the selective drivers of skin pigmentation differences among human populations, including reduced UV radiation at high latitudes and vitamin D deficiency [Jablonski and Chaplin, 2000, Loomis, 1967]. Estimated TMRCAs for the variants at the OCA2 and EDAR loci among East Asians appear to be as young or younger than the KITLG variant, but older than the LCT and ADH1B locus. This suggests a selective history in East Asian populations leading to adaptive responses for these traits occurring after an initial colonization. In some cases, the dispersion of replicate MCMC estimates make it difficult to describe the historical significance of an observed order for TMRCA values. However, the consistency of estimates among different populations for particular variants add some confidence to our model's ability to reproduce the ages which are relevant to those loci or certain geographic regions.

To assess the relative concordance of our estimates with those from previous approaches, we compared our results to a compilation of previously published estimates based on the time of mutation, time since fixation, or TMRCA of variants

associated with the genes studied here (Figure S.8 in Appendix A.3). The range of confidence intervals for these studies is largely a reflection of the assumptions invoked or relaxed for any one method, as well as the sample size and quality of the data used. In principle, our method extracts more information than approaches that use summary statistics such as ABC. In our empirical application, we found that our method provides a gain in accuracy while accounting for uncertainty in both the ancestral haplotype and its length on each chromosome. Notably, our method provides narrower credible intervals by incorporating the full information from ancestral haplotype lengths, derived mutations, and a reference panel of non-carrier haplotypes.

Another caveat of our method is its dependence on the reference panel, which is intended to serve as a representative sample of non-ancestral haplotypes in the population during the selected allele's increase in frequency. Four possible challenges can arise: (1) segments of the ancestral selected haplotype may be present in the reference panel due to recombination, (this is more likely for alleles that have reached higher frequency), (2) the reference panel may contain haplotypes that are similar to the ancestral haplotype due to low levels of genetic diversity, (3) the reference panel may be too diverged from the focal population, and (4) population connectivity and turnover may lead the "local" reference panel to be largely composed of migrant haplotypes which were not present during the selected allele's initial increase in frequency.

Under scenarios 1 and 2, the background haplotypes will be too similar to the ancestral haplotype and it may be difficult for the model to discern a specific ancestry

switch location. This leads to fewer differences (mutations) than expected between the ancestral haplotype and each beneficial allele carrier. The simulation results are consistent with this scenario: our method tends to underestimate the true age across a range of selection intensities and allele frequencies when using a local reference panel.

Conversely, under scenarios 3 and 4 the model will fail to describe a recombinant haplotype in the sample of beneficial allele carriers as a mosaic of haplotypes in the reference panel. As a result, the model will infer more mutation events to explain observed differences from the ancestral haplotype. Our simulation results show this to be the case with reference panels diverged by N generations: posterior mean estimates are consistently older than their true value. Our simulations are perhaps pessimistic though - we chose reference panel divergence times of N and $0.5N$ generations, approximately corresponding to F_{ST} values of 0.4 and 0.2, respectively. For the smaller F_{ST} values observed in humans, we expect results for diverged panels to be closer to those obtained with the local reference panel. Nonetheless, future extensions to incorporate multiple populations within the reference panel would be helpful and possible by modifying the approach of Price et al. [2009]. Such an approach would also enable the analysis of admixed populations (we excluded admixed samples from our analysis of the 1000 Genomes data above).

Aside from the challenges imposed by the choice of reference panel, another potential source of bias lies in our transition probabilities, which are not conditioned on the frequency of the selected variant. In reality, recombination events at some distance away from the selected site will only result in a switch from the ancestral

to background haplotypes at a rate proportional to $1 - p_l$, where p_l is the frequency of the ancestral haplotype alleles at locus l. In this way, some recombination events may go unobserved – as the beneficial allele goes to high frequency the probability of an event leading to an observable ancestral to background haplotype transition decreases. One solution may be to include the frequency-dependent transition probabilities derived by Chen et al. (2015). Under their model, the mutation time is estimated by assuming a deterministic, logistic frequency trajectory starting at $\frac{1}{2N}$. An additional benefit of using frequency trajectories would be the ability to infer posterior distributions on selection coefficients. While the selection coefficient is typically assumed to be related to the time since mutation by $t_1 = log(Ns)/s$, we do not have an equivalent expression for time to the common ancestor. Rather than the initial frequency being $\frac{1}{2N}$ for a new mutation, our initial frequency must correspond to the frequency at which the TMRCA occurs. Griffiths and Tavare (1994) derive a framework to model a genealogy under arbitrary population size trajectories, which should be analogous to the problem of an allele frequency trajectory, and additional theory on intra-allelic genealogies may be useful here as well [Griffiths and Tavare, 1994, Slatkin and Rannala, 2000, Wiuf, 2000, Wiuf and Donnelly, 1999].

Our model also assumes independence among all haplotypes in the sample in a composite-likelihood framework, which is equivalent to assuming a star-genealogy [Larribe and Fearnhead, 2011, Varin et al., 2011]. This is unlikely to be the case when sample sizes are large or the TMRCA is old. It is also unlikely to be true if the beneficial allele existed on multiple haplotypes preceding the onset of selection, was introduced by multiple migrant haplotypes from other populations, or occurred by

multiple independent mutation events [Berg and Coop, 2015, Hermisson and Pennings, 2005, Innan and Kim, 2004, Prezeworski et al., 2005, Pritchard et al., 2010]. Methods for distinguishing selection from standing variation versus de novo mutation are available that should make it easier distinguish these cases [Garud et al., 2015, Messer and Neher, 2012, Messer and Petrov, 2013, Peter et al., 2012].

If the underlying allelic genealogy is not star-like, one can expect different estimates of the TMRCA for different subsets of the data. Here, we performed multiple MCMCs on resampled subsets of the data to informally diagnose whether there are violations from the star-like genealogy assumption. We speculate that exactly how the TMRCAs vary may provide insight to the underlying history. In cases where the TMRCA estimates for a particular population are old and more variable than other populations, the results may be explained by structure in the genealogy, whereby recent coalescent events have occurred among the same ancestral haplotype before the common ancestor. When estimates are dispersed among resampled datasets the presence of multiple ancestral haplotypes prior to the variant's increase in frequency may be a better explanation. Further support for this explanation might come from comparisons to other population samples which show little to no dispersion of estimates from resampled datasets. Future work might make it possible to formalize this inference process.

A final caveat regards the misspecification of mutation and recombination rates. TMRCA estimates are largely determined by the use of accurate measures for these two parameters. In a way, this provides some robustness to our method. Our age estimates depend on mutation and recombination rates, so accurate specification

for one of the values can compensate for slight misspecification in the other. As previously noted, in cases where a fine-scale recombination map is unavailable we suggest using a uniform recombination rate specific to the locus of interest (Figure **??** in Appendix). Choosing an appropriate mutation rate will continue to depend on current and future work which tries to resolve discrepancies in published mutation rate estimates inferred by various approaches [Ségurel et al., 2014].

One future direction for our method may be to explicitly incorporate the possibility of multiple ancestral haplotypes within the sample. Under a disease mapping framework, Morris et al. [2002] implement a similar idea in the case where independent disease causing mutations arise at the same locus leading to independent genealogies, for which they coin the term "shattered coalescent". For our case, beneficial mutations may also be independently derived on different haplotypes. Alternatively, a single mutation may be old enough to reside on different haplotypes due to a sufficient amount of linked variation existing prior to the onset of selection. Berg and Coop [2015] model selection from standing variation to derive the distribution of haplotypes that the selected allele is present on.

While we have treated the TMRCA as a parameter of interest, our method also produces a sample of the posterior distribution on the ancestral haplotype. This could provide useful information to estimate the frequency spectrum of derived mutations on the ancestral haplotype. Similarly, the frequency of shared recombination breakpoints could shed light on the genealogy and how well it conforms to the star-shape assumption. The extent of the ancestral haplotype in each individual may also prove useful for identifying deleterious alleles that have increased in frequency as a

result of strong positive selection on linked beneficial alleles [Chun and Fay, 2011, Hartfield and Otto, 2011]. For example, Huff et al. [2012] describe a risk allele for Crohn's disease at high frequency in European populations which they suggest is linked to a beneficial allele under recent selection. Similar to an admixture mapping approach, our method could be used to identify risk loci by testing for an association between the ancestral haplotype and disease status. As another application, identifying the ancestral haplotype may be useful in the context of identifying a source population (or species) for a beneficial allele prior to its introduction and subsequent increase in frequency in the recipient population (see Chapter 3).

In many cases, the specific site under selection may be unknown or restricted to some set of putative sites. While our method requires the position of the selected site be specified, future extensions could treat the selected site as a random variable to be estimated under the same MCMC framework. This framework would also be amenable to marginalizing over uncertainty on the selected site.

While we focus here on inference from modern DNA data, the increased accessibility of ancient DNA has added a new dimension to population genetic datasets [Allentoft et al., 2015, Haak et al., 2015, Lazaridis et al., 2014, Mathieson et al., 2015a,b, Skoglund et al., 2014]. Because it will remain difficult to use ancient DNA approaches in many species with poor sample preservation, we believe methods based on modern DNA will continue to be useful going forward. That said, ancient DNA is providing an interesting avenue for comparative work between inference from modern and ancient samples. For example, Nakagome et al. [2016] use simulations to assess the fit of this ancient DNA polymorphism to data simulated under their in-

ferred parameter values for allele age and selection intensity and they find reasonable agreement. Much work still remains to fully leverage ancient samples into population genetic inference while accounting for new sources of uncertainty and potential for sampling bias.

Despite these challenges, it is clear that our understanding of adaptive history will continue to benefit from new computational tools which extract insightful information from a diverse set of data sources.

2.7 Materials and Methods

We generated data using the software mssel (Dick Hudson, personal communication), which simulates a sample of haplotypes conditioned on the frequency trajectory of a selected variant under the structured coalescent [Hudson and Kaplan, 1988, Kaplan et al., 1988]. Trajectories were first simulated forwards in time under a Wright-Fisher model for an additive locus with varying strengths of selection and different ending frequencies of the selected variant. Trajectories were then truncated to end at the first time the allele reaches a specified frequency. See Table S.1 in Appendix A.4 for relative ages of simulated TMRCA values for different end frequencies and selection strengths. For the results in Figure 2.2, 100 simulations were performed for each parameter combination. MCMCs were run for 5000 iterations with a burn-in excluding the first 3000 iterations. A standard deviation of 10 was used for the proposal distribution of t_{ca}. The red boxplots indicate local reference panels. The blue and green boxplots indicate reference panels diverged by .5Ne generations and 1Ne generations, respectively. Each data set was simulated for a 1 Mbp locus with a

mutation rate of 1×10^{-8}, recombination rate of 1×10^{-8} and population size of 10000. Sample sizes for the selected and reference panels were 100 and 20, respectively.

For more efficient run times of the MCMC, we set a maximum number of individuals to include in the selected and reference panels to be 100 and 20, respectively. In cases where the true number of haplotypes for either panel was greater than this in the full data set, we resampled a subset of haplotypes from each population for a total of five replicates per population. For simulation results supporting the use of this resampling strategy, see Figure S.5 in Appendix A.3. The MCMCs were run for 15000 iterations with a standard deviation of 20 for the $TMRCA$ proposal distribution. The first 9000 iterations were removed as a burn-in, leading to 6000 iterations for a sample of the posterior. Convergence was assessed by comparison of MCMC replicates. Figure 2.3 and Figure S.8 in Appendix A.3 show the results for all five variants along with previous point estimates and 95% confidence intervals assuming a generation time of 29 years [Fenner, 2005]. Tables S.3 and S.4 in Appendix A.4 list the mean and 95% credible intervals for estimates with the highest mean posterior probability which we refer to in the text. Tables S.5 and S.6 in Appendix A.4 list the previous estimates and confidence intervals with additional details of the different approaches taken.

To model recombination rate variation, we used recombination rates from the Decode sex-averaged recombination map inferred from pedigrees among families in Iceland [Kong et al., 2010]. Because some populations may have recombination maps which differ from the Decode map at fine scales, we used a mean uniform recombination rate inferred from the 1 megabase region surrounding each variant.

The motivation for this arises from how recombination rates have been previously shown to remain relatively consistent among recombination maps inferred for different populations at the megabase-scale [Auton and McVean, 2012, Baudat et al., 2010, Broman et al., 1998, Kong et al., 2010]. Further, we found our estimates depend mostly on having the megabase-scale recombination rate appropriately set, with little difference in most cases for estimates obtained by modeling fine-scale recombination at each locus (Figure S.7 in Appendix A.3). We specify the switching rate among background haplotypes after recombining off of the ancestral haplotype to be $4Nr$, where $N = 10,000$ and r is the mean recombination rate for the 1Mb locus.

For modeling mutation, a challenge is that previous mutation rate estimates vary depending on the approach used [Ségurel et al., 2014]. Estimates using the neutral substitution rate between humans and chimps are more than 2×10^{-8} per bp per generation, while estimates using whole genome sequencing are closer to 1×10^{-8}. As a compromise, we specify a mutation rate of 1.6×10^{-8}.

Acknowledgements

We would like to thank Hussein Al-Asadi, Arjun Biddanda, Anna Di Rienzo, Dick Hudson, Choongwon Jeong, Evan Koch, Joseph Marcus, Shigeki Nakagome, Ben Peter, Mark Reppel, Alex White, members of the Coop lab at UC Davis, members of the Przeworski lab at Columbia University, and members of the He and Stephens labs at the University of Chicago for helpful comments. JS was supported by an NSF Graduate Research Fellowship and National Institute Of General Medical Sci-

ences of the National Institutes of Health under award numbers DGE-1144082 and T32GM007197, respectively. This work was also supported by the National Institute of General Medical Sciences of the National Institutes of Health under award numbers RO1GM83098 and RO1GM107374 to GC, as well as R01HG007089 to JN.

CHAPTER 3

ESTIMATING THE TIMING OF ADAPTIVE INTROGRESSION

Joel Smith[1], Rena Schweizer[2,3], Olivia Gray [4], Robert Wayne [3], Anna Di Rienzo [4], John Novembre[1,4]

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

2 Division of Biological Sciences, University of Montana, Missoula, MT

3 Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA

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

3.1 Abstract

There are a growing number of studies which describe genomic patterns of adaptation that are consistent with beneficial alleles being shared between species through hybridization. These patterns can also be used to describe the time at which the allele was introduced into the receiving species or population. Below, we use our method of estimating the time to the most recent common ancestor (described in Chapter 2) as a proxy for the timing of adaptive introgression. When assumptions of the model are reasonably approximated by the data, this approach leads to dramatic gains in accuracy relative to more commonly used Approximate Bayesian Computation methods. We describe these assumptions and present results for the timing of coat color adaptation in North American wolves and high altitude adaptation in

Tibetans.

3.2 Introduction

Adaptation is often described as either occurring by selection on genetic variants introduced by recent mutation or older genetic variants at intermediate frequency. The distinction between these two scenarios is known to have important implications for the resulting patterns of genetic diversity around the selected allele [Berg and Coop, 2015]. While these models are typically considered in the context of a single randomly mating population, population structure can lead to other scenarios which might be viewed as a combination of these two patterns.

An older beneficial allele may reach high frequency in one population, but in cases where migration events among populations of a species are rare, the introduction of a beneficial variant from one population to another is analogous to a new mutation having occurred in the receiving population. Subsequent propagation of this variant among individuals in the receiving population leads to a distinct pattern of variation around the selected allele: the ancestral haplotype of the donor population, or linked sequence of genetic variants on which the beneficial allele resides, will be at high frequency and embedded within a genomic background of the receiving population. This process of repeated crossing of a favored allele into a new genomic background is a natural analog to the artificial selection technique used in plant and animal breeding known as introgression [Anderson et al., 2009, Burbank, 1921, Mendez et al., 2012, Simmonds, 1993, Song et al., 2011].

A more strict definition of adaptive introgression will further specify that the

donor and receiving populations be different species. In practice, the pattern of adaptive introgression is most recognizable when there is sufficient sequence divergence between the introduced haplotype and its new genomic background. This is more likely to be the case if some degree of reproductive isolation exists between the populations. In an effort to avoid making species designations, we will use the pattern of an ancestral haplotype with shared ancestry from another population (or species) as criteria for labeling the process as introgression.

Studies which aim to describe this pattern can use the length of the introgressed haplotype and a count of derived mutations on its background to infer the time to the common ancestor (TMRCA). When the beneficial allele is strongly selected, the TMRCA indicates the time at which the allele began increasing in frequency after first arriving in the population. Here, we present estimates for the timing of introgression using the method we describe in Chapter 2.

Our model assumes that the selected allele resides on a single ancestral haplotype. One complication that may arise in the context of introgression is that admixture events involving many individuals may introduce multiple ancestral haplotypes into the receiving population. This could lead to an over-counting of derived mutations which would yield an over-estimate for the TMRCA. Frequent admixture would, however, lead to larger admixture proportions from the donor population across the genome. This pattern might serve as an indicator for some degree of model-misspecification. Keeping this caveat in mind, we estimated the timing of introgression events for two recently described examples of coat color adaptation in North American wolves and high altitude adaptation in Tibetans.

3.3 Coat Color Adaptation in North American Wolves

Interest in understanding the evolutionary history of coat color polymorphism in North American wolves was initially prompted by the observation that coat color varied on a latitudinal cline with light coat color being associated with open tundra habitat in the north, and dark coat color being more strongly associated with southern forest habitat [Gipson et al., 2002]. This polymorphism was subsequently found to co-segregate with a 3-bp deletion at the K locus (*CBD103*) which leads to dominant inheritance of the dark phenotype [Anderson et al., 2009]. The allele conferring dark coat color (K^B) occurs on a single haplotype, defined on a scale of 100 kb, that is shared among domesticated dogs and wolves. In addition to conferring dark coat color in canids, the K locus is known to be functionally relevant to microbial immune response [Pazgier et al., 2006, Yang et al., 1999]. Current evidence suggests that the K^B allele provides a heterozygote advantage to individuals at lower latitudes where the risk of infection by pathogens is higher [Coulson et al., 2011, Ducrest et al., 2008, Stahler et al., 2013].

Using a count of derived mutations, initial estimates for the TMRCA of the K^B allele in dogs was found to be comparable to estimates among dog and wolf chromosomes when considered together (46,886 years 95% confidence interval: 12,779 - 121,182, assuming a mutation rate of 1×10^{-9} and generation time of 3 years). The K^B allele TMRCA among wolves was not formally estimated, but a comparison of genetic diversity around the K locus indicated that the estimated time ranged from 500 to 14 kya, consistent with the hypothesis that the K^B allele was introduced into North American wolf populations by domesticated dogs of Native Americans. While

these haplotype patterns also suggest that natural selection increased the frequency of the K^B allele, further analyses and data were necessary to provide a more detailed picture of the evolutionary history of this locus.

Schweizer et al. [2018] performed targeted capture sequencing at the K locus for a larger sample of wolf populations across North America. The first goal was to validate previous evidence for natural selection operating on the K^B allele. This was done using comparisons of nucleotide diversity (π), Tajima's D, Watterson's estimator (θ_w) and extended haplotype homozygosity (EHH) between the ancestral allele (K^y) and the derived K^B allele [Nei and Li, 1979, Sabeti et al., 2002, Tajima, 1989, Watterson, 1975]. All of these metrics confirmed an adaptive hypothesis whereby a single haplotype was introduced from domesticated dogs into wolves and increased in frequency leading to a selective sweep signature.

To describe the relative timing of the K^B allele's spatial spread and/or timing of selection among wolf populations in North America, we estimated the TMRCA for the deletion in each of the four samples using our method (described in Chapter 2) which leverages both the decay in LD between the selected allele and nearby sites, as well as the accumulation of new mutations on the selected allele's ancestral haplotype [Smith et al., 2018]. Provided that assumptions of the model are reasonably approximated by the data, this approach leads to dramatic gains in accuracy relative to more commonly used Approximate Bayesian Computation (ABC) methods. Specifically, the method assumes a "star-shaped" genealogy among sites linked to the selected allele's ancestral haplotype. This is a reasonable assumption in cases where the focal allele is subject to strong positive selection; which for our case is

appropriate given the results of reduced nucleotide diversity, Tajima's D and EHH scores observed at the K locus.

Another assumption involves the specification of an appropriate reference panel of haplotypes that do not have the selected allele. This reference panel should approximate the background haplotypes with which the selected haplotype recombined during its increase in frequency. Simulation results show that a reference panel which is too similar to the selected haplotype can lead to under-estimates of the true TMRCA while a misspecified reference panel that is too diverged from the true reference panel will over-estimate the TMRCA (Figure 2.2). For this reason, we excluded the sample of dogs for which a suitable reference panel was not available. We used the local reference panel of haplotypes without the selected allele in each of the four samples from natural populations of North American wolves.

TMRCA estimates can vary depending on the mutation and recombination rates used. To account for locus specific variation in recombination rates across the K locus region, we used a recombination map inferred for dogs based on patterns of LD [Auton et al., 2013]. This recombination map, however, does not include the entire sequenced region downstream of the selected site. We predicted the unobserved recombination rates at these sites using the adjacent 4 Mb of observed recombination rates and the smooth spline function in R with the smoothing parameter set to 0.95.

Several estimates of the per basepair per generation mutation rate have been inferred using different approaches. Skoglund et al. [2015] use ancient DNA from a 35 ky old wolf and to infer a rate of 0.4×10^{-8} per basepair per generation. Frantz et al. [2016] calibrate a molecular clock using radiocarbon dating on an ancient dog

to infer similar a mutation rate between 0.3×10^{-8} and 0.45×10^{-8} per basepair per generation. We report TMRCA estimates using these values in addition to one higher rate of 1×10^{-8} per basepair per generation.

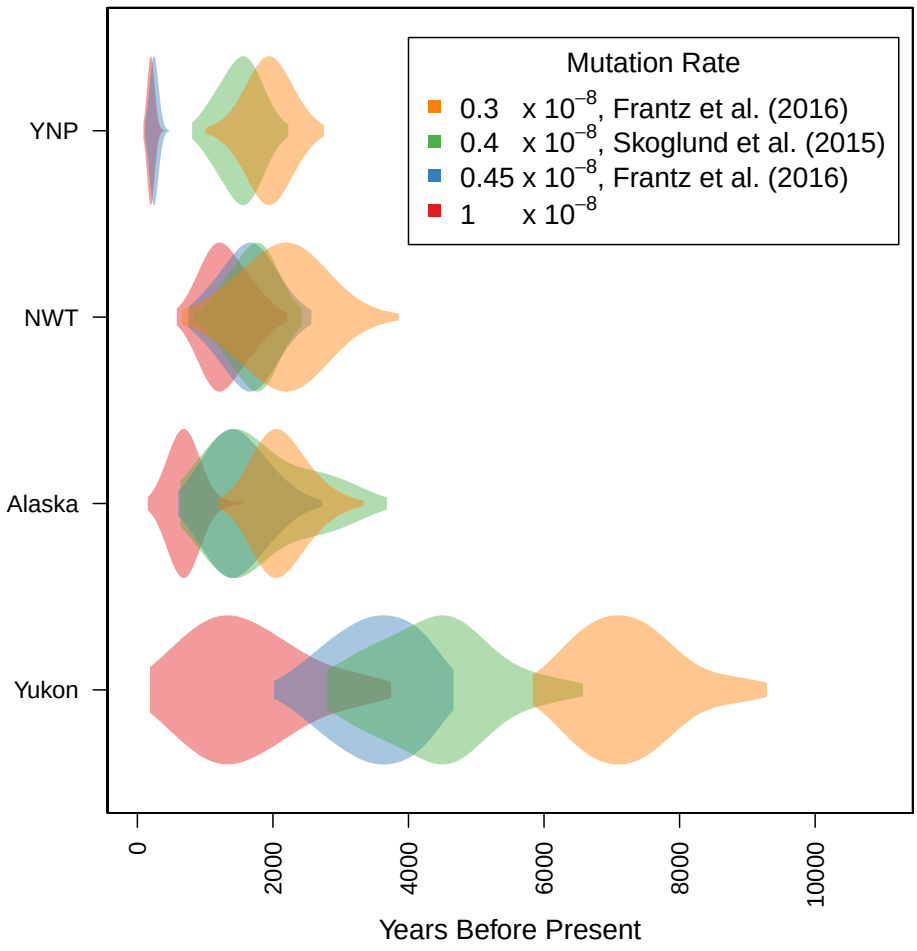

Figure 3.1: TMRCA estimates of the K^B allele in the 4 North American populations using 4 different mutation rates assuming a generation time of 3 years. The mutation rates in the legend are in units of per basepair per generation. The violin plots are samples from the posterior distribution of TMRCAs drawn from a Markov chain Monte Carlo run for 50000 iterations with a standard deviation of 10 for the proposal distribution. The locus includes 3 Mbp of flanking sequence around the selected site.

Population	Sample Size		Mutation Rate	TMRCA	95% C.I.
	Sel.	Ref.			
Alaska	6	20	1×10^{-8}	694	376 - 1100
Alaska	6	20	0.45×10^{-8}	1477	831 - 2329
Alaska	6	20	0.4×10^{-8}	1801	826 - 3263
Alaska	6	20	0.3×10^{-8}	2135	1602 - 2951
NWT	9	20	1×10^{-8}	1301	822 - 1931
NWT	9	20	0.45×10^{-8}	1635	932 - 2333
NWT	9	20	0.4×10^{-8}	1704	1137 - 2234
NWT	9	20	0.3×10^{-8}	2155	1105 - 3250
YNP	16	20	1×10^{-8}	202	133 - 285
YNP	16	20	0.45×10^{-8}	250	162 - 352
YNP	16	20	0.4×10^{-8}	1513	982 - 2030
YNP	16	20	0.3×10^{-8}	1942	1346 - 2503
Yukon	2	8	1×10^{-8}	1598	393 - 3389
Yukon	2	8	0.45×10^{-8}	3500	2338 - 4398
Yukon	2	8	0.4×10^{-8}	4378	2979 - 6169
Yukon	2	8	0.3×10^{-8}	7248	6219 - 8963

Table 3.1: Mean posteriors and credible intervals for TMRCA estimates (in years) of the K^B allele in the 4 North American populations using 4 different mutation rates assuming a generation time of 3 years. Mutation rates are per basepair per generation.

Providing TMRCA estimates among different samples can shed light on the relative order in which the mutation spread through North American wolf populations. We find a relatively consistent order of TMRCA estimates, where the oldest values are found in Yukon with posterior means ranging from 1598 to 7248 ya, depending on the mutation rate used (See Figure 3.1 and Table 3.1). The youngest TMRCA values are consistently found in Yellowstone, ranging from 202 to 1942 ya. Estimates for Alaska, Northwest Territories, and two Yellowstone posteriors have significant overlap and all fall roughly between 694 and 2135 ya.

In some cases, the choice of different mutation rates did not affect the variability of time estimates (Table 3.1). Among the Alaska and Northwest Territories samples, mutation rates from 0.4×10^{-8} to 0.45×10^{-8} impose similar time estimates. The same is true for the Yellowstone samples at mutation rates of 1×10^{-8} and 0.45×10^{-8}. However, all estimates using a mutation rate of 1×10^{-8} did result in younger TMRCA values.

3.4 High Altitude Adaptation in Tibetans

Much of our current progress toward understanding recent human evolution has been driven by the increased scale of sequencing efforts and methods development which continue to describe wide representations of human genetic diversity [Cann et al., 2002, Genomes Project Consortium, 2012, HapMap Consortium, 2003]. Genome scans for signatures of natural selection have provided long lists of candidate loci, many of which have known relevance to phenotypes of interest that are specific to particular populations or geographic regions [Akey et al., 2002, Coop et al., 2009, Enard et al., 2014, Johnson and Voight, 2018, Liu et al., 2013, Pickrell et al., 2009, Sabeti et al., 2002, Voight et al., 2006]. These discoveries have prompted follow-up studies which aim to identify the functional mechanisms which underly these adaptations, as well as characterize their demographic and evolutionary context [Fumagalli et al., 2015, Kamberov et al., 2013, Tishkoff et al., 2007]. One of the most well-known cases is that of Tibetans' high altitude adaptation to hypoxia [Huerta-Sánchez et al., 2014, Jeong et al., 2014].

The low-oxygen environments found at high altitude in which many human pop-

ulations have persisted are known to have negative fitness consequences for childbearing individuals with ancestral origins from low-altitude populations [Moore et al., 2004, 2001, Niermeyer et al., 2009]. Some of the consequences which have been described include lower birth weight and an increased rate of hypertension during pregnancy relative to individuals with high-elevation ancestry. The genetic architecture for high altitude adaptation is known to involve a variety of genes which affect several traits; however, the transcription factor *EPAS1* shows the strongest signature of recent selection among Tibetans [Beall et al., 2010, Bigham et al., 2010, Huerta-Sánchez et al., 2014, Peng et al., 2010b, Simonson et al., 2010, Wang et al., 2011, Xu et al., 2010, Yi et al., 2010].

To more carefully characterize the haplotype patterns at this locus, Huerta-Sánchez et al. [2014] resequenced 40 Tibetan and 40 Han Chinese individuals and found that the *EPAS1* selected haplotype is highly differentiated with respect to the Han Chinese and all other modern populations in the Human Genome Diversity Panel. Interestingly, the only sample with a nearly identical haplotype was from an archaic Denisovan individual which was found in the Altai Mountains of Southern Siberia [Reich et al., 2010]. The proposed demographic model most consistent with this pattern is a scenario in which admixture between Denisovan individuals which had already adapted to hypoxic environments and an ancestral Tibetan population led to a selective sweep for the beneficial *EPAS1* haplotype. The resulting haplotypic signature is characteristic of adaptive introgression, whereby a highly differentiated population has contributed genetic variation which is beneficial to the receiving population. The introduction and propagation of this distinct haplotype among Tibetan

individuals has driven it to high frequency in the Tibetan plateau and nowhere else.

We estimated the TMRCA for 10 candidate SNPs at the *EPAS1* locus in a sample of 59 Tibetans (Figure 3.2). This set of variants was identified by both GWAS and selection scans to be the best functional candidates to modulate *EPAS1* activity. 3 SNPs from this set (rs188801636, rs76242811 and rs375554942) are found in an upstream enhancer of *EPAS1* and show several signs of experimental validation for functional relevance (Di Rienzo Lab at University of Chicago, personal communication) including: (1) transcriptional differences between alleles using a luciferase reporter assay on transformed epithelial cells, (2) Hi-C experiments which indicate that this enhancer is looping to touch the *EPAS1* transcription start site, and (3) ATAC-seq results confirming that this enhancer is in a region of open chromatin.

To compute samples from the posterior distribution of the TMRCA, we used a mutation rate of 1.6×10^{-8} per basepair per generation and estimated a mean recombination rate of 1.4×10^{-8} per basepair per generation from the sex-averaged Decode recombination map across the 1 Mbp region that we considered [Kong et al., 2010]. We generated 5 replicate MCMCs for each SNP to ensure convergence among runs. Assuming a generation time of 29 years, mean posterior estimates are approximately 18850 years ago (95% credible interval: 16907 - 21257). Estimates are consistent across all SNPs in large part due to the high levels of linkage disequilibrium between them on the selected haplotype. Results for each SNP are summarized in Table 3.2.

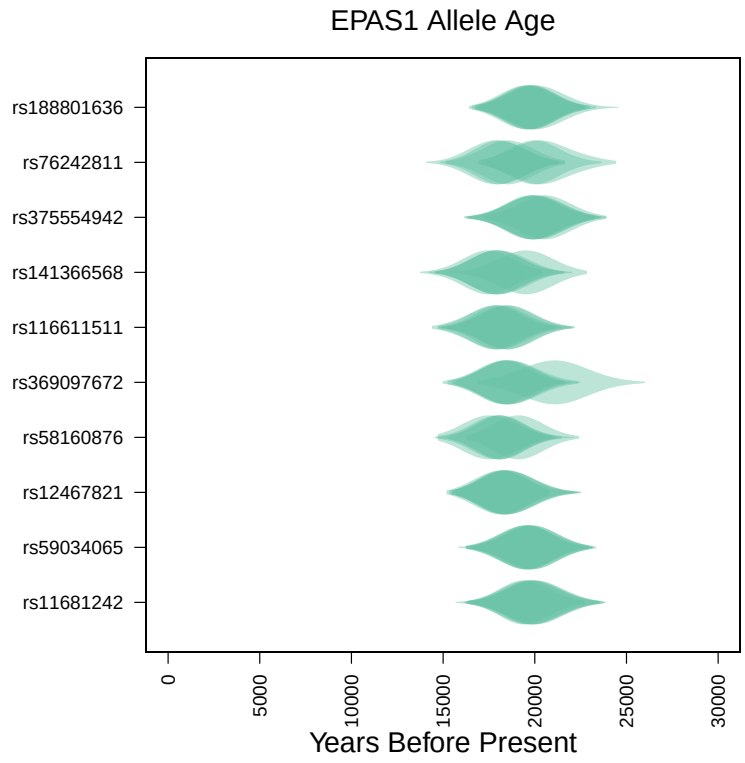

Figure 3.2: Comparison of TMRCA estimates for the candidate SNPs assuming a generation time of 29 years. The violin plots are samples from the posterior distribution of TMRCAs drawn from a Markov chain Monte Carlo run for 15000 iterations with a burn-in of 10000 iterations and standard deviation of 10 for the proposal distribution. Replicate MCMC runs are plotted with transparency.

3.5 Discussion

While our method to estimate the TMRCA of a beneficial variant was originally intended to describe adaptations for the same population in which the mutation occurred, many cases of adaptive introgression remain amenable to describing the timing of introgression. The primary cause of model misspecification for cases of adaptive introgression would be the result of ongoing or more frequent admixture

	Sample Size			
SNP ID	Sel.	Ref.	TMRCA	95% C.I.
rs188801636	97	21	19815	17594 - 22155
rs76242811	97	21	19660	17512 - 21843
rs375554942	95	23	18419	16369 - 20672
rs141366568	90	28	18200	16191 - 20345
rs116611511	90	28	19022	16798 - 21403
rs369097672	90	28	18180	16057 - 20355
rs58160876	90	28	18190	16118 - 20325
rs12467821	93	25	20090	17842 - 22357
rs59034065	90	28	19015	16820 - 21358
rs11681242	93	25	19786	17656 - 21971

Table 3.2: A summary of the results from Figure 3.2. TMRCA estimates and 95% credible intervals are mean estimates across MCMC replicates scaled to a generation time of 29 years.

events which contribute more than one ancestral haplotype. In addition to being a violation of the "star-shaped" genealogy assumption, multiple ancestral haplotypes would lead to an overestimate of the TMRCA due to a greater number of ancestral variants being counted as derived mutations. For the human and wolf applications considered here, there is sufficient evidence to show that admixture events which contributed beneficial variants to the receiving populations were rare enough to only contribute a single ancestral haplotype [Huerta-Sánchez et al., 2014, Schweizer et al., 2018]. In this way, admixture events which lead to adaptive introgression are analogous to mutation events for non-introgression scenarios of adaptation.

The rarity of admixture events among domesticated dogs and North American wolves is corroborated by the introgression timing estimates which, when considering all four mutation rates, place the oldest mean TMRCAs between 1598 to 7248 years

ago in the Yukon population. Domesticated dogs arrived in North America 8.5k to 10k years ago and there was limited European contact before the 19th century, which suggests that the only dogs contributing to *K* locus variation to North American wolves were those from Native Americans [Leonard et al., 2002]. The population density of Native Americans (and their dogs) was relatively low, which is consistent with the observation of a single ancestral haplotype across North American wolf samples, possibly derived from a single introgression event.

The demographic history of modern Tibetans has received much recent attention due to a combination of studies which have described their unique adaptations to high altitude and their complicated history of admixture among several modern and archaic populations [Aldenderfer, 2011, Jeong and Di Rienzo, 2014, Lu et al., 2016]. Archaeological and genetic data from mitochondria and Y chromosomes have suggested that the Tibetan plateau may have been colonized as early as 30,000 years ago; and while the *EPAS1* haplotype shows close resemblance to the Denisovan sample, genome-wide patterns of ancestry suggest that Tibetans show levels of archaic admixture which are comparable to other East Asian populations [Jeong et al., 2014].

Jeong et al. [2014] used genotype data from 540,000 SNPs to describe patterns of relatedness in a sample of 69 Sherpa, 96 Tibetans and additional world-wide samples from the HGDP (Human Genome Diversity Panel) and HapMap phase 3 (HapMap3) datasets. They find Tibetans to be an admixed population of Sherpa and lowland East Asian populations. Because this admixture event is estimated to have occurred relatively recently (400 - 600 years ago), genetic data from the Sherpa serves as a proxy to study the ancestral high altitude population that contributed to Tibetan

ancestry. To do so, Jeong et al. [2014] inferred the history of population size changes using the Pairwise Sequentially Markovian Coalescent (PSMC) on whole-genome sequence data from two Sherpa and two low altitude East Asians (Han and Dai), and find that the population size trajectories between the high altitude and low altitude individuals begin to diverge around 40,000 years ago. When considering the X chromosomes, split-time estimates are closer to 20,000 years ago. Together, these estimates are consistent with previous colonization times of around 30,000 years.

Despite our own estimates for the TMRCA of the derived *EPAS1* allele being on the early side of this range, this timeline can be explained by either early colonization of the Tibetan plateau with a delayed arrival of the beneficial allele, or by its maintenance at low frequency until increasing around 20,000 years ago. The consistency of our estimates among replicate resampled haplotypes suggests that the star-genealogy assumption is well-approximated by the data. Based on simulation results and other applications of this method in humans (see Chapter 2), older or less strongly selected alleles with genealogical structure are more likely to yield different TMRCA estimates from different subsamples of the data.

Moving forward, ancient DNA samples for both the North American canid and Tibetan populations will likely provide even more evidence to resolve their adaptive and demographic histories. While these two applications are well-suited to the underlying assumptions of our model for estimating the TMRCA of beneficial alleles, future studies that are not consistent with our model will require more thorough simulation and testing to determine the kinds of biases that might arise from multiple ancestral haplotypes which enter the receiving population at one or multiple times.

Acknowledgements

This work was supported by an 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 JS.

CHAPTER 4

EXPECTED PATTERNS OF LOCAL ANCESTRY IN A HYBRID ZONE

Joel Smith[1] and John Novembre[1,2]

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

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

4.1 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.

4.2 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 de-

pends 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, 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 against derived and ancestral allele combinations. 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.

4.2.1 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 4.1 and Table 4.1). Much of our current understanding for how relevant any of these genetic architectures might be in nature has been driven by theoretical arguments and simulations. There are, however, a modest number of 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].

		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 4.1: 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 allele substitutions fix at two different loci among populations that are geographically isolated. The top row in Figure 4.1 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 diversity of fitness configurations that might result from this model. The fitness matrix in Figure 4.1a is an example where the the derived substitutions were fixed by positive selection, such that the ancestral genotype suffers a fitness cost. Figures 4.1a and 4.1b are examples where the derived alleles interact dominantly; whereas in Figure 4.1c, derived alleles interact recessively.

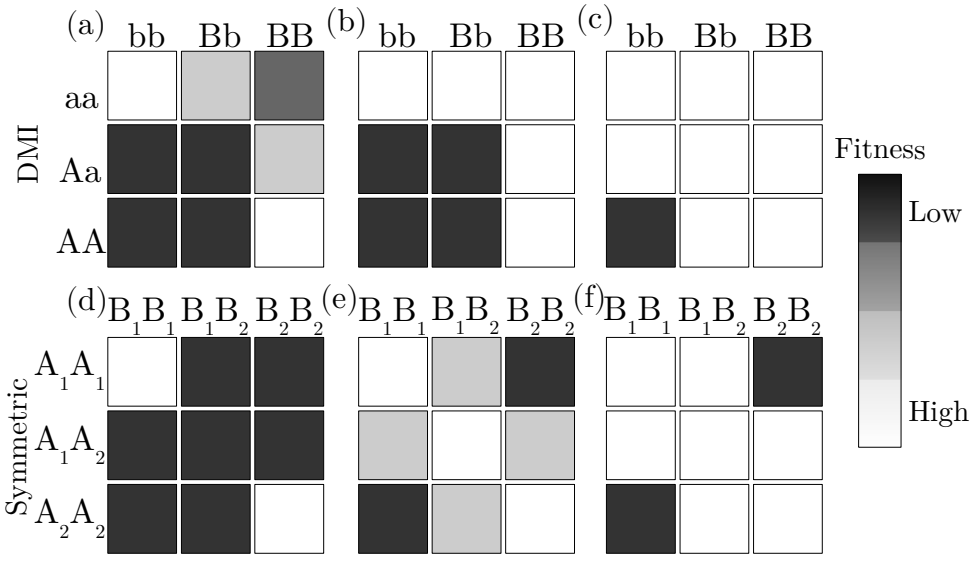

Figure 4.1: Two-locus fitness matrices for six models of genetic incompatibility. Each matrix includes the fitnesses of all possible two-locus genotypes where each locus is biallelic. Shaded boxes represent genotypes with a fitness cost that varies positively with the amount of shading. The top row of 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.

Lindtke and Buerkle [2015] draw attention to a different model of genetic incom-

patibility where allele substitutions occur at two loci in both populations leading to a symmetric pattern of fitnesses between the two derived genotypes $\mathbf{A_1A_1B_1B_1}$ and $\mathbf{A_2A_2B_2B_2}$ (Figure 4.1d, 4.1e, 4.1f). 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 4.1 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 4.2 (using the software dfuse from Lindtke and Buerkle [2015]) illustrates the effect of the DMI model in Figure 4.1b 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.

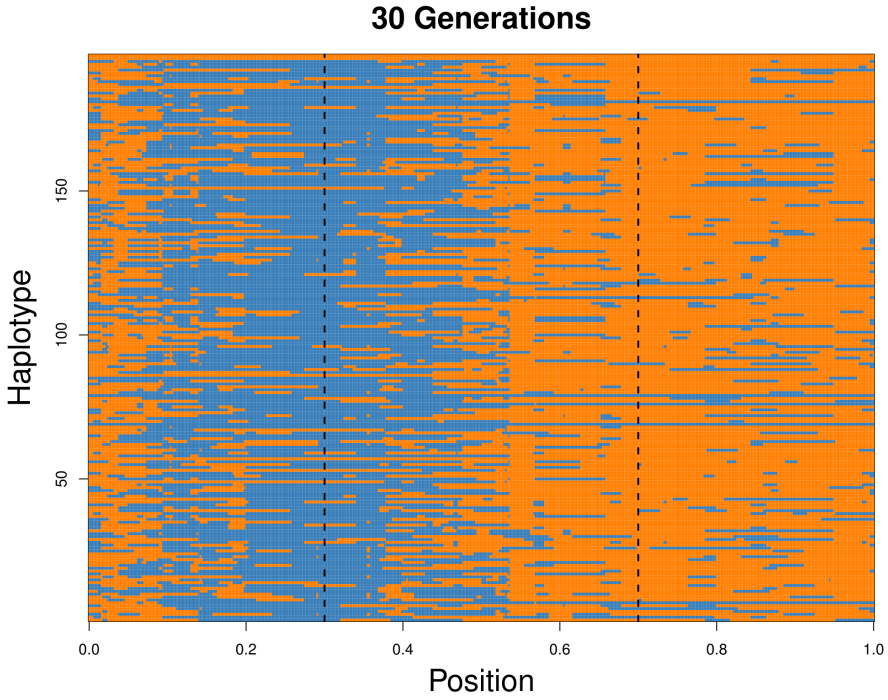

Figure 4.2: Haplotype data simulated using the software dfuse with the fitness matrix in Figure 4.1b. The forward-in-time simulation begins with two infinite source populations contributing equal fractions of ancestry (0.5) to a target population of 100 individuals 30 generations in the past. Each generation 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 interacting loci are indicated by the vertical dotted lines.

4.3 Model Description

4.3.1 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

$$P(p, t|\tau) = m_p(t) \prod_{t'=\tau+1}^{t-1} (1 - m(t')). \tag{4.1}$$

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

$$R(p, t \to p', t') = \sum_{\tau=1}^{min(t,t')-1} \frac{P(p', t'|\tau)}{(t-1)} \tag{4.2}$$

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:

$$Q(p, t \to p', t') = m_{p'}(t') \sum_{\tau=1}^{min(t,t')-1} \prod_{s=\tau+1}^{t'-1} (1 - m(s)). \tag{4.3}$$

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:

$$\mathbb{E}_{k,Q_0}(l) = \frac{Q_0^k l^{k-1} e^{-Q_0 l}}{(k-1)!} \tag{4.4}$$

This leads to the tract length distribution:

$$\phi(l) = \sum_{k=1}^{\Lambda+1} b_k \, \mathbb{E}_{k,Q_0}(l) \tag{4.5}$$

4.3.2 A Locus-Specific Tract Length Distribution With Selection

Equation 4.5 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 4.3). 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:

$$Q_v(p, t \to p', t') = m_{p'}(t') \sum_{\tau=1}^{min(t,t')-1} \Psi_v^\tau \prod_{s=\tau+1}^{t'-1} (1 - m(s)). \tag{4.6}$$

Equation 4.6 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 4.3 provides some intuition for how the following transition probabilities in $\mathbf{P}_{,,}^{t,t-1}$ are computed.

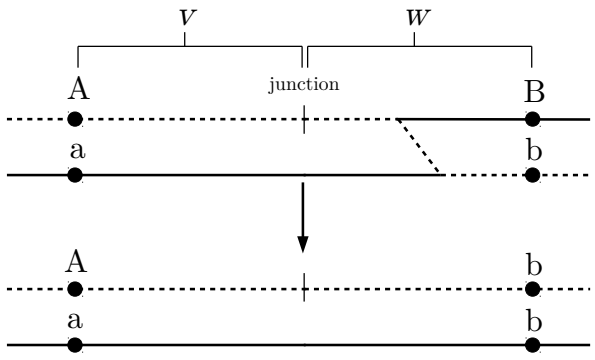

Figure 4.3: A visual description of the transition probability $\mathbf{P}_{1,5}^{t,t-1}$. For the first state in S, **AB**$|ab$, the transition probability to state **Ab**$|ab$, is a product of the probability that the bold haplotype (**AB**) is chosen (0.5), a recombination event occurs between the 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 **Ab**$|ab$ survives, ω_{14}.

$$\mathbf{P}_{1,j}^{t,t-1} = \begin{cases} .5\omega_{14}(1-r)x_4^{t-1} & \text{if } j = 1; \\ .5\omega_{14}(1-r)x_2^{t-1} & \text{if } j = 2; \\ .5\omega_{14}(1-r)x_3^{t-1} & \text{if } j = 3; \\ .5\omega_{14}(1-r)x_1^{t-1} & \text{if } j = 4; \\ .5\omega_{14}r\frac{w}{v+w}x_4^{t-1} & \text{if } j = 5; \\ .5\omega_{14}r\frac{w}{v+w}x_2^{t-1} & \text{if } j = 6; \\ .5\omega_{14}r\frac{w}{v+w}x_3^{t-1} & \text{if } j = 7; \\ .5\omega_{14}r\frac{w}{v+w}x_1^{t-1} & \text{if } j = 8 \\ .5\omega_{14}r\frac{v}{v+w}x_4^{t-1} & \text{if } j = 9; \\ .5\omega_{14}r\frac{v}{v+w}x_2^{t-1} & \text{if } j = 10; \\ .5\omega_{14}r\frac{v}{v+w}x_3^{t-1} & \text{if } j = 11; \\ .5\omega_{14}r\frac{v}{v+w}x_1^{t-1} & \text{if } j = 12 \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{1,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.7}$$

$$\mathbf{P}_{2,j}^{t,t-1} = \begin{cases} .5\omega_{12}(((1-r)x_4^{t-1}) + (r\frac{v}{v+w}x_4^{t-1})) & \text{if } j = 1; \\ .5\omega_{12}(((1-r)x_2^{t-1}) + (r\frac{v}{v+w}x_2^{t-1})) & \text{if } j = 2; \\ .5\omega_{12}(((1-r)x_3^{t-1}) + (r\frac{v}{v+w}x_3^{t-1})) & \text{if } j = 3; \\ .5\omega_{12}(((1-r)x_1^{t-1}) + (r\frac{v}{v+w}x_1^{t-1})) & \text{if } j = 4; \\ .5\omega_{12}r\frac{w}{v+w}x_4^{t-1} & \text{if } j = 5; \\ .5\omega_{12}r\frac{w}{v+w}x_2^{t-1} & \text{if } j = 6; \\ .5\omega_{12}r\frac{w}{v+w}x_3^{t-1} & \text{if } j = 7; \\ .5\omega_{12}r\frac{w}{v+w}x_1^{t-1} & \text{if } j = 8 \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12 \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{2,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.8}$$

$$\mathbf{P}_{3,j}^{t,t-1} = \begin{cases} .5\omega_{13}(((1-r)x_4^{t-1}) + (r\frac{w}{v+w}x_4^{t-1})) & \text{if } j = 1; \\ .5\omega_{13}(((1-r)x_2^{t-1}) + (r\frac{w}{v+w}x_2^{t-1})) & \text{if } j = 2; \\ .5\omega_{13}(((1-r)x_3^{t-1}) + (r\frac{w}{v+w}x_3^{t-1})) & \text{if } j = 3; \\ .5\omega_{13}(((1-r)x_1^{t-1}) + (r\frac{w}{v+w}x_1^{t-1})) & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ .5\omega_{13}r\frac{v}{v+w}x_4^{t-1} & \text{if } j = 9; \\ .5\omega_{13}r\frac{v}{v+w}x_2^{t-1} & \text{if } j = 10; \\ .5\omega_{13}r\frac{v}{v+w}x_3^{t-1} & \text{if } j = 11; \\ .5\omega_{13}r\frac{v}{v+w}x_1^{t-1} & \text{if } j = 12 \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{3,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.9}$$

$$\mathbf{P}_{4,j}^{t,t-1} = \begin{cases} .5\omega_{11}(((1-r)x_4^{t-1}) + (rx_4^{t-1})) & \text{if } j = 1; \\ .5\omega_{11}(((1-r)x_2^{t-1}) + (rx_2^{t-1})) & \text{if } j = 2; \\ .5\omega_{11}(((1-r)x_3^{t-1}) + (rx_3^{t-1})) & \text{if } j = 3; \\ .5\omega_{11}(((1-r)x_1^{t-1}) + (rx_1^{t-1})) & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{4,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.10}$$

$$\mathbf{P}_{5,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ .5\omega_{24}(((1-r)x_4^{t-1}) + (r\frac{w}{v+w}x_4^{t-1})) & \text{if } j = 5; \\ .5\omega_{24}(((1-r)x_2^{t-1}) + (r\frac{w}{v+w}x_2^{t-1})) & \text{if } j = 6; \\ .5\omega_{24}(((1-r)x_3^{t-1}) + (r\frac{w}{v+w}x_3^{t-1})) & \text{if } j = 7; \\ .5\omega_{24}(((1-r)x_1^{t-1}) + (r\frac{w}{v+w}x_1^{t-1})) & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ .5\omega_{24}r\frac{v}{v+w}x_4^{t-1} & \text{if } j = 13; \\ .5\omega_{24}r\frac{v}{v+w}x_2^{t-1} & \text{if } j = 14; \\ .5\omega_{24}r\frac{v}{v+w}x_3^{t-1} & \text{if } j = 15; \\ .5\omega_{24}r\frac{v}{v+w}x_1^{t-1} & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{5,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.11}$$

$$\mathbf{P}_{6,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ .5\omega_{22}(((1-r)x_4^{t-1}) + (rx_4^{t-1})) & \text{if } j = 5; \\ .5\omega_{22}(((1-r)x_2^{t-1}) + (rx_2^{t-1})) & \text{if } j = 6; \\ .5\omega_{22}(((1-r)x_3^{t-1}) + (rx_3^{t-1})) & \text{if } j = 7; \\ .5\omega_{22}(((1-r)x_1^{t-1}) + (rx_1^{t-1})) & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{6,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.12}$$

$$\mathbf{P}_{7,j}^{t,t-1} = \begin{cases} .5\omega_{23} r \frac{w}{v+w} x_4^{t-1} & \text{if } j = 1; \\ .5\omega_{23} r \frac{w}{v+w} x_2^{t-1} & \text{if } j = 2; \\ .5\omega_{23} r \frac{w}{v+w} x_3^{t-1} & \text{if } j = 3; \\ .5\omega_{23} r \frac{w}{v+w} x_1^{t-1} & \text{if } j = 4; \\ .5\omega_{23} (1-r) x_4^{t-1} & \text{if } j = 5; \\ .5\omega_{23} (1-r) x_2^{t-1} & \text{if } j = 6; \\ .5\omega_{23} (1-r) x_3^{t-1} & \text{if } j = 7; \\ .5\omega_{23} (1-r) x_1^{t-1} & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ .5\omega_{23} r \frac{v}{v+w} x_4^{t-1} & \text{if } j = 13; \\ .5\omega_{23} r \frac{v}{v+w} x_2^{t-1} & \text{if } j = 14; \\ .5\omega_{23} r \frac{v}{v+w} x_3^{t-1} & \text{if } j = 15; \\ .5\omega_{23} r \frac{v}{v+w} x_1^{t-1} & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{7,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.13}$$

$$\mathbf{P}_{8,j}^{t,t-1} = \begin{cases} .5\omega_{12} r \frac{w}{v+w} x_4^{t-1} & \text{if } j = 1; \\ .5\omega_{12} r \frac{w}{v+w} x_2^{t-1} & \text{if } j = 2; \\ .5\omega_{12} r \frac{w}{v+w} x_3^{t-1} & \text{if } j = 3; \\ .5\omega_{12} r \frac{w}{v+w} x_1^{t-1} & \text{if } j = 4; \\ .5\omega_{12}(((1-r)x_4^{t-1}) + (r\frac{v}{v+w} x_4^{t-1})) & \text{if } j = 5; \\ .5\omega_{12}(((1-r)x_2^{t-1}) + (r\frac{v}{v+w} x_2^{t-1})) & \text{if } j = 6; \\ .5\omega_{12}(((1-r)x_3^{t-1}) + (r\frac{v}{v+w} x_3^{t-1})) & \text{if } j = 7; \\ .5\omega_{12}(((1-r)x_1^{t-1}) + (r\frac{v}{v+w} x_1^{t-1})) & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{8,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.14}$$

$$\mathbf{P}_{9,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ .5\omega_{34}(((1-r)x_4^{t-1}) + (r\frac{v}{v+w}x_4^{t-1})) & \text{if } j = 9; \\ .5\omega_{34}(((1-r)x_2^{t-1}) + (r\frac{v}{v+w}x_2^{t-1})) & \text{if } j = 10; \\ .5\omega_{34}(((1-r)x_3^{t-1}) + (r\frac{v}{v+w}x_3^{t-1})) & \text{if } j = 11; \\ .5\omega_{34}(((1-r)x_1^{t-1}) + (r\frac{v}{v+w}x_1^{t-1})) & \text{if } j = 12; \\ .5\omega_{34}r\frac{w}{v+w}x_4^{t-1} & \text{if } j = 13; \\ .5\omega_{34}r\frac{w}{v+w}x_2^{t-1} & \text{if } j = 14; \\ .5\omega_{34}r\frac{w}{v+w}x_3^{t-1} & \text{if } j = 15; \\ .5\omega_{34}r\frac{w}{v+w}x_1^{t-1} & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{9,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.15}$$

$$\mathbf{P}_{10,j}^{t,t-1} = \begin{cases} .5\omega_{23} r \frac{v}{v+w} x_4^{t-1} & \text{if } j = 1; \\ .5\omega_{23} r \frac{v}{v+w} x_2^{t-1} & \text{if } j = 2; \\ .5\omega_{23} r \frac{v}{v+w} x_3^{t-1} & \text{if } j = 3; \\ .5\omega_{23} r \frac{v}{v+w} x_1^{t-1} & \text{if } j = 4 \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ .5\omega_{23}(1-r) x_4^{t-1} & \text{if } j = 9; \\ .5\omega_{23}(1-r) x_2^{t-1} & \text{if } j = 10; \\ .5\omega_{23}(1-r) x_3^{t-1} & \text{if } j = 11; \\ .5\omega_{23}(1-r) x_1^{t-1} & \text{if } j = 12; \\ .5\omega_{23} r \frac{w}{v+w} x_4^{t-1} & \text{if } j = 13; \\ .5\omega_{23} r \frac{w}{v+w} x_2^{t-1} & \text{if } j = 14; \\ .5\omega_{23} r \frac{w}{v+w} x_3^{t-1} & \text{if } j = 15; \\ .5\omega_{23} r \frac{w}{v+w} x_1^{t-1} & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{10,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.16}$$

$$\mathbf{P}_{11,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ .5\omega_{33}(((1-r)x_4^{t-1}) + (rx_4^{t-1})) & \text{if } j = 9; \\ .5\omega_{33}(((1-r)x_2^{t-1}) + (rx_2^{t-1})) & \text{if } j = 10; \\ .5\omega_{33}(((1-r)x_3^{t-1}) + (rx_3^{t-1})) & \text{if } j = 11; \\ .5\omega_{33}(((1-r)x_1^{t-1}) + (rx_1^{t-1})) & \text{if } j = 12; \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{11,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.17}$$

$$\mathbf{P}_{12,j}^{t,t-1} = \begin{cases} .5\omega_{13} r \frac{v}{v+w} x_4^{t-1} & \text{if } j = 1; \\ .5\omega_{13} r \frac{v}{v+w} x_2^{t-1} & \text{if } j = 2; \\ .5\omega_{13} r \frac{v}{v+w} x_3^{t-1} & \text{if } j = 3; \\ .5\omega_{13} r \frac{v}{v+w} x_1^{t-1} & \text{if } j = 4 \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ .5\omega_{13}(((1-r)x_4^{t-1}) + (r\frac{w}{v+w} x_4^{t-1})) & \text{if } j = 9; \\ .5\omega_{13}(((1-r)x_2^{t-1}) + (r\frac{w}{v+w} x_2^{t-1})) & \text{if } j = 10; \\ .5\omega_{13}(((1-r)x_3^{t-1}) + (r\frac{w}{v+w} x_3^{t-1})) & \text{if } j = 11; \\ .5\omega_{13}(((1-r)x_1^{t-1}) + (r\frac{w}{v+w} x_1^{t-1})) & \text{if } j = 12; \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16 \\ 1 - \sum_j^{16} \mathbf{P}_{12,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.18}$$

$$\mathbf{P}_{13,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ .5\omega_{44}(((1-r)x_4^{t-1}) + (rx_4^{t-1})) & \text{if } j = 13; \\ .5\omega_{44}(((1-r)x_2^{t-1}) + (rx_2^{t-1})) & \text{if } j = 14; \\ .5\omega_{44}(((1-r)x_3^{t-1}) + (rx_3^{t-1})) & \text{if } j = 15; \\ .5\omega_{44}(((1-r)x_1^{t-1}) + (rx_1^{t-1})) & \text{if } j = 16; \\ 1 - \sum_j^{16} \mathbf{P}_{13,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.19}$$

$$\mathbf{P}_{14,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ .5\omega_{24} r \frac{v}{v+w} x_4^{t-1} & \text{if } j = 5; \\ .5\omega_{24} r \frac{v}{v+w} x_2^{t-1} & \text{if } j = 6; \\ .5\omega_{24} r \frac{v}{v+w} x_3^{t-1} & \text{if } j = 7; \\ .5\omega_{24} r \frac{v}{v+w} x_1^{t-1} & \text{if } j = 8 \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ .5\omega_{24}(((1-r)x_4^{t-1}) + (r\frac{w}{v+w}x_4^{t-1})) & \text{if } j = 13; \\ .5\omega_{24}(((1-r)x_2^{t-1}) + (r\frac{w}{v+w}x_2^{t-1})) & \text{if } j = 14; \\ .5\omega_{24}(((1-r)x_3^{t-1}) + (r\frac{w}{v+w}x_3^{t-1})) & \text{if } j = 15; \\ .5\omega_{24}(((1-r)x_1^{t-1}) + (r\frac{w}{v+w}x_1^{t-1})) & \text{if } j = 16; \\ 1 - \sum_j^{16} \mathbf{P}_{14,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.20}$$

$$\mathbf{P}_{15,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ .5\omega_{34} r \frac{w}{v+w} x_4^{t-1} & \text{if } j = 9; \\ .5\omega_{34} r \frac{w}{v+w} x_2^{t-1} & \text{if } j = 10; \\ .5\omega_{34} r \frac{w}{v+w} x_3^{t-1} & \text{if } j = 11; \\ .5\omega_{34} r \frac{w}{v+w} x_1^{t-1} & \text{if } j = 12 \\ .5\omega_{34}(((1-r)x_4^{t-1}) + (r \frac{v}{v+w} x_4^{t-1})) & \text{if } j = 13; \\ .5\omega_{34}(((1-r)x_2^{t-1}) + (r \frac{v}{v+w} x_2^{t-1})) & \text{if } j = 14; \\ .5\omega_{34}(((1-r)x_3^{t-1}) + (r \frac{v}{v+w} x_3^{t-1})) & \text{if } j = 15; \\ .5\omega_{34}(((1-r)x_1^{t-1}) + (r \frac{v}{v+w} x_1^{t-1})) & \text{if } j = 16; \\ 1 - \sum_j^{16} \mathbf{P}_{15,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.21}$$

$$\mathbf{P}_{16,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ .5\omega_{44} r \frac{v}{v+w} x_4^{t-1} & \text{if } j = 9; \\ .5\omega_{44} r \frac{v}{v+w} x_2^{t-1} & \text{if } j = 10; \\ .5\omega_{44} r \frac{v}{v+w} x_3^{t-1} & \text{if } j = 11; \\ .5\omega_{44} r \frac{v}{v+w} x_1^{t-1} & \text{if } j = 12 \\ .5\omega_{44} r \frac{w}{v+w} x_4^{t-1} & \text{if } j = 9; \\ .5\omega_{44} r \frac{w}{v+w} x_2^{t-1} & \text{if } j = 10; \\ .5\omega_{44} r \frac{w}{v+w} x_3^{t-1} & \text{if } j = 11; \\ .5\omega_{44} r \frac{w}{v+w} x_1^{t-1} & \text{if } j = 12 \\ .5\omega_{44}(1-r) x_4^{t-1} & \text{if } j = 13; \\ .5\omega_{44}(1-r) x_2^{t-1} & \text{if } j = 14; \\ .5\omega_{44}(1-r) x_3^{t-1} & \text{if } j = 15; \\ .5\omega_{44}(1-r) x_1^{t-1} & \text{if } j = 16; \\ 1 - \sum_j^{16} \mathbf{P}_{16,j}^{t,t-1} & \text{if } j = 17 \end{cases} \tag{4.22}$$

$$\mathbf{P}_{17,j}^{t,t-1} = \begin{cases} 0 & \text{if } j = 1; \\ 0 & \text{if } j = 2; \\ 0 & \text{if } j = 3; \\ 0 & \text{if } j = 4; \\ 0 & \text{if } j = 5; \\ 0 & \text{if } j = 6; \\ 0 & \text{if } j = 7; \\ 0 & \text{if } j = 8; \\ 0 & \text{if } j = 9; \\ 0 & \text{if } j = 10; \\ 0 & \text{if } j = 11; \\ 0 & \text{if } j = 12; \\ 0 & \text{if } j = 13; \\ 0 & \text{if } j = 14; \\ 0 & \text{if } j = 15; \\ 0 & \text{if } j = 16; \\ 1 & \text{if } j = 17 \end{cases} \tag{4.23}$$

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

$$\pi_v^\tau = \pi_0^\tau \prod_{t=0}^{\tau} \mathbf{P}_v^{t,t-1}. \tag{4.24}$$

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

$$\Psi_v^{\tau} = 1 - \pi_v^{\tau}\eta. \tag{4.25}$$

The transition matrix Q_v can now be computed using Equation 4.6 for all values of v where $v \in \{1...L\}$. In contrast to the transition matrix Q defined in Equation 4.3, 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, $\{b_n\}_{n=1,...,\Lambda}$, can be computed and used with Equations 4.4 and 4.5 to calculate the tract length distribution.

4.4 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].

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 4.6, the matrix multiplication required both for Equation 4.24 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 becomes 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].

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 Bret Payseur and Megan Frayer 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 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.

CHAPTER 5

DO HELICONIUS BUTTERFLY SPECIES EXCHANGE MIMICRY ALLELES?

Joel Smith and Marcus Kronforst

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

5.1 Abstract

Hybridization has the potential to transfer beneficial alleles across species boundaries and there are a growing number of examples in which this has apparently occurred. Recent studies suggest that *Heliconius* butterflies have transferred wing pattern mimicry alleles between species via hybridization, but ancestral polymorphism could also produce a signature of shared ancestry around mimicry genes. To distinguish between these alternative hypotheses, we measured DNA sequence divergence around putatively introgressed mimicry loci and compared this to the rest of the genome. Our results reveal that putatively introgressed regions show strongly reduced sequence divergence between co-mimetic species, suggesting that their divergence times are younger than the rest of the genome. This is consistent with introgression and not ancestral variation. We further show that this signature of introgression occurs at sites throughout the genome, not just around mimicry genes.

5.2 Introduction

Genetic exchange between species is gaining ground as a potentially important source of variation for adaptation and speciation [Abbott et al., 2013, Arnold et al., 2012, Seehausen, 2004]. A recent paper by the *Heliconius* Genome Consortium demonstrated that *Heliconius* butterfly species with similar mimetic wing patterns exhibit shared ancestry around wing patterning loci, suggesting that hybridization and introgression have transferred mimicry alleles from one species to another [Dasmahapatra et al., 2012]. The results of a companion study further support this conclusion [Pardo-Diaz et al., 2012]. However, there are two entirely distinct phenomena that could produce a signal of shared ancestry around mimicry loci; introgression (Figure 5.1a) and shared ancestral polymorphism (Figure 5.1b). While trans-species polymorphisms due to shared ancestral variation are widespread in nature, this alternative has not been carefully considered in the case of *Heliconius* mimicry [Klein et al., 1998].

The D-statistic used by the HGC to infer introgression is capable of distinguishing biased allele sharing from random sorting of ancestral polymorphism [Durand et al., 2011]. This test statistic uses nucleotide sites that fall into two categories: "ABBA" and "BABA", where each letter refers to an allele in each of four taxa. "A" refers to the outgroup, or ancestral allele, and "B" refers to the derived allele (Supplementary Material). ABBA and BABA configurations should occur with equal frequency in cases where there has been no admixture between taxa and ancestral populations mate randomly. When tallied across the genome, skew toward ABBA or BABA sites would indicate a systematic bias in allele sharing between taxa. In cases where allele

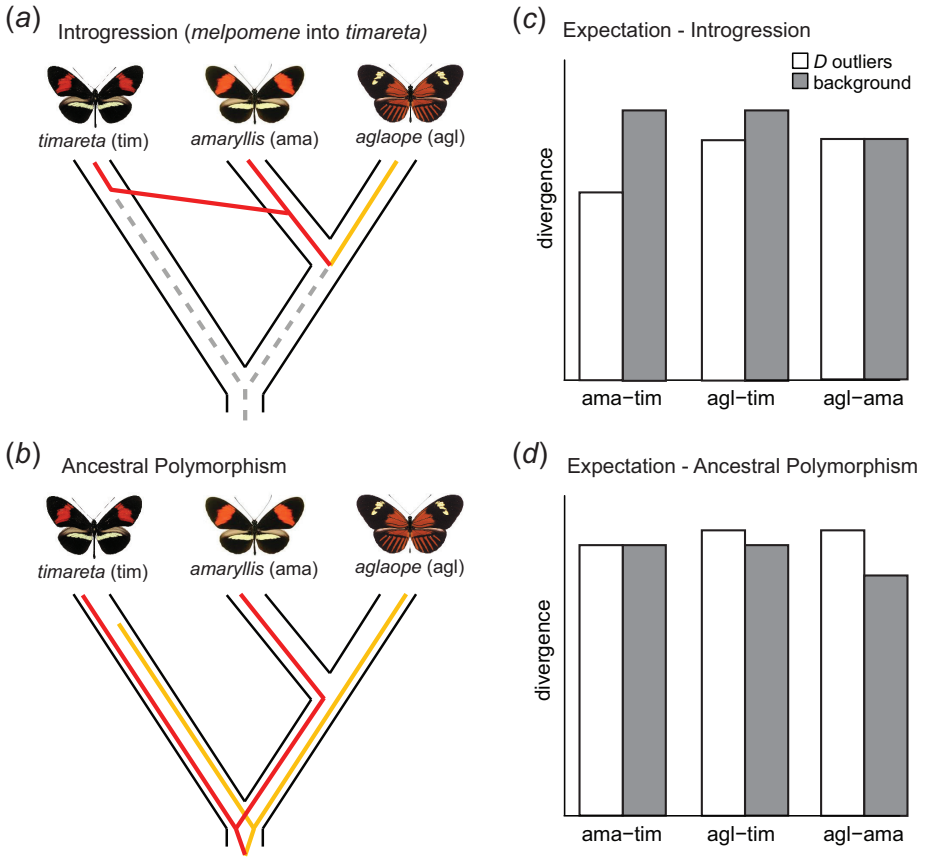

Figure 5.1: Sequence divergence distinguishes between ancestral polymorphism and introgression as the source of shared haplotypes between species. (a) Introgression predicts young divergence times between shared haplotypes resulting in (c) reduced sequence divergence between species, as compared to the genomic background. (b) Ancestral polymorphism predicts older divergence times resulting in (d) greater sequence divergence between species.

sharing is enriched between sympatric taxa, introgression is a potential explanation.

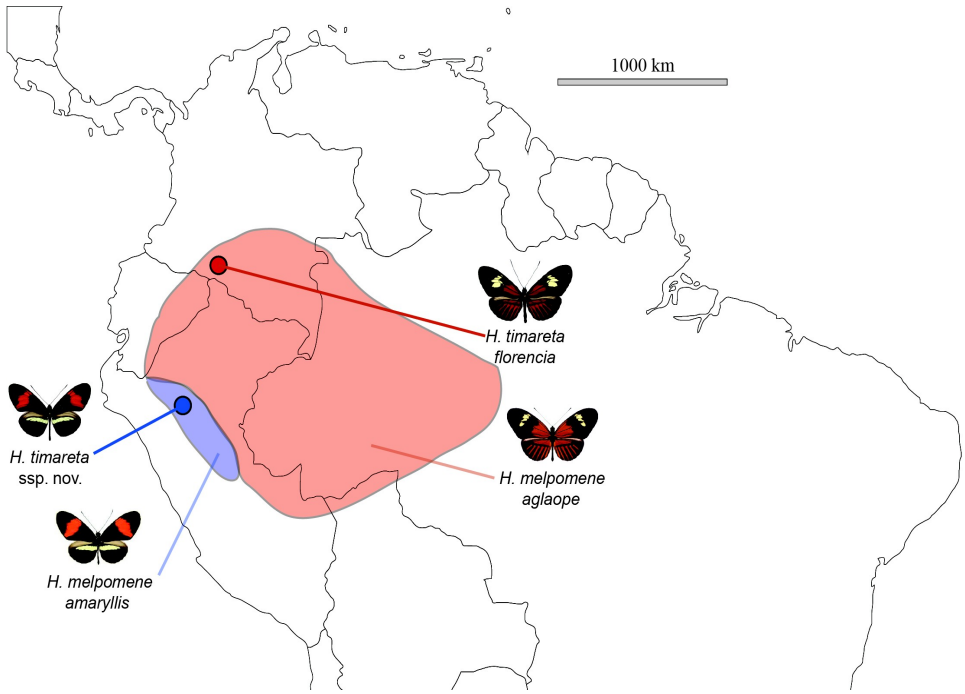

Figure 5.2: Map of South America showing distributions of *H. melpomene* and *H. timareta* used in this study.

However, ancestral polymorphism could interact with selection to give the false appearance of introgression in a mimicry system. If an ancestral species were polymorphic for two mimetic wing patterns, which it then passed to each of two descendant species, subsequent directional selection for local mimicry could result in co-occurring sister species with shared mimicry phenotypes controlled by homologous haplotypes. As an example, polymorphism in the ancestor could be the product of co-occurring mimicry models or co-mimics, a phenomenon well-known to promote in-

traspecific mimetic polymorphism [Joron and Mallet, 1998, Mallet and Joron, 1999]. Subsequent divergent selection could then simply be a product of shifting ranges for one or both of the models/co-mimics. This scenario is analogous to well-studied molecular trans-species polymorphisms such as MHC alleles and the ABO blood group, except that variation is locally monomorphic today as a result of selection for mimicry [Klein et al., 1998, Ségurel et al., 2012].

This alternative explanation could account for a variety of HGC results, such as localized elevation of the D-statistic around mimicry loci and gene trees that group taxa by mimicry phenotype as opposed to species (but it is unlikely to explain the genome-wide elevation in the D-statistic noted between sympatric species) [Dasmahapatra et al., 2012]. The fact that ancestral polymorphism has not been considered as a potential explanation for shared variation around mimicry genes has been controversial, with subsequent published work and blog discussions [Brower [2013], http://gcbias.org/2012/05/23/journal-tea-may-21st/] calling for clarification. *Heliconius* butterflies tend to be locally monomorphic so it is difficult to imagine how a wing pattern polymorphism could be maintained through a speciation event. However, there are multiple examples of local polymorphism in *Heliconius*, including one of the focal species here, *H. timareta*, which has as many as four co-existing phenotypes in Ecuador [Chamberlain et al., 2009, Joron et al., 2011, Mallet, 1999]. Here we specifically test the hypothesis that shared mimicry phenotypes are due to ancestral variation, as opposed to introgression, by examining DNA sequence divergence in putatively introgressed regions of the genome.

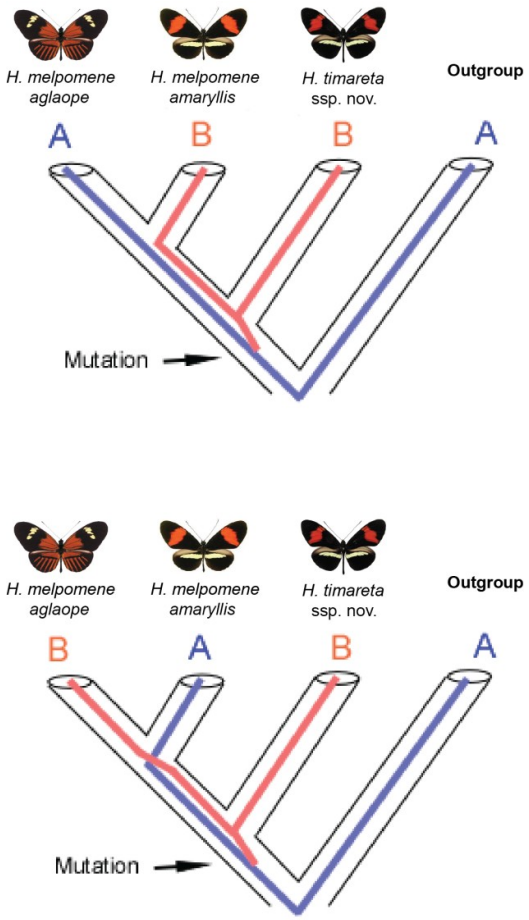

Figure 5.3: A Schematic of taxa distributed on the phylogeny for calculating the D-statistic. Note that an enrichment of ABBA sites, resulting in a positive D value, is indicative of biased allele sharing between sympatric *H. melpomene amaryllis* and *H. timareta ssp. nov.*

5.3 Materials and Methods

We reanalyzed the HGC data, focusing on DNA sequence divergence between species. Our approach considers the impact of introgression vs. ancestral polymorphism on sequence divergence between shared alleles, as compared to the genomic background.

Introgression should result in recent splitting of alleles, and low sequence divergence, compared to the genomic background (Figure 5.1c). In contrast, ancestral polymorphism should result in more ancient splitting of alleles and greater sequence divergence (Figure 5.1d). Note that a third potential explanation, convergent molecular evolution, was excluded by HGC analyses but would result in mimicry allele divergence times that match the genomic background [Dasmahapatra et al., 2012].

The HGC data consist of targeted resequencing data around two mimicry loci (B/D and N/Yb) for two *H. melpomene*/*H. timareta* comparisons, co-mimetic *H. m. amaryllis* and *H. t. ssp. nov.* in Peru and *H. m. aglaope* and *H. t. florencia* in Peru and Colombia (Figure 5.2). There are also targeted resequencing data for outgroup taxa as well as genome-wide RAD data for all taxa except *H. timareta florencia*. For our analyses, we focused on sympatric *H. m. amaryllis* and *H. timareta* in Peru, using *H. m. aglaope* as an allopatric comparison, because these are the samples for which both resequencing and RAD data exist. Each ingroup taxon consists of four individuals in the resequencing datasets and five individuals in the RAD dataset. The silvaniform outgroup includes one individual from each of the following taxa: *H. hecale*, *H. numata silvana*, *H. ethilla*, *H. pardalinus sergestus*, and *H. pardalinus ssp. nov.* Because the D-statistic depends on identifying derived alleles that are shared between taxa, the inclusion of multiple taxa in the outgroup provides additional confidence in identifying ancestral alleles.

To estimate allele frequencies, we used biallelic sites having a GATK quality score greater than 30 (99.9% accuracy) [McKenna et al., 2010]. In accordance with the HGC analysis, only sites with at least 50% of the genotypes available for each

of the four groups were used. Alleles were polarized with respect to the outgroup major allele and sites with an outgroup frequency of 50% were excluded. After filtering, the B/D, N/Yb and RAD datasets contained 55847, 79549 and 142869 SNPs, respectively.

Using fixed sites among four taxa (P_1, P_2, P_3, and outgroup P_4) at a locus with n "ABBA" and "BABA" sites in total, Patterson's D-statistic is calculated as:

$$D(P_1, P_2, P_3, P_4) = \frac{\sum_{i=1}^{n} C_{ABBA}(i) - C_{BABA}(i)}{\sum_{i=1}^{n} C_{ABBA}(i) + C_{BABA}(i)} \tag{5.1}$$

where $C_{ABBA}(i)$ and $C_{BABA}(i)$ take on values of 1 or 0 for each site compatible with an "ABBA" or "BABA" configuration, respectively. The D-statistic can be extended to include polymorphic sites using frequency estimates ($\hat{p}_{i1}$, $\hat{p}_{i2}$, $\hat{p}_{i3}$, $\hat{p}_{i4}$) at each site, i, for a locus with n SNPs. We followed the HGC in using this approach:

$$D(P_1, P_2, P_3, P_4) = \frac{\sum_{i=1}^{n} \hat{p}_{i3}(1-\hat{p}_{i4})((1-\hat{p}_{i1})\hat{p}_{i2} - \hat{p}_{i1}(1-\hat{p}_{i2}))}{\sum_{i=1}^{n} \hat{p}_{i3}(1-\hat{p}_{i4})((1-\hat{p}_{i1})\hat{p}_{i2} + \hat{p}_{i1}(1-\hat{p}_{i2}))}. \tag{5.2}$$

We calculated Patterson's D-statistic in non-overlapping 5 kbp windows across the B/D (717 kbp) and the N/Yb (1.15 Mbp) scaffolds from the *H. melpomene* reference genome. We used the following four group comparison to calculate D: *H. melpomene aglaope*, *H. melpomene amaryllis*, *H. timareta ssp. nov.*, silvaniform outgroup (Figure 5.3). Positive D outlier windows are indicative of allele sharing between sympatric *H. m. amaryllis* and *H. timareta* in Peru. We used a similar approach for the rest of the genome, calculating D and divergence based on RAD data. For the genome-wide

analysis we excluded the B/D and N/Yb regions.

Finally, we compared mean DNA sequence divergence (d_{xy}) between the top 10% D outliers (positive values) and the remaining windows in each region. We estimated the mean pairwise sequence divergence between taxa for a window of length n basepairs as:

$$d_{xy} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{ix}(1 - \hat{p}_{iy}) + \hat{p}_{iy}(1 - \hat{p}_{ix}) \tag{5.3}$$

where $\hat{p}_x$ and $\hat{p}_y$ refer to the reference allele frequency in taxon x and y, respectively. Note that for RAD data, sequenced loci were clustered into windows based on variants being within 5 kbp of each other and divergence was calculated by dividing by the total basepairs in the window. For this reason, divergence estimates in Figure 2c are artificially low but internally consistent. We used Welch's t-test to compare divergence between outlier and background intervals.

5.4 Results

When we considered the putatively introgressed regions around two wing patterning loci, the B/D locus (which controls red patterning) and the N/Yb locus (which controls yellow patterning), we found distinct signals of reduced divergence consistent with introgression (Figure 5.4a and 5.4b). In addition, results from the HGC suggested that introgression may also occur in parts of the genome other than these two wing patterning loci [Dasmahapatra et al., 2012]. We scanned the genome and identified widespread signatures of elevated D-statistic consistent with genome-wide

introgression between *H. m. amaryllis* and *H. timareta* in Peru. When we examined sequence divergence in these additional regions, excluding the B/D and N/Yb loci, we again found reduced divergence indicative of introgression (Figure 5.4c).

The results reveal an additional pattern consistent with introgression. The HGC analyses suggest that mimicry introgression may have also occurred between *H. m. aglaope* and *H. timareta* in Colombia [Dasmahapatra et al., 2012]. Furthermore, directionality of introgression is suspected to be from *H. melpomene* into *H. timareta* [Pardo-Diaz et al., 2012]. If so, then our comparison between *H. m. aglaope* and Peruvian *timareta*, in putatively introgressed regions of the genome (middle white bars in Figure 5.4), should effectively be a comparison between *aglaope* and *amaryllis*. The fact that this level of divergence is similar to *H. m. aglaope* and *H. m. amaryllis* background divergence (right-most bars in Figure 5.4) lends additional support to the introgression hypothesis.

5.5 Discussion

Hybridization is widespread among *Heliconius* species and previous work suggests that this may result in interspecific gene flow [Brower, 2011, Bull et al., 2006, Kronforst, 2008, Kronforst et al., 2006, Mallet et al., 2007]. This, combined with examples of mimicry between species known to hybridize, makes the mimicry introgression hypothesis appealing [Brower, 1996, Gilbert, 2003, Giraldo et al., 2008]. However, the alternative hypothesis of trans-species polymorphism due to shared ancestral variation must also be considered. This alternative explanation is relevant when it comes to *Heliconius* wing patterns because *H. melpomene* and *H. timareta* are closely re-

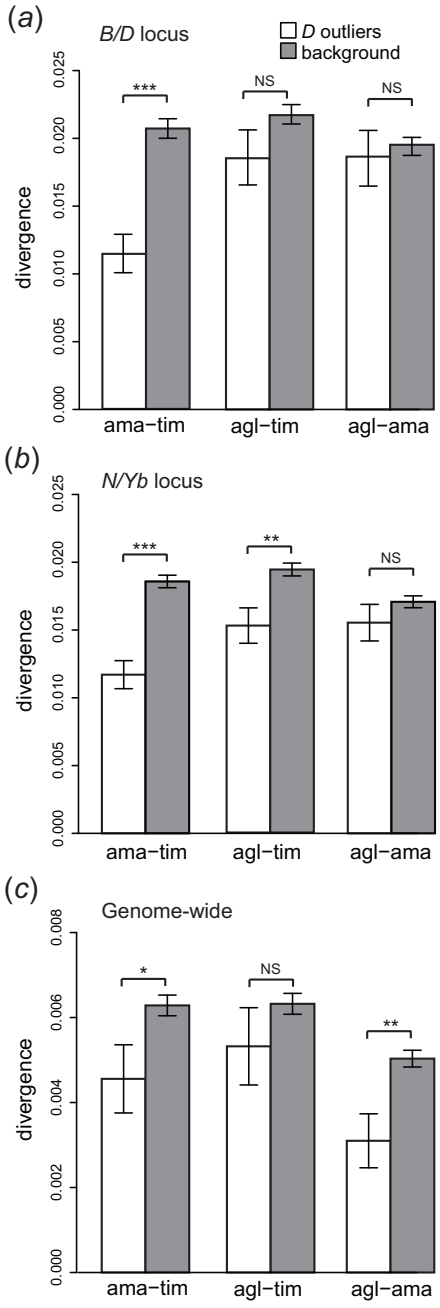

Figure 5.4: Sequence divergence is reduced between co-mimetic *H. melpomene amaryllis* and *H. timareta* at (a) the B/D mimicry locus, (b) the N/Yb mimicry locus, and (c) genome-wide regions of elevated Patterson's D-statistic. $***P < 0.001$, ** $P < 0.01$, * $P < 0.05$, NS $P > 0.05$.

lated, having diverged 1-1.5 mya, and both species are polymorphic for the same wing pattern variation, suggesting their common ancestor may have been so as well [Kronforst et al., 2006]. Despite these features that make ancestral polymorphism plausible, our results ultimately support the hypothesis that introgression has moved mimicry alleles between *Heliconius* species [Dasmahapatra et al., 2012, Gilbert, 2003, Pardo-Diaz et al., 2012]. By comparing two alternative hypotheses we are able to substantiate this previously tentative conclusion. Furthermore, given the widespread signatures of introgression found across the genome, it is quite likely that interspecific gene flow has impacted aspects of *Heliconius* biology beyond wing patterning.

Acknowledgements

We thank Kanchon Dasmahapatra and Jim Mallet for sharing data and for their comments on a previous version of the manuscript. We also thank Wei Zhang and Grace Lee for assistance with the analysis and reviewers for comments on previous versions of the manuscript. This research was supported by NSF grant DEB-1316037 to MRK.

APPENDIX

A.1 Initializing the Ancestral Haplotype for the MCMC

To decrease run times for the MCMC, we initialize the starting sequence for the ancestral haplotype using a heuristic algorithm which exploits the decrease in polymorphism near the selected site. Let A^0 denote the initial ancestral haplotype to be estimated, and let the indicator variable I_{ij} denote whether chromosome i is part of the ancestral haplotype at site j:

$$I_{ij} = \begin{cases} 1 & \text{if } X_{ij} = A^0_j; \\ 0 & \text{if } X_{ij} \neq A^0_j \end{cases} \tag{4}$$

The algorithm proceeds as follows:

1. At $j = 1$ all chromosomes with the beneficial allele are specified to be on the ancestral haplotype at the selected site, i.e. $\sum_{i=1}^{n} I_{i1} = n$ and $A^0_j = 1$.

2. Moving to the next adjacent SNP, we calculate the allele frequency, F_j, among chromosomes on the ancestral haplotype at the previous site:

$$F_j = \frac{\sum_{i=1}^{n} X_{ij} I_{i(j-1)}}{\sum_{i=1}^{n} I_{i(j-1)}} \tag{5}$$

3. The major allele among advantageous allele carriers is assumed to be the putative ancestral allele and minor alleles are assumed to be the result of a putative recombination event off of the ancestral haplotype in the previous SNP interval. For $j > 0$,

$$A_j^0 = \begin{cases} 1 & \text{if } F_j > 0.5; \\ 0 & \text{if } F_j < 0.5 \end{cases} \tag{6}$$

Because we expect there to be some rare or singleton variants on the ancestral haplotype, singletons are removed before step 1 in an effort to improve estimates of the ancestral haplotype at more distant sites. In addition, major and minor alleles can't be identified at sites with alleles at 0.5 frequency and are also removed initially. Steps 2 and 3 are computed iteratively until reaching the end of the locus ($j = L$) on both sides flanking the selected site. The sites that were removed ($F_j = 0.5$ and singletons) are then added back in and take values of I_{ij} from I_{ij+1}. A_j^0 for the added sites are computed using equations 12 and 13. At sites for which $\sum_{i=1}^{n} I_{i1} = 0$, $A_j^0 = \text{Binomial}(1, P_j)$, where $P_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$. After getting the initial estimate A_j^0, the MCMC is run and evaluated for convergence by visual inspection of trace plots.

A.2 Modeling Singletons and Invariant Sites on Background Haplotypes

We implemented an approach for modeling invariant sites on the background haplotypes among carriers of the beneficial allele, the accuracy of which are summarized in Table S.8 (Appendix A.4). This portion of the likelihood is denoted β_{iw} in the model description (Equation 2.5), and corresponds to the Li and Stephens [2003] haplotype copying model. The original formulation of this model ignores invariant

sites, which is equivalent to assuming the probability of observing them is 1. We modified our likelihood by first noting that, under the star genealogy assumption, all variants which are found in the carrier panel but absent in the reference panel should be singletons. Thus we can estimate the probability of a invariant site by considering the rate of singletons in the reference panel. Specifically, we estimate the probability as $1 - (S/nL)$, where S is the number of singletons in the reference panel, n is the number of reference panel haplotypes and L is the number of basepairs at the locus. For a given haplotype i and SNP w in β_{iw}, the probability of observing d invariant sites is $(1 - (S/nL))^d$ (Model B in Table S.8, Appendix A.4).

Despite this attempt to more accurately model the background haplotypes in the carrier panel, we did not find any consistent improvement in bias or accuracy.

A.3 Chapter 2 Supplementary Figures

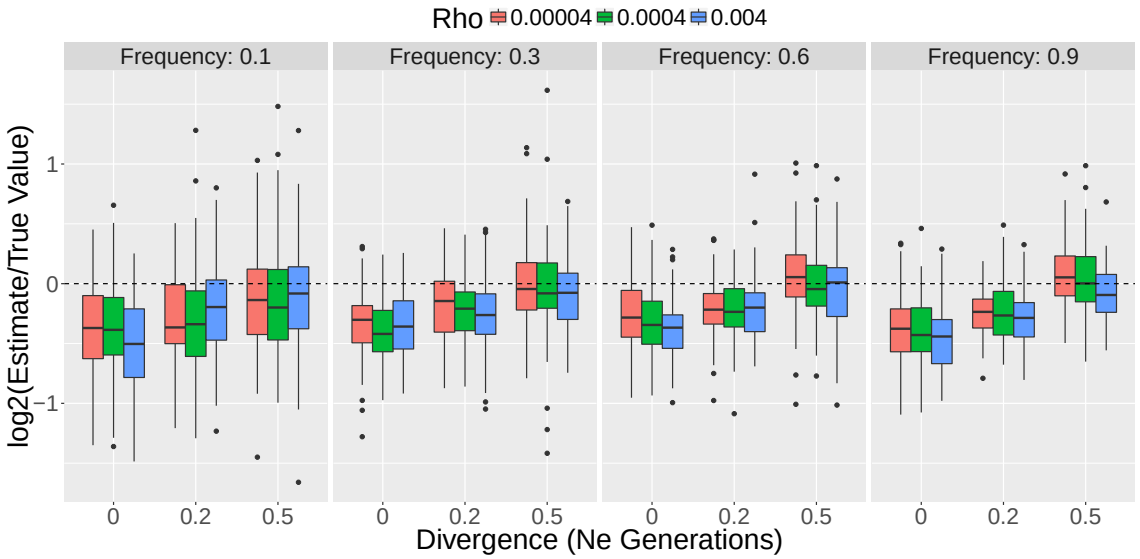

Figure S.1: Effect of mis-specifying rho. Accuracy results for 3 different values of rho used in the Li and Stephens [2003] copying model for background haplotypes. All other parameter values are identical to Figure 2.2. The divergence value of 0 refers to a local reference panel. Allele frequency refers to the end frequency of the beneficial allele trajectory.

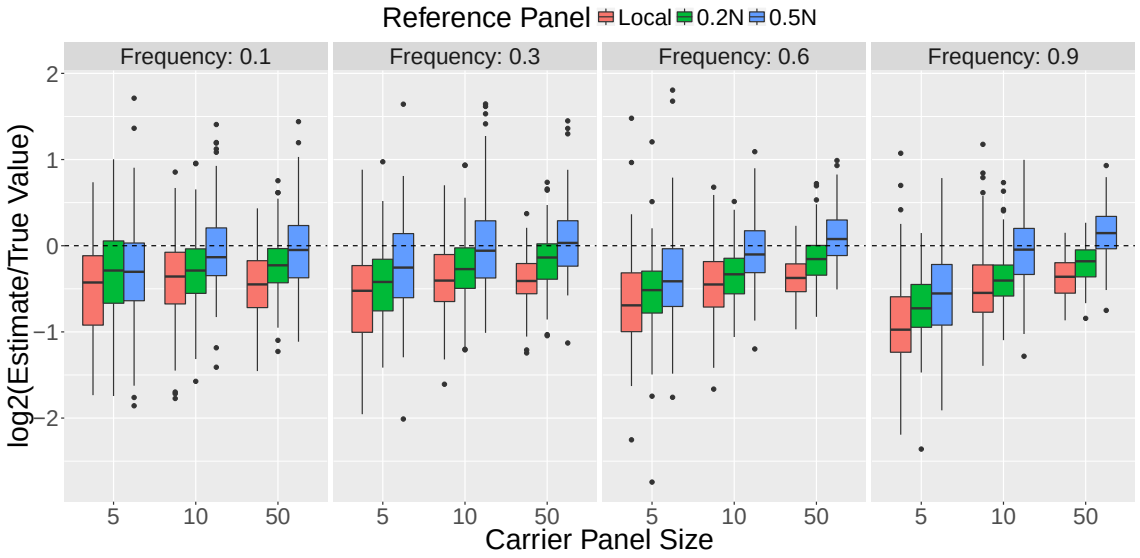

Figure S.2: Effect of beneficial allele carrier sample size. Accuracy results for 3 different sample sizes for the panel of haplotypes carrying the beneficial allele. The selection strength for all simulations was set to 0.01. All other parameter values are identical to Figure 2.2. Allele frequency refers to the end frequency of the beneficial allele trajectory.

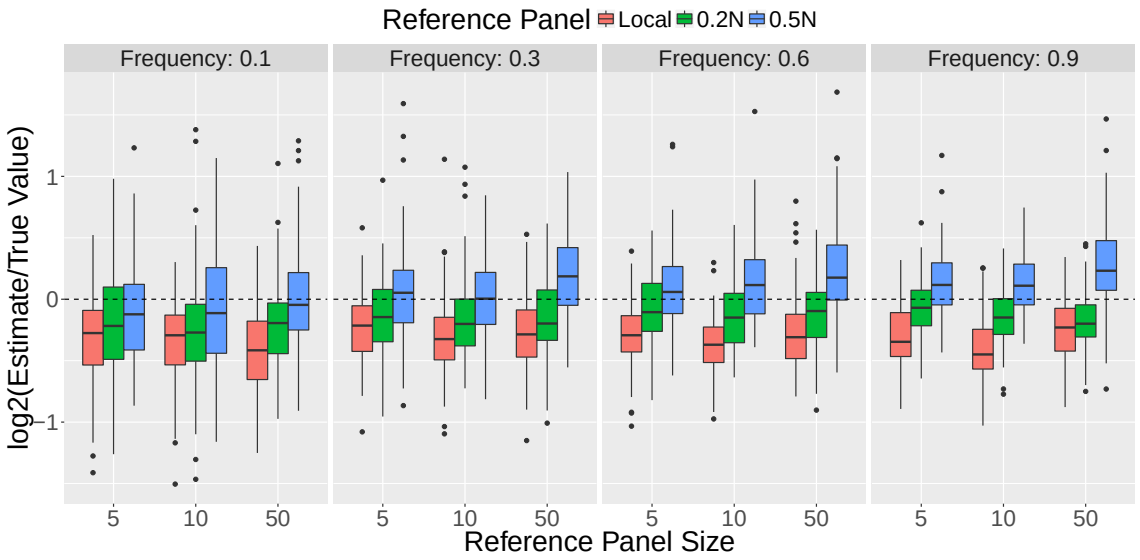

Figure S.3: Effect of reference panel sample size. Accuracy results for 3 different sample sizes for the reference panel of haplotypes without the selected allele. The selection strength for all simulations was set to 0.01. All other parameter values are identical to Figure 2.2. Allele frequency refers to the end frequency of the beneficial allele trajectory.

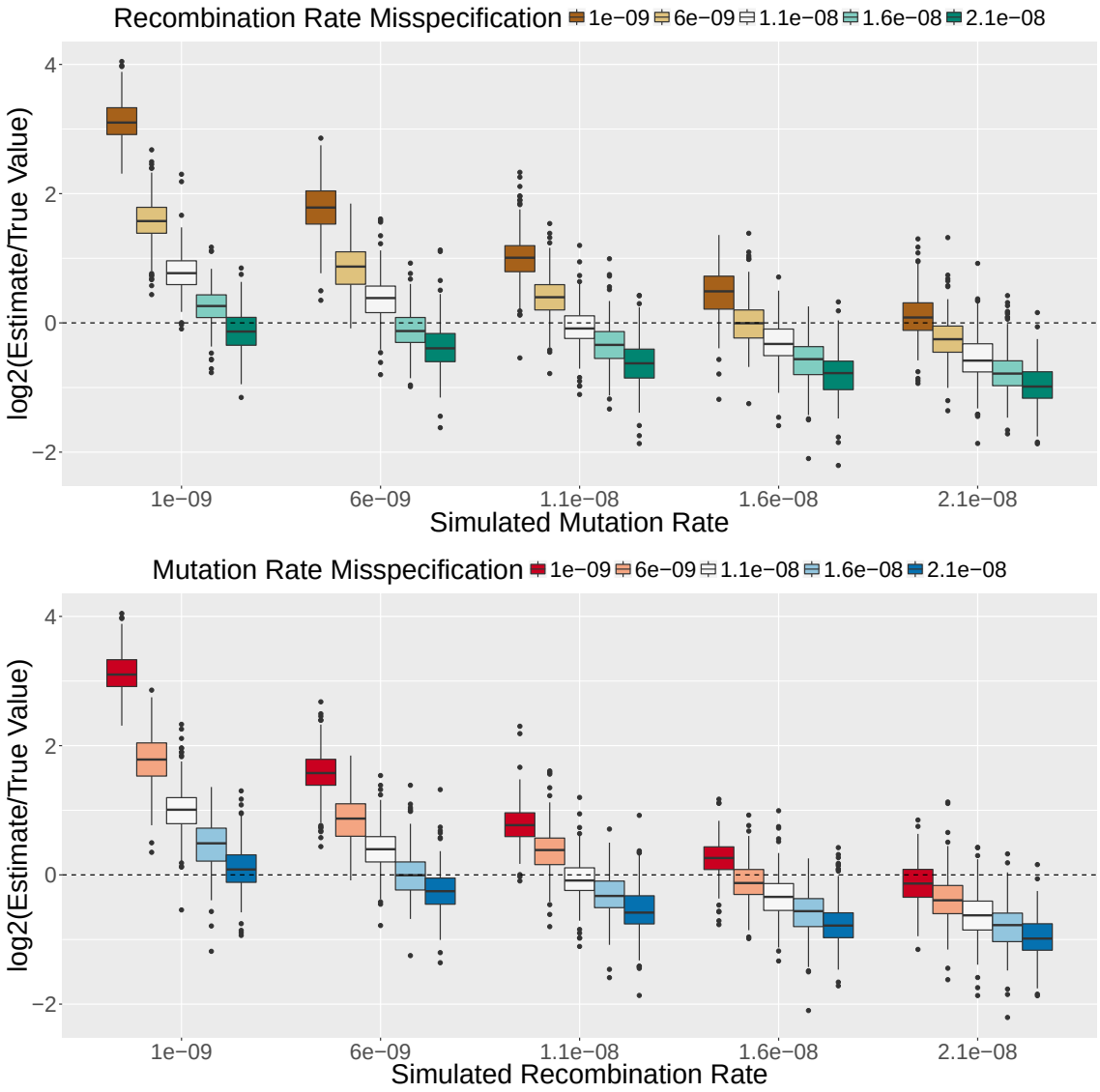

Figure S.4: Effect of mis-specifying the mutation and recombination rates. Accuracy results for varying degrees of mutation and recombination rate misspecification. In both panels, the parameter values on the x-axis were used both for simulation and inference. For the colored boxplots, the true values are in white (1.1×10^{-8}) and the colors refer to different degrees of misspecification used for inference. Simulations were performed with a local reference panel and a selection strength of 0.01. All other parameter values are identical to Figure 2.2.

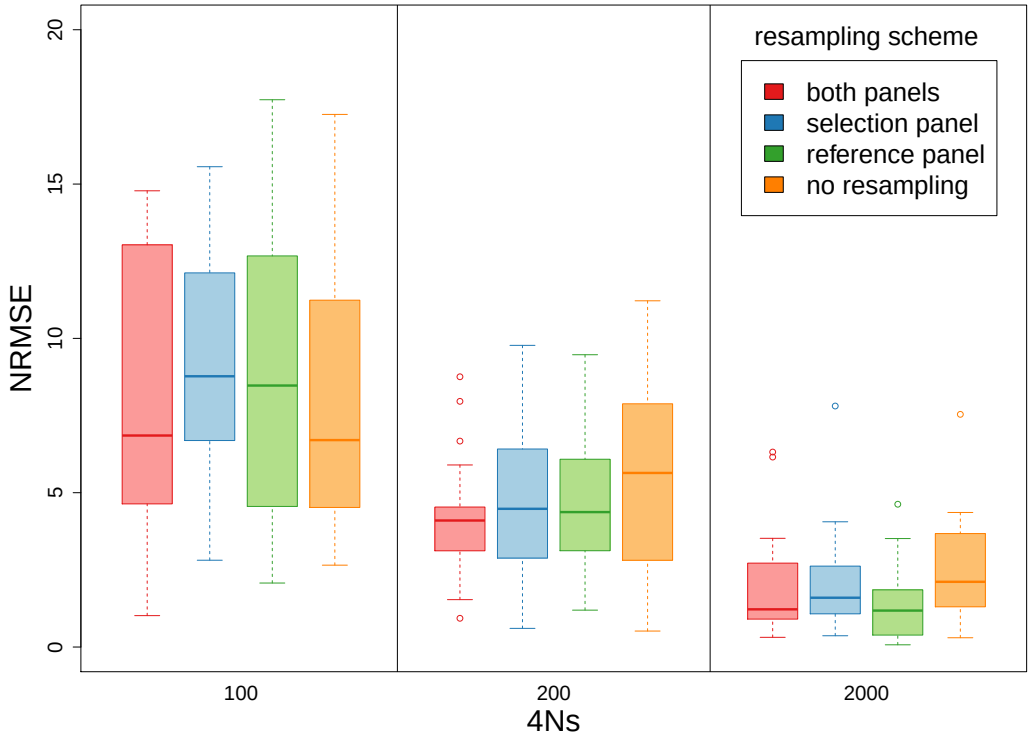

Figure S.5: Effect of resampling subsets of complete data. Estimated accuracy and among independent MCMC runs for different resampling schemes. Frequency trajectories were simulated to an end frequency of 0.1. Under each 2Ns value and resampling scheme indicated in the legend, 20 data sets were simulated and inference was performed on the 5 replicate MCMCs. In each simulation, the full dataset includes sample sizes of 100 for the selected and reference panels. Inference for each replicate was then performed on 50 selected haplotypes and 20 reference haplotypes according to the sampling scheme in the legend. Normalized RMSE values are calculated using the estimates and true TMRCA value, while the standard deviations are calculated using the estimates and their mean.

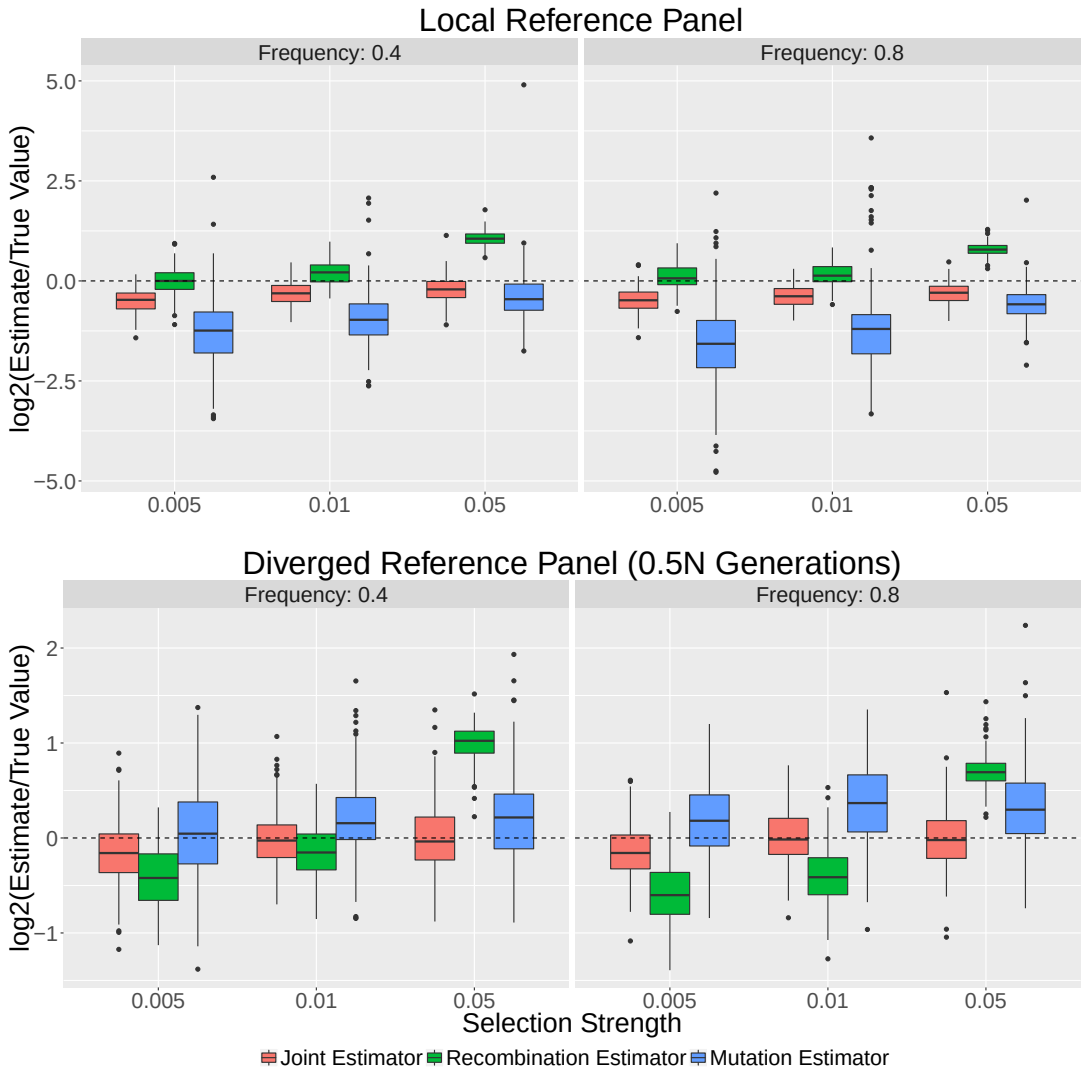

Figure S.6: Comparison to heuristic estimates. We compared our TMRCA estimator (joint estimator) to an estimate which uses the mean length of haplotype lengths and another estimate which uses number of derived mutations on the ancestral haplotype. In all simulations a selection strength of 0.01 was used. All other parameter values are identical to Figure 2.2. Frequency refers to the end frequency of the beneficial allele trajectory.

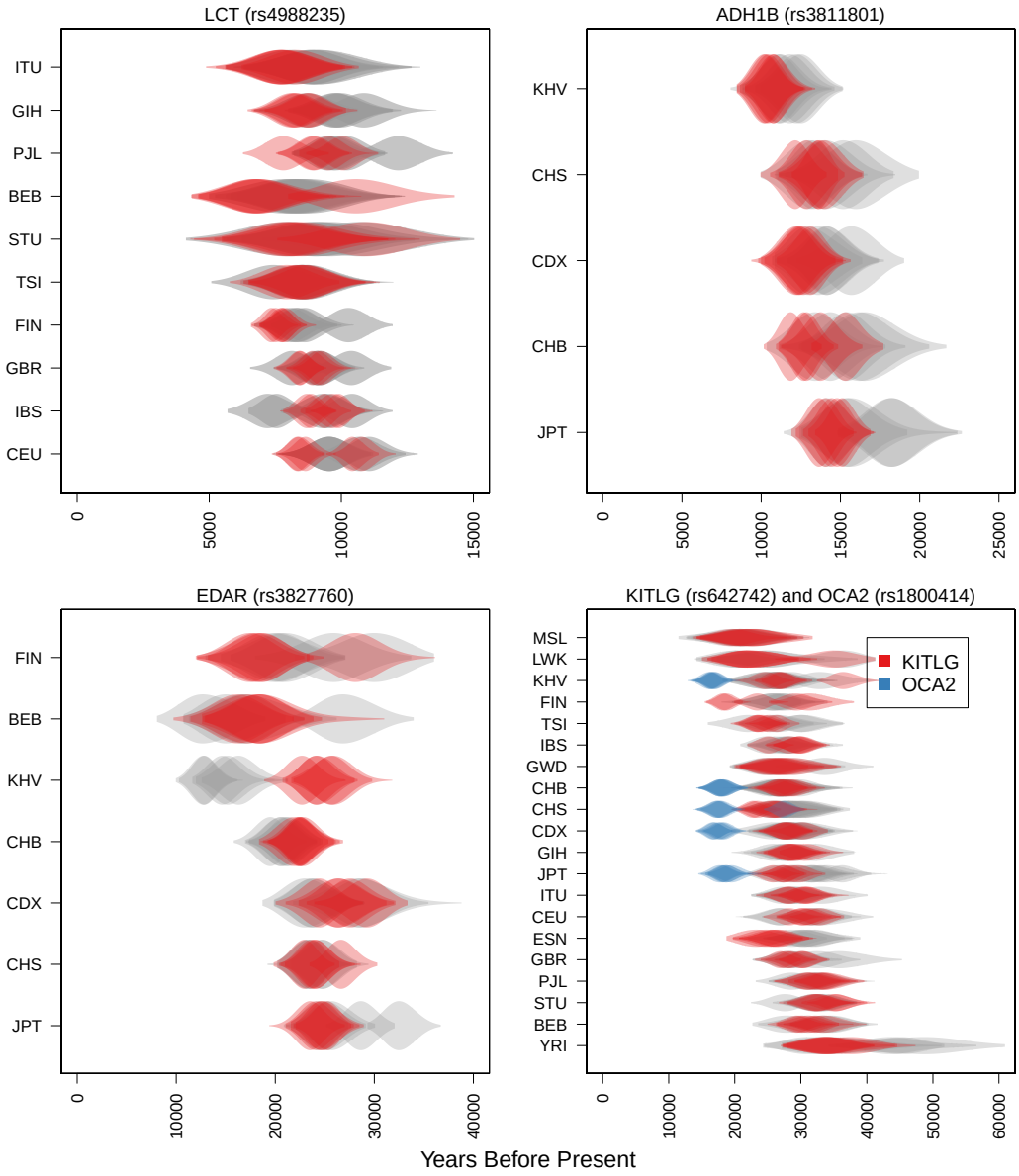

Figure S.7: Comparison of fine-scale and Mbp-scale recombination maps. A comparison between estimates made using the fine-scale Decode recombination map (gray) and a uniform recombination rate (red and blue). The uniform recombination rate used for each gene is the mean rate for the 1Mb region around each variant indicated by the rs number. Five replicate MCMCs were performed for each variant and population by resampling the selected and reference panels with replacement.

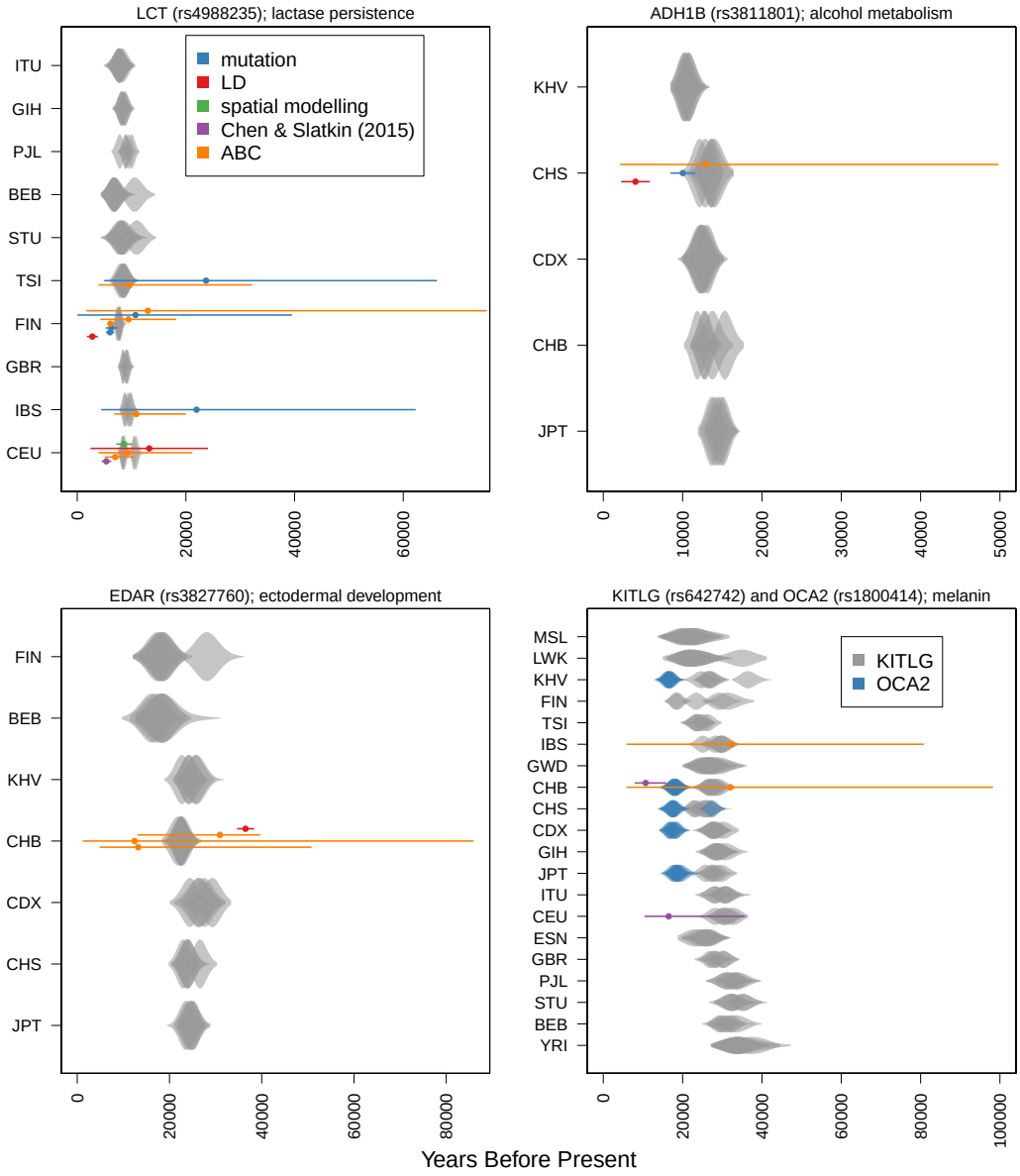

Figure S.8: Comparison of TMRCA estimates and previous estimate approaches. Results from Figure 2.2 sorted into different plots for different variants. Previous estimates are colored by an abbreviated description of the type of information used in the data. The blue violin plots in the KITLG/OCA2 plot are estimates for the OCA2 variant. The purple and orange previous estimates for CHB in the KITLG/OCA2 plot refer to OCA2 and KITLG, respectively.

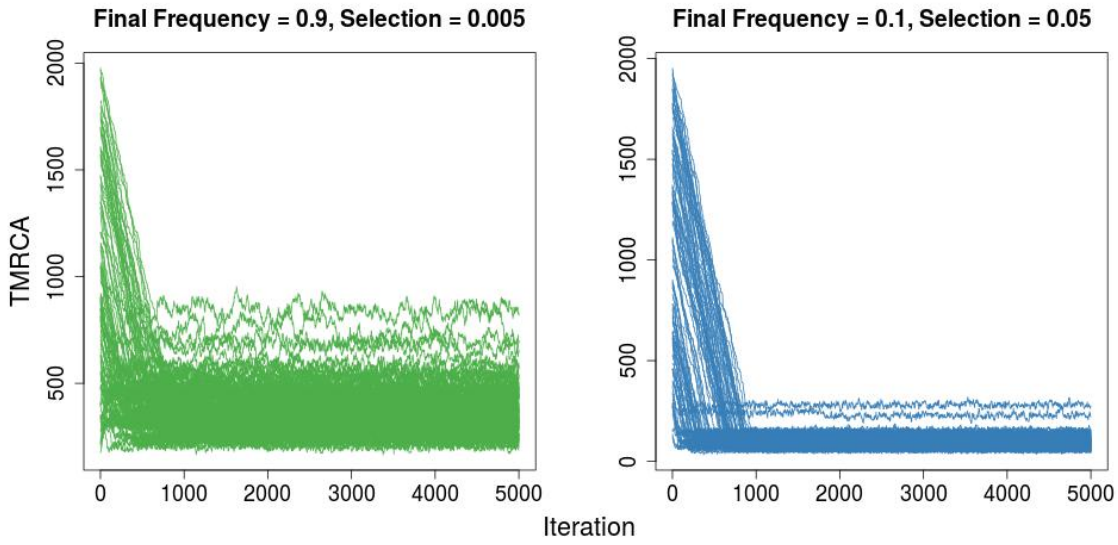

Figure S.9: Traces of MCMC results from simulated data. Results from Figure 2.2 for data simulated in a single population using a local reference panel. Each plot is the result of MCMC runs performed on 100 simulated data sets. The simulated parameter values in the left plot represent the oldest TMRCAs and those in the right are the youngest.

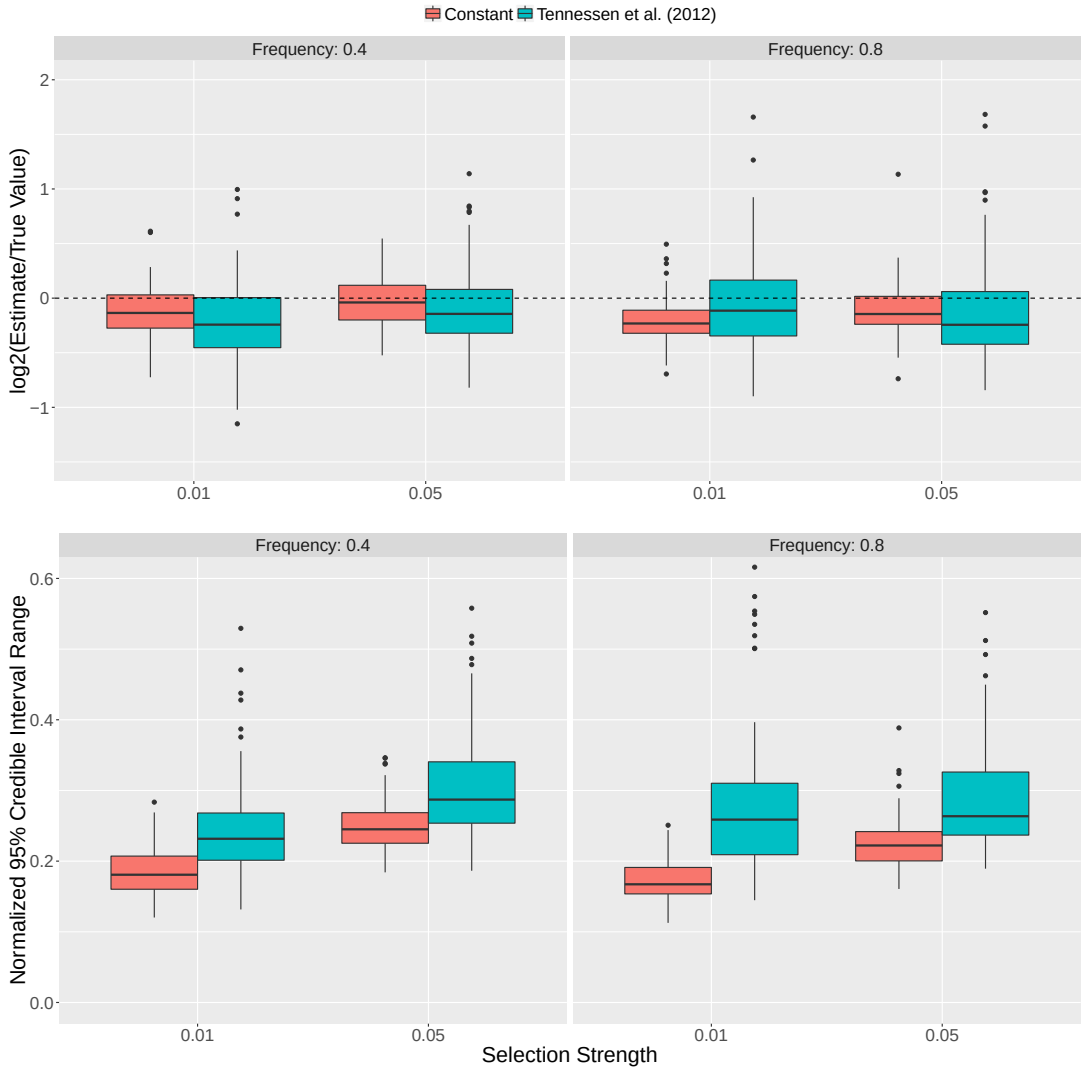

Figure S.10: Effects of non-equilibrium demographic history on estimate accuracy. A comparison of estimate accuracy and credible interval ranges using data simulated under the European demographic history inferred by Tennessen et al. [2012] and a constant population size model. To decrease computation time, we used a present day population size of 150,000 rather than 500,000. All relative changes in growth rate and bottleneck sizes are identical to those inferred by Tennessen et al. [2012]. We used a local reference panel for both demographic histories, and other parameter values are identical to those used for Figure 2.2 in the main text.

A.4 Chapter 2 Supplementary Tables

Selection Strength	Frequency	TMRCA
0.005	0.1	525
0.005	0.3	789
0.005	0.6	1030
0.005	0.9	1411
0.01	0.1	322
0.01	0.3	461
0.01	0.6	596
0.01	0.9	772
0.05	0.1	94
0.05	0.3	120
0.05	0.6	144
0.05	0.9	179

Table S.1: Simulated TMRCA values (mean generations). These are mean TMRCA values from simulations using 3 selection strengths and 4 ending frequencies for the beneficial allele. Each mean TMRCA is computed with 300 simulations.

Abbreviation	Sample
CHB	Han Chinese in Bejing, China
JPT	Japanese in Tokyo, Japan
CHS	Southern Han Chinese
CDX	Chinese Dai in Xishuangbanna, China
KHV	Kinh in Ho Chi Minh City, Vietnam
CEU	Utah Residents with Northern and Western European Ancestry
TSI	Toscani in Italia
FIN	Finnish in Finland
GBR	British in England and Scotland
IBS	Iberian Population in Spain
YRI	Yoruba in Ibadan, Nigeria
LWK	Luhya in Webuye, Kenya
GWD	Gambian in Western Divisions in the Gambia
MSL	Mende in Sierra Leone
ESN	Esan in Nigeria
GIH	Gujarati Indian from Houston, Texas
PJL	Punjabi from Lahore, Pakistan
BEB	Bengali from Bangladesh
STU	Sri Lankan Tamil from the UK
ITU	Indian Telugu from the UK

Table S.2: Sample abbreviations for the 1000 Genomes Project panel.

		Mbp-Scale Rec. Map		Fine-Scale Rec. Map	
Gene	Pop.	t_{ca} (years)	95% C.I.	t_{ca} (years)	95% C.I.
KITLG	FIN	18733	16675 - 20816	26343	21185 - 31439
KITLG	MSL	22339	15723 - 28949	20244	15042 - 26063
KITLG	LWK	22783	17921 - 28011	24200	16839 - 30730
KITLG	KHV	24544	21643 - 27192	26697	22249 - 31526
KITLG	ESN	26254	22854 - 29657	31791	26440 - 36543
KITLG	TSI	26427	24109 - 28905	22776	18379 - 27588
KITLG	CHS	26535	23456 - 29651	28396	23284 - 33294
KITLG	CHB	26772	24297 - 30141	28968	24451 - 34718
KITLG	GBR	26785	23841 - 29252	36132	31410 - 41697
KITLG	GWD	27669	21900 - 33664	25833	20134 - 32433
KITLG	ITU	28093	24607 - 31040	280906	24212 - 32250
KITLG	CDX	28362	25128 - 31245	29010	24457 - 33869
KITLG	JPT	28636	26351 - 31139	31634	27551 - 36471
KITLG	GIH	29029	25862 - 32439	28935	24599 - 33752
KITLG	IBS	29730	26169 - 32812	25373	21393 - 30031
KITLG	CEU	31287	27866 - 34512	34009	29818 - 38072
KITLG	STU	32021	8243 - 36318	27693	23968 - 32516
KITLG	BEB	32030	29000 - 34975	34375	29254 - 39578
KITLG	PJL	33719	30137 - 37310	31384	26814 - 36005
KITLG	YRI	33947	28861 - 39098	44437	36047 - 54074
EDAR	FIN	17386	13887 - 20794	20176	15053 - 25838
EDAR	BEB	18370	14325 - 22871	18418	13680 - 25409
EDAR	CHB	22192	19682 - 25735	19262	16921 - 21521
EDAR	JPT	23508	21595 - 25644	25730	23096 - 28826
EDAR	CHS	24058	22005 - 26678	24813	22493 - 27204
EDAR	CDX	24360	21572 - 27044	24346	21214 - 28019
EDAR	KHV	25683	23169 - 28379	12686	11001 - 14645

Table S.3: TMRCA estimates from the 1000 Genomes Project panel using the Mbp and fine-scale recombination rate. These results represent the distributions with the highest posterior probability among the 5 replicates shown with transparency in Figure 2.2 and Figure S.8 in Appendix A.3. All estimates are scaled to a generation time of 29 years.

		Mbp-Scale Rec. Map		Fine-Scale Rec. Map	
Gene	Pop.	t_{ca} (years)	95% C.I.	t_{ca} (years)	95% C.I
OCA2	KHV	16370	14439 - 18102	26904	22093 - 32402
OCA2	CHS	17316	14913 - 19799	26377	21217 - 31921
OCA2	CHB	17838	15336 - 20174	25159	20764 - 29688
OCA2	CDX	18083	6231 - 20253	28644	24241 - 33819
OCA2	JPT	18598	16110 - 20785	31582	27875 - 35522
ADH1B	KHV	10841	9720 - 12147	11186	9503 - 12862
ADH1B	CHS	12101	10668 - 13479	15352	12969 - 17974
ADH1B	CDX	12176	10678 - 136992	13568	11183 - 15941
ADH1B	JPT	13996	12670 - 15278	18317	15995 - 20911
ADH1B	CHB	15377	13763 - 17281	13526	11280 - 16210
LCT	BEB	6869	5143 - 8808	7971	5893 - 10443
LCT	FIN	7545	6982 - 8112	10332	9349 - 11427
LCT	ITU	7795	6199 - 9419	8972	7043 - 11015
LCT	TSI	7936	6616 - 9435	8630	70843 - 10230
LCT	STU	8197	6167 - 10338	7671	5205 - 10364
LCT	GBR	8412	7754 - 9084	8185	7111 - 9226
LCT	CEU	8662	80642 - 9340	10701	9579 - 11839
LCT	GIH	8732	7724 - 9921	9926	8596 - 11379
LCT	IBS	9341	8687 - 9988	7593	6602 - 8681
LCT	PJL	9514	8596 - 10382	9500	8511 - 10618

Table S.4: *Continued.*

Gene	Population	Years Before Present	Estimate	Information	Reference
LCT	CEU	8560 (7328 - 9861)	t_1	LD, freq.	Chen et al. [2015]
LCT	CEU	7466 (5516 - 11019)	t_1	LD, mut., freq.	Nakagome et al. [2016]
LCT	CEU	9277 (4021 - 21102)	t_1	LD, mut., freq.	Tishkoff et al. [2007]
LCT	CEU	13246 (2538 - 23954)	t_{ca}	LD	Bersaglieri et al. [2004]
LCT	Finland	2791 (1885 - 3698)	t_{ca}	LD	Bersaglieri et al. [2004]
LCT	Finland	5921 (5266 - 6576)	t_{ca}	mut.	Enattah et al. [2008]
LCT	Finland	6177 (5209 - 7145)	t_{ca}	mut.	Enattah et al. [2008]
LCT	Finland	6119 (5655 - 6542)	t_{ca}	LD	Enattah et al. [2007]
LCT	Finland	9425 (4350 - 18125)	t_{ca}	LD	Coelho et al. [2005]
LCT	Finland	7155 (77 - 26293)	t_{ca}	mut.	Enattah et al. [2007]
LCT	Finland	10730 (0 - 39440)	t_{ca}	mut.	Coelho et al. [2005]
LCT	Italy	9645 (3990 - 32120)	t_{ca}	LD	Coelho et al. [2005]
LCT	Italy	23710 (5000 - 66120)	t_{ca}	mut.	Coelho et al. [2005]
LCT	Portugal	10869 (6890 - 19940)	t_{ca}	LD	Coelho et al. [2005]
LCT	Portugal	21959 (4489 - 62199)	t_{ca}	mut.	Coelho et al. [2005]
LCT	Finland	12992 (1740 - 75284)	t_1	LD, mut., freq.	Peter et al. [2012]
LCT	European	8632 (7257 - 10020)	t_1	spatial	Itan et al. [2009]

Table S.5: Previous allele age point estimates and 95% confidence intervals for the loci considered in this study. All estimates are scaled to a generation time of 29 years and, where possible for SNP data, scaled to a mutation rate of 1.6×10^{-8}. For the times estimated in each case, t_1 refers to the time of mutation, t_{ca} is time to the common ancestor and t^{fix} is time since fixation [Przeworski, 2003].

Gene	Population	Years Before Present	Estimate	Information	Reference
KITLG	Portugal	32277 (6003 - 80683)	t_1	LD, mut., freq.	Beleza et al. [2013b]
KITLG	Han Chinese	32045 (6032 - 98165)	t_1	LD, mut., freq.	Beleza et al. [2013b]
KITLG	CEU	26386 (16846 - 56928)	t_1	LD, freq.	Chen et al. [2015]
OCA2	Han Chinese	17056 (12912 - 25246)	t_1	LD, freq.	Chen et al. [2015]
ADH1B	East Asians	4060 (2320 - 5800)	t_{ca}	mut.	Li et al. [2011]
ADH1B	East Asians	10026 (8512 - 11540)	t_{ca}	LD	Peng et al. [2010a]
ADH1B	Han Chinese	8957 (1533 - 34618)	t_1	LD, mut., freq.	Peter et al. [2012]
EDAR	East Asians	8666 (914 - 59711)	t^{fix}	LD, mut.	Bryk et al. [2008]
EDAR	Han Chinese	8463 (3192 - 32443)	t_1	LD, mut., freq.	Peter et al. [2012]
EDAR	Han Chinese	35873 (15283 - 45907)	t_1	freq. spec.	Kamberov et al. [2013]
EDAR	Han Chinese	42328 (40339 - 44317)	t_1	freq. spec.	Kamberov et al. [2013]

Table S.6: *Continued*

	startmrca									
	local ref. panel		diverged ref. panel		Chen et al. [2015]		ForSim		IS-Age	
	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE
Freq = 80%										
s = 0.005	-0.4786	0.5638	-0.1365	**0.3186**	-0.0410	0.3637	-1.2113	1.2313	0.0330	0.7255
s = 0.01	-0.3830	0.4657	0.0132	**0.3014**	-0.0190	0.3053	-0.2830	0.4125	-0.0984	0.6384
s = 0.05	-0.3074	0.4044	0.0027	0.3289	0.1469	**0.3253**	1.7866	1.8039	0.2719	0.4296
Freq = 40%										
s = 0.005	-0.5070	**0.5882**	-0.1607	**0.3801**	-0.0450	0.7080	-0.4826	0.6284	-0.3158	0.6331
s = 0.01	-0.3198	0.4294	-0.0132	**0.2897**	-0.0621	0.5949	0.2720	0.4275	-0.1586	0.8631
s = 0.05	-0.2009	**0.3731**	-0.0024	**0.3497**	0.2884	0.6554	2.1822	2.1994	0.4936	0.6391

Table S.7: Comparison of accuracy and bias results between our estimator "startmrca" and previously reported results from Chen et al. [2015] under different end frequencies (Freq) and selection strengths (s). Root mean squared errors (RMSE) and means were computed using $\log_2$(Estimated/True) TMRCA values. Results in bold indicate the method with lower RMSE values than the others. Simulations were matched to include a sample size of 200 haplotypes of length 1Mbp with a mutation and recombination rate of 1×10^{-8}. The diverged reference panel is sampled from a population that split with the beneficial allele carrier population .5N generations in the past. ForSim is the forward simulation method by Beleza et al. [2013a]; and IS-Age is the importance sampling-based method by Chen and Slatkin [2013].

			Model A		Model B	
Reference Panel	Selection	Frequency	Mean	RMSE	Mean	RMSE
Diverged	0.005	0.4	-0.16070	**0.38010**	-0.17842	0.38143
Diverged	0.01	0.4	-0.01320	**0.28970**	-0.00003	0.34665
Diverged	0.05	0.4	-0.00240	**0.34970**	0.03975	0.40088
Diverged	0.005	0.8	-0.13650	**0.31860**	-0.16190	0.33011
Diverged	0.01	0.8	0.01320	0.30140	0.05729	**0.28370**
Diverged	0.05	0.8	0.00270	**0.32890**	0.04515	0.36350
Local	0.005	0.4	-0.50700	**0.58820**	-0.52444	0.61450
Local	0.01	0.4	-0.31980	**0.42940**	-0.35912	0.44024
Local	0.05	0.4	-0.20090	**0.37310**	-0.23170	0.37473
Local	0.005	0.8	-0.47860	**0.56380**	-0.52219	0.60239
Local	0.01	0.8	-0.38300	**0.46570**	-0.44445	0.50871
Local	0.05	0.8	-0.30740	0.40440	-0.30716	**0.40357**

Table S.8: Comparison of accuracy and bias results between different approaches for modelling invariant sites among background haplotypes in the carrier panel (β_{iw}). Model A refers to the original Li and Stephens [2003] model. Model B uses the singleton rate in the reference panel (see Appendix A.2). As in Table S.7, root mean squared errors (RMSE) and means were computed using $\log_2$(Estimated/True) TMRCA values. Results in bold indicate the model with lowest RMSE value. Frequency refers to the end frequency of the beneficial allele trajectory.

	Posterior		Bootstraps	
	Mean	95% C.I.	Mean	95% C.I.
Freq = 80%				
s = 0.01	728	(643 - 813)	721	(623 - 838)
s = 0.1	76	(62 - 90)	77	(67 - 88)
Freq = 40%				
s = 0.01	388	(345 - 434)	413	(360 - 485)
s = 0.1	52	(43 - 62)	53	(44 - 64)

Table S.9: Bootstrap Estimate Comparisons. Comparison of TMRCA estimates from posterior results of simulated data versus estimates from 100 bootstrap replicates of those same datasets. For each dataset, we simulated 100 beneficial allele carriers and 20 non-carriers for the reference panel. Bootstrap replicates were generated by resampling among the beneficial allele carriers. We used mutation and recombination rates of 1×10^{-8} and a population size of 10000.

A.5 Gamete Frequency Trajectories

We can use the epistatic selection model described in Gavrilets [1997] to generate expected genotype frequencies at two focal loci separated by a recombination distance of r in a hybrid zone. Let x_1, x_2, x_3, x_4 refer to the frequencies of gametes **AB**, **Ab**, **aB** and **ab** at the beginning of a generation. In a single randomly mating population, gamete frequencies after a generation of selection and recombination are

$$x'_g = \begin{cases} \frac{\omega_g x_g - r\omega_{14} D}{\bar{\omega}} & \text{if } g = 1, 4; \\ \\ \frac{\omega_g x_g + r\omega_{14} D}{\bar{\omega}} & \text{if } g = 2, 3 \end{cases} \tag{7}$$

where $\omega_i = \sum_j \omega_{ij} x_j$ is the fitness of gamete i, $\bar{\omega} = \sum_{ij} w_{ij} x_i x_j$ is mean fitness, and D is the linkage disequilibrium computed as $D = x_1 x_4 - x_2 x_3$. The fitness matrix

of two locus genotypes is given in Table 4.1. There are a variety of hypotheses for different specifications of the fitness matrix. For Dobzhansky-Muller type interactions, **aaBB** is the ancestral genotype from which two derived genotypes (**aabb** and **AABB**) are descended in different species or populations.

The Gavrilets [1997] model uses a linear stepping stone scheme of two infinite source populations with fixed ancestry for either species (**aabb** and **AABB**) contributing to two or more subpopulations between them. We simplify this model to a single hybrid population with equal migrant contributions from both source populations. Let the migration rate from each of the source populations (A or B) into the hybrid population be m. Let $x''_{g,h}$, $x''_{g,A}$ and $x''_{g,B}$ denote the frequency of gamete g in the hybrid zone and the two source populations. After migration,

$$x''_{g,h} = (1-2m)x'_g + mx'_{g,A} + mx'_{g,B} \tag{8}$$

We also consider a random sampling component to incorporate the effects of genetic drift on the random union of gametes in a finite population. Before selection and migration in a population of size N, we model the probability of a particular pairing of gametes in a zygote as multinomial with sample size $2N$, and probabilities x_1^2, $2x_1x_2$, x_2^2, $2x_1x_3$, $2x_1x_4 + 2x_2x_3$, $2x_2x_4$, x_3^2, $2x_3x_4$, and x_4^2 for each zygote **AABB**, **AABb**, **AAbb**, **AaBB**, **AaBb**, **Aabb**, **aaBB**, **aaBb**, and **aabb**.

REFERENCES

Richard Abbott, Dirk Albach, Stephen Ansell, Jan W Arntzen, Stuart JE Baird, Nicolas Bierne, Jenny Boughman, Alan Brelsford, C Alex Buerkle, Richard Buggs, et al. Hybridization and speciation. *Journal of Evolutionary Biology*, 26(2):229–246, 2013.

Joshua M Akey, Ge Zhang, Kun Zhang, Li Jin, and Mark D Shriver. Interrogating a high-density snp map for signatures of natural selection. *Genome Research*, 12 (12):1805–1814, 2002.

Joshua M Akey, Michael A Eberle, Mark J Rieder, Christopher S Carlson, Mark D Shriver, Deborah A Nickerson, and Leonid Kruglyak. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology*, 2:1591–1599, 2004.

Mark Aldenderfer. Peopling the tibetan plateau: insights from archaeology. *High Altitude Medicine & Biology*, 12(2):141–147, 2011.

Morten E Allentoft, Martin Sikora, Karl-Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B Damgaard, Hannes Schroeder, Torbjörn Ahlström, Lasse Vinner, et al. Population genomics of Bronze Age Eurasia. *Nature*, 522(7555):167–172, 2015.

Tovi M Anderson, Sophie I Candille, Marco Musiani, Claudia Greco, Daniel R Stahler, Douglas W Smith, Badri Padhukasahasram, Ettore Randi, Jennifer A Leonard, Carlos D Bustamante, et al. Molecular and evolutionary history of melanism in north american gray wolves. *Science*, 323(5919):1339–1343, 2009.

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

Michael L Arnold, Jennafer AP Hamlin, Amanda N Brothers, and Evangeline S Ballerini. Natural hybridization as a catalyst of rapid evolutionary change. *Rapidly Evolving Genes and Genetic Systems*, 256:265, 2012.

Adam Auton and Gil McVean. Estimating recombination rates from genetic variation in humans. In *Evolutionary Genomics*, pages 217–237. Springer, 2012.

Adam Auton, Ying Rui Li, Jeffrey Kidd, Kyle Oliveira, Julie Nadel, J Kim Holloway, Jessica J Hayward, Paula E Cohen, John M Greally, Jun Wang, et al. Genetic

recombination is targeted towards gene promoter regions in dogs. *PLoS Genetics*, 9(12):e1003984, 2013.

Zachary Baker, Molly Schumer, Yuki Haba, Lisa Bashkirova, Chris Holland, Gil G Rosenthal, and Molly Przeworski. Repeated losses of prdm9-directed recombination despite the conservation of prdm9 across vertebrates. *Elife*, 6, 2017.

Rowan DH Barrett and Hopi E Hoekstra. Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, 12(11):767–780, 2011.

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.

Frédéric Baudat, Jérôme Buard, Corinne Grey, Adi Fledel-Alon, Carole Ober, Molly Przeworski, Graham Coop, and Bernard De Massy. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327(5967): 836–840, 2010.

Cynthia M Beall, Gianpiero L Cavalleri, Libin Deng, Robert C Elston, Yang Gao, Jo Knight, Chaohua Li, Jiang Chuan Li, Yu Liang, Mark McCormack, et al. Natural selection on epas1 (hif2α) associated with low hemoglobin concentration in tibetan highlanders. *Proceedings of the National Academy of Sciences*, 107(25): 11459–11464, 2010.

Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Sandra Beleza, Nicholas A Johnson, Sophie I Candille, Devin M Absher, Marc A Coram, Jailson Lopes, Joana Campos, Isabel Inês Araújo, Tovi M Anderson, Bjarni J Vilhjálmsson, et al. Genetic architecture of skin and eye color in an African-European admixed population. *PLoS Genetics*, 9(3):e1003372, 2013a.

Sandra Beleza, António M Santos, Brian McEvoy, Isabel Alves, Cláudia Martinho, Emily Cameron, Mark D Shriver, Esteban J Parra, and Jorge Rocha. The timing of pigmentation lightening in Europeans. *Molecular Biology and Evolution*, 30(1): 24–35, 2013b.

Jeremy J Berg and Graham Coop. A Coalescent model for a sweep of a unique standing variant. *Genetics*, 201(2):707–725, 2015.

Todd Bersaglieri, Pardis C Sabeti, Nick Patterson, Trisha Vanderploeg, Steve F Schaffner, Jared A Drake, Matthew Rhodes, David E Reich, and Joel N Hirschhorn. Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74(6):1111–1120, 2004.

Abigail Bigham, Marc Bauchet, Dalila Pinto, Xianyun Mao, Joshua M Akey, Rui Mei, Stephen W Scherer, Colleen G Julian, Megan J Wilson, David López Herráez, et al. Identifying signatures of natural selection in tibetan and andean populations using dense genome scan data. *PLoS Genetics*, 6(9):e1001116, 2010.

Karl W Broman, Jeffrey C Murray, Val C Sheffield, Raymond L White, and James L Weber. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *The American Journal of Human Genetics*, 63(3):861–869, 1998.

Andrew VZ Brower. A new mimetic species of heliconius (lepidoptera: Nymphalidae), from southeastern colombia, revealed by cladistic analysis of mitochondrial dna sequences. *Zoological Journal of the Linnean Society*, 116(3):317–332, 1996.

Andrew VZ Brower. Hybrid speciation in heliconius butterflies? a review and critique of the evidence. *Genetica*, 139(5):589–609, 2011.

Andrew VZ Brower. Introgression of wing pattern alleles and speciation via homoploid hybridization in heliconius butterflies: a review of evidence from the genome. *Proceedings of the Royal Society of London B: Biological Sciences*, 280 (1752):20122302, 2013.

Jarosław Bryk, Emilie Hardouin, Irina Pugach, David Hughes, Rainer Strotmann, Mark Stoneking, and Sean Myles. Positive selection in East Asians for an EDAR allele that enhances NF-κB activation. *PLoS One*, 3(5):e2209–e2209, 2008.

Vanessa Bull, Margarita Beltrán, Chris D Jiggins, W Owen McMillan, Eldredge Bermingham, and James Mallet. Polyphyly and gene flow between non-sibling heliconius species. *BMC Biology*, 4(1):11, 2006.

Luther Burbank. *How Plants are Trained to Work for Man: Trees. Bibliography. Index*, volume 8. PF Collier & Son Company, 1921.

Howard M Cann, Claudia De Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, Walter F Bodmer, Batsheva Bonne-Tamir, Anne Cambon-Thomsen, et al. A human genome diversity cell line panel. *Science*, 296(5566):261–262, 2002.

Rebecca L Cann, Mark Stoneking, and Allan C Wilson. Mitochondrial dna and human evolution. *Nature*, 325(6099):31–36, 1987.

Nicola L Chamberlain, Ryan I Hill, Durrell D Kapan, Lawrence E Gilbert, and Marcus R Kronforst. Polymorphic butterfly reveals the missing link in ecological speciation. *Science*, 326(5954):847–850, 2009.

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`.

Hua Chen and Montgomery Slatkin. Inferring selection intensity and allele age from multilocus haplotype structure. *G3: Genes— Genomes— Genetics*, 3(8):1429–1442, 2013.

Hua Chen, Jody Hey, and Montgomery Slatkin. A hidden Markov model for investigating recent positive selection through haplotype structure. *Theoretical Population Piology*, 99:18–30, 2015.

Sung Chun and Justin C Fay. Evidence for hitchhiking of deleterious mutations within the human genome. *PLoS Genetics*, 7(8):e1002240, 2011.

Andrew G Clark, Xu Wang, and Tara Matise. Contrasting methods of quantifying fine structure of human recombination. *Annual Review of Genomics and Human Genetics*, 11:45–64, 2010.

Margarida Coelho, Donata Luiselli, Giorgio Bertorelle, Ana Isabel Lopes, Susana Seixas, Giovanni Destro-Bisol, and Jorge Rocha. Microsatellite variation and evolution of human lactase persistence. *Human Genetics*, 117(4):329–339, 2005.

Graham Coop and Robert C Griffiths. Ancestral inference on gene trees under selection. *Theoretical Population Piology*, 66(3):219–232, 2004.

Graham Coop and Molly Przeworski. An evolutionary view of human recombination. *Nature Reviews Genetics*, 8(1):23, 2007.

Graham Coop, Kevin Bullaughey, Francesca Luca, and Molly Przeworski. The timing of selection at the human FOXP2 gene. *Molecular Biology and Evolution*, 25(7): 1257–1259, 2008.

Graham Coop, Joseph K Pickrell, John Novembre, Sridhar Kudaravalli, Jun Li, Devin Absher, Richard M Myers, Luigi Luca Cavalli-Sforza, Marcus W Feldman, and Jonathan K Pritchard. The role of geography in human adaptation. *PLoS Genetics*, 5(6):e1000500, 2009.

Tim Coulson, Daniel R MacNulty, Daniel R Stahler, Robert K Wayne, Douglas W Smith, et al. Modeling effects of environmental change on wolf population dynamics, trait evolution, and life history. *Science*, 334(6060):1275–1278, 2011.

JA Coyne and HA Orr. Speciation. sinauer. *Sunderland, MA*, 2004.

Anne C Dalziel, Sean M Rogers, and Patricia M Schulte. Linking genotypes to phenotypes and fitness: how mechanistic biology can inform molecular ecology. *Molecular Ecology*, 18(24):4997–5017, 2009.

Kanchon K Dasmahapatra, James R Walters, Adriana D Briscoe, John W Davey, Annabel Whibley, Nicola J Nadeau, Aleksey V Zimin, Daniel ST Hughes, Laura C Ferguson, Simon H Martin, et al. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405):94, 2012.

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

Kevin P Donnelly. The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*, 23(1):34–63, 1983.

Anne-Lyse Ducrest, Laurent Keller, and Alexandre Roulin. Pleiotropy in the melanocortin system, coloration and behavioural syndromes. *Trends in Ecology & Evolution*, 23(9):502–510, 2008.

Eric Y Durand, Nick Patterson, David Reich, and Montgomery Slatkin. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8):2239–2252, 2011.

Richard Durrett and Jason Schweinsberg. Approximating selective sweeps. *Theoretical Population Piology*, 66(2):129–138, 2004.

Katherine Eaton, Melissa Edwards, S Krithika, Gillian Cook, Heather Norton, and Esteban J Parra. Association study confirms the role of two OCA2 polymorphisms in normal skin pigmentation variation in East Asian populations. *The American Journal of Human Biology*, 2015.

Melissa Edwards, Abigail Bigham, Jinze Tan, Shilin Li, Agnes Gozdzik, Kendra Ross, Li Jin, and Esteban J Parra. Association of the OCA2 polymorphism His615Arg with melanin content in east Asian populations: further evidence of convergent evolution of skin pigmentation. *PLoS Genetics*, 6(3):e1000867, 2010.

Kathryn R Elmer and Axel Meyer. Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology & Evolution*, 26(6): 298–306, 2011.

David Enard, Philipp W Messer, and Dmitri A Petrov. Genome-wide signals of positive selection in human evolution. *Genome Research*, 24(6):885–895, 2014.

Nabil Sabri Enattah, Timo Sahi, Erkki Savilahti, Joseph D Terwilliger, Leena Peltonen, and Irma Järvelä. Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, 30(2):233–237, 2002.

Nabil Sabri Enattah, Aimee Trudeau, Ville Pimenoff, Luigi Maiuri, Salvatore Auricchio, Luigi Greco, Mauro Rossi, Michael Lentze, JK Seo, Soheila Rahgozar, et al. Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. *The American Journal of Human Genetics*, 81(3):615–625, 2007.

Nabil Sabri Enattah, Tine GK Jensen, Mette Nielsen, Rikke Lewinski, Mikko Kuokkanen, Heli Rasinpera, Hatem El-Shanti, Jeong Kee Seo, Michael Alifrangis, Insaf F Khalil, et al. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *The American Journal of Human Genetics*, 82(1):57–72, 2008.

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.

Mimy Y Eng, Susan E Luczak, and Tamara L Wall. ALDH2, ADH1B, and ADH1C genotypes in Asians: a literature review. *Alcohol Research and Health*, 30(1):22, 2007.

Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003.

Jack N Fenner. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *The American Journal of Physical Anthropology*, 128(2):415–423, 2005.

Laurent AF Frantz, Victoria E Mullin, Maud Pionnier-Capitan, Ophélie Lebrasseur, Morgane Ollivier, Angela Perri, Anna Linderholm, Valeria Mattiangeli, Matthew D Teasdale, Evangelos A Dimopoulos, et al. Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science*, 352(6290):1228–1231, 2016.

Akihiro Fujimoto, Jun Ohashi, Nao Nishida, Taku Miyagawa, Yasuyuki Morishita, Tatsuhiko Tsunoda, Ryosuke Kimura, and Katsushi Tokunaga. A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. *Human Genetics*, 124(2):179–185, 2008.

Matteo Fumagalli, Ida Moltke, Niels Grarup, Fernando Racimo, Peter Bjerregaard, Marit E Jørgensen, Thorfinn S Korneliussen, Pascale Gerbault, Line Skotte, Allan Linneberg, et al. Greenlandic inuit show genetic signatures of diet and climate adaptation. *Science*, 349(6254):1343–1347, 2015.

Stacey B Gabriel, Stephen F Schaffner, Huy Nguyen, Jamie M Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, et al. The structure of haplotype blocks in the human genome. *Science*, 296 (5576):2225–2229, 2002.

Nandita R Garud, Philipp W Messer, Erkan O Buzbas, and Dmitri A Petrov. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genetics*, 11(2):e1005004, 2015.

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

Hilda Geiringer. On the probability theory of linkage in mendelian heredity. *The Annals of Mathematical Statistics*, 15(1):25–57, 1944.

1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012.

LE Gilbert. Adaptive novelty through introgression in heliconius wing patterns: evidence for shared genetic "tool box" from synthetic hybrid zones and a theory of diversification. *Ecology and evolution taking flight: butterflies as model systems*, pages 281–318, 2003.

Philip S Gipson, Edward E Bangs, Theodore N Bailey, Diane K Boyd, H Dean Cluff, Douglas W Smith, and Michael D Jiminez. Color patterns among wolves in western north america. *Wildlife Society Bulletin*, pages 821–830, 2002.

Nathalia Giraldo, Camilo Salazar, Chris D Jiggins, Eldredge Bermingham, and Mauricio Linares. Two sisters in the same dress: Heliconius cryptic species. *BMC Evolutionary Biology*, 8(1):324, 2008.

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.

Simon Gravel, Fouad Zakharia, Andres Moreno-Estrada, Jake K Byrnes, Marina Muzzio, Juan L Rodriguez-Flores, Eimear E Kenny, Christopher R Gignoux, Brian K Maples, Wilfried Guiblet, et al. Reconstructing native american migrations from whole-genome and whole-exome data. *PLoS genetics*, 9(12):e1004023, 2013.

Robert C Griffiths and Simon Tavare. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 344(1310):403–410, 1994.

Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stewardson, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 2015.

International HapMap Consortium. The international hapmap project. *Nature*, 426 (6968):789, 2003.

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.

Matthew Hartfield and Sarah P Otto. Recombination and hitchhiking of deleterious alleles. *Evolution*, 65(9):2421–2434, 2011.

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.

Joachim Hermisson and Pleuni S Pennings. Soft sweeps molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4):2335–2352, 2005.

Anjali G Hinch, Arti Tandon, Nick Patterson, Yunli Song, Nadin Rohland, Cameron D Palmer, Gary K Chen, Kai Wang, Sarah G Buxbaum, Ermeg L Akylbekova, et al. The landscape of recombination in African Americans. *Nature*, 476 (7359):170–175, 2011.

Richard R Hudson. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7(1):44, 1990.

Richard R Hudson. The variance of coalescent time estimates from DNA sequences. *Journal of Molecular Evolution*, 64(6):702–705, 2007.

Richard R Hudson and Norman L Kaplan. The coalescent process in models with selection and recombination. *Genetics*, 120(3):831–840, 1988.

RR Hudson. Linkage disequilibrium and recombination. *Handbook of Statistical Genetics*, 2001.

Emilia Huerta-Sánchez, Xin Jin, Zhuoma Bianba, Benjamin M Peter, Nicolas Vinckenbosch, Yu Liang, Xin Yi, Mingze He, Mehmet Somel, Peixiang Ni, et al. Altitude adaptation in tibetans caused by introgression of denisovan-like dna. *Nature*, 512 (7513):194, 2014.

Chad D Huff, David J Witherspoon, Yuhua Zhang, Chandler Gatenbee, Lee A Denson, Subra Kugathasan, Hakon Hakonarson, April Whiting, Chadwick T Davis, Wilfred Wu, et al. Crohn's disease and genetic hitchhiking at IBD5. *Molecular Biology and Evolution*, 29(1):101–111, 2012.

Hideki Innan and Yuseob Kim. Pattern of polymorphism after strong artificial selection in a domestication event. *Proceedings of the National Academy of Sciences*, 101(29):10667–10672, 2004.

Yuval Itan, Adam Powell, Mark A Beaumont, Joachim Burger, Mark G Thomas, et al. The origins of lactase persistence in Europe. *PLoS Computational Biology*, 5(8):e1000491–e1000491, 2009.

Nina G Jablonski and George Chaplin. The evolution of human skin coloration. *The Journal of Human Evolution*, 39(1):57–106, 2000.

Nina G Jablonski and George Chaplin. Human skin pigmentation, migration and disease susceptibility. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1590):785–792, 2012.

Choongwon Jeong and Anna Di Rienzo. Adaptations to local environments in modern human populations. *Current Opinion in Genetics & Development*, 29:1–8, 2014.

Choongwon Jeong, Gorka Alkorta-Aranburu, Buddha Basnyat, Maniraj Neupane, David B Witonsky, Jonathan K Pritchard, Cynthia M Beall, and Anna Di Rienzo. Admixture facilitates genetic adaptations to high altitude in tibet. *Nature Communications*, 5:3281, 2014.

Kelsey Elizabeth Johnson and Benjamin F Voight. Patterns of shared signatures of recent positive selection across human populations. *Nature Ecology & Evolution*, page 1, 2018.

Mathieu Joron and James LB Mallet. Diversity in mimicry: paradox or paradigm? *Trends in Ecology & Evolution*, 13(11):461–466, 1998.

Mathieu Joron, Lise Frezal, Robert T Jones, Nicola L Chamberlain, Siu F Lee, Christoph R Haag, Annabel Whibley, Michel Becuwe, Simon W Baxter, Laura Ferguson, et al. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, 477(7363):203, 2011.

Yana G Kamberov, Sijia Wang, Jingze Tan, Pascale Gerbault, Abigail Wark, Longzhi Tan, Yajun Yang, Shilin Li, Kun Tang, Hua Chen, et al. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell*, 152(4):691–702, 2013.

Norman L Kaplan, Thomas Darden, and Richard R Hudson. The coalescent process in models with selection. *Genetics*, 120(3):819–829, 1988.

Norman L Kaplan, RR Hudson, and CH Langley. The" hitchhiking effect" revisited. *Genetics*, 123(4):887–899, 1989.

Jeffrey M Kidd, Simon Gravel, Jake Byrnes, Andres Moreno-Estrada, Shaila Musharoff, Katarzyna Bryc, Jeremiah D Degenhardt, Abra Brisbin, Vrunda Sheth, Rong Chen, et al. Population genetic inference from personal genome data: impact

of ancestry and admixture on human genomic variation. *The American Journal of Human Genetics*, 91(4):660–671, 2012.

Ryosuke Kimura, Tetsutaro Yamaguchi, Mayako Takeda, Osamu Kondo, Takashi Toma, Kuniaki Haneji, Tsunehiko Hanihara, Hirotaka Matsukusa, Shoji Kawamura, Koutaro Maki, et al. A common variation in EDAR is a genetic determinant of shovel-shaped incisors. *The American Journal of Human Genetics*, 85(4): 528–535, 2009.

Jan Klein, Akie Sato, Sandra Nagl, and Colm O'hUigín. Molecular trans-species polymorphism. *Annual Review of Ecology and Systematics*, 29(1):1–21, 1998.

Augustine Kong, Gudmar Thorleifsson, Daniel F Gudbjartsson, Gisli Masson, Asgeir Sigurdsson, Aslaug Jonasdottir, G Bragi Walters, Adalbjorg Jonasdottir, Arnaldur Gylfason, Kari Th Kristinsson, et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103, 2010.

Marcus R Kronforst. Gene flow persists millions of years after speciation in heliconius butterflies. *BMC Evolutionary Biology*, 8(1):98, 2008.

Marcus R Kronforst, Laura G Young, Lauren M Blume, and Lawrence E Gilbert. Multilocus analyses of admixture and introgression among hybridizing heliconius butterflies. *Evolution*, 60(6):1254–1268, 2006.

Mikko Kuokkanen, Jorma Kokkonen, Nabil Sabri Enattah, Tero Ylisaukko-oja, Hanna Komu, Teppo Varilo, Leena Peltonen, Erkki Savilahti, and Irma Järvelä. Mutations in the translated region of the lactase gene (LCT) underlie congenital lactase deficiency. *The American Journal of Human Genetics*, 78(2):339–344, 2006.

Fabrice Larribe and Paul Fearnhead. On composite likelihoods in statistical genetics. *Statistica Sinica*, pages 43–69, 2011.

Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1):e1002453, 2012.

Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H Sudmant, Joshua G Schraiber, Sergi Castellano, Mark Lipson, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, 2014.

Jennifer A Leonard, Robert K Wayne, Jane Wheeler, Raúl Valadez, Sonia Guillén, and Carles Vila. Ancient dna evidence for old world origin of new world dogs. *Science*, 298(5598):1613–1616, 2002.

Stephen Leslie, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, Ellen C Royrvik, Barry Cunliffe, Daniel J Lawson, et al. The fine-scale genetic structure of the british population. *Nature*, 519(7543):309, 2015.

Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

Hui Li, Namita Mukherjee, Usha Soundararajan, Zsanett Tárnok, Csaba Barta, Shagufta Khaliq, Aisha Mohyuddin, Sylvester LB Kajuna, S Qasim Mehdi, Judith R Kidd, et al. Geographically separate increases in the frequency of the derived ADH1B* 47His allele in eastern and western Asia. *The American Journal of Human Genetics*, 81(4):842–846, 2007.

Hui Li, Sheng Gu, Yi Han, Zhi Xu, Andrew J Pakstis, Li Jin, Judith R Kidd, and Kenneth K Kidd. Diversification of the ADH1B gene during expansion of modern humans. *Annals of Human Genetics*, 75(4):497–507, 2011.

Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165 (4):2213–2233, 2003.

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.

Xuanyao Liu, Rick Twee-Hee Ong, Esakimuthu Nisha Pillai, Abier M Elzein, Kerrin S Small, Taane G Clark, Dominic P Kwiatkowski, and Yik-Ying Teo. Detecting and characterizing genomic signatures of positive selection in global populations. *The American Journal of Human Genetics*, 92(6):866–881, 2013.

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.

WF Loomis. Skin-pigment regulation of vitamin-D biosynthesis in man. *Science*, 157(3788):501–506, 1967.

Dongsheng Lu, Haiyi Lou, Kai Yuan, Xiaoji Wang, Yuchen Wang, Chao Zhang, Yan Lu, Xiong Yang, Lian Deng, Ying Zhou, et al. Ancestral origins and genetic history of tibetan highlanders. *The American Journal of Human Genetics*, 99(3): 580–594, 2016.

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

James Mallet. Causes and consequences of a lack of coevolution in müllerian mimicry. *Evolutionary Ecology*, 13(7-8):777–806, 1999.

James Mallet and Mathieu Joron. Evolution of diversity in warning color and mimicry: polymorphisms, shifting balance, and speciation. *Annual Review of Ecology and Systematics*, 30(1):201–233, 1999.

James Mallet, Margarita Beltrán, Walter Neukirchen, and Mauricio Linares. Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evolutionary Biology*, 7(1):28, 2007.

T Maruyama. The age of a rare mutant gene in a large population. *American Journal of Human Genetics*, 26(6):669, 1974.

Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Bastien Llamas, Joseph Pickrell, Harald Meller, Manuel A Rojo Guerra, Johannes Krause, David Anthony, et al. Eight thousand years of natural selection in Europe. *BioRxiv*, page 016477, 2015a.

Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson, Songül Alpaslan Roodenberg, Eadaoin Harney, Kristin Stewardson, Daniel Fernandes, Mario Novak, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503, 2015b.

Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

Mary Sara McPeek and Andrew Strahs. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *The American Journal of Human Genetics*, 65(3):858–875, 1999.

Loukia Meligkotsidou and Paul Fearnhead. Maximum-likelihood estimation of coalescence times in genealogical trees. *Genetics*, 171(4):2073–2084, 2005.

Fernando L Mendez, Joseph C Watkins, and Michael F Hammer. A haplotype at stat2 introgressed from neanderthals and serves as a candidate of positive selection in papua new guinea. *The American Journal of Human Genetics*, 91(2):265–274, 2012.

Philipp W Messer and Richard A Neher. Estimating the strength of selective sweeps from deep population diversity data. *Genetics*, 191(2):593–605, 2012.

Philipp W Messer and Dmitri A Petrov. Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution*, 28(11):659–669, 2013.

Craig T Miller, Sandra Beleza, Alex A Pollen, Dolph Schluter, Rick A Kittles, Mark D Shriver, and David M Kingsley. cis-Regulatory changes in kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell*, 131(6):1179–1189, 2007.

LG Moore, M Shriver, L Bemis, B Hickler, M Wilson, T Brutsaert, E Parra, and E Vargas. Maternal adaptation to high-altitude pregnancy: an experiment of nature—a review. *Placenta*, 25:S60–S71, 2004.

Lorna G Moore, David Young, Robert E McCullough, Tarshi Droma, and Stacy Zamudio. Tibetan protection from intrauterine growth restriction (iugr) and reproductive loss at high altitude. *American Journal of Human Biology*, 13(5): 635–644, 2001.

Andrés Moreno-Estrada, Simon Gravel, Fouad Zakharia, Jacob L McCauley, Jake K Byrnes, Christopher R Gignoux, Patricia A Ortiz-Tello, Ricardo J Martínez, Dale J Hedges, Richard W Morris, et al. Reconstructing the population genetic history of the caribbean. *PLoS genetics*, 9(11):e1003925, 2013.

Andrew P Morris, John C Whittaker, and David J Balding. Bayesian fine-scale mapping of disease loci, by hidden Markov models. *The American Journal of Human Genetics*, 67(1):155–169, 2000.

Andrew P Morris, John C Whittaker, and David J Balding. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *The American Journal of Human Genetics*, 70(3):686–707, 2002.

Simon Myers, Rory Bowden, Afidalina Tumian, Ronald E Bontrop, Colin Freeman, Tammie S MacFie, Gil McVean, and Peter Donnelly. Drive against hotspot motifs in primates implicates the prdm9 gene in meiotic recombination. *Science*, 327 (5967):876–879, 2010.

Shigeki Nakagome, Gorka Alkorta-Aranburu, Roberto Amato, Bryan Howie, Benjamin M Peter, Richard R Hudson, and Anna Di Rienzo. Estimating the ages of selection signals from different epochs in human history. *Molecular Biology and Evolution*, 33(3):657–669, 2016.

Masatoshi Nei and Wen-Hsiung Li. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10):5269–5273, 1979.

Rasmus Nielsen. Molecular signatures of natural selection. *Annual Review of Genetics*, 39:197–218, 2005.

Susan Niermeyer, P Andrade Mollinedo, and L Huicho. Child health and living at high altitude. *Archives of Disease in Childhood*, 94(10):806–811, 2009.

Lynne C Olds and Eric Sibley. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Human Molecular Genetics*, 12(18):2333–2340, 2003.

Louise Ormond, Matthieu Foll, Gregory B Ewing, Susanne P Pfeifer, and Jeffrey D Jensen. Inferring the age of a fixed beneficial allele. *Molecular Ecology*, 25(1): 157–169, 2016.

Michael V Osier, Andrew J Pakstis, Himla Soodyall, David Comas, David Goldman, Adekunle Odunsi, Friday Okonofua, Josef Parnas, Leslie O Schulz, Jaume Bertranpetit, et al. A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. *The American Journal of Human Genetics*, 71(1):84–99, 2002.

Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Pe'er. Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, 91(5):809–822, 2012.

Carolina Pardo-Diaz, Camilo Salazar, Simon W Baxter, Claire Merot, Wilsea Figueiredo-Ready, Mathieu Joron, W Owen McMillan, and Chris D Jiggins. Adaptive introgression across species boundaries in heliconius butterflies. *PLoS Genetics*, 8(6):e1002752, 2012.

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.

M Pazgier, DM Hoover, D Yang, W Lu, and J Lubkowski. Human β-defensins. *Cellular and Molecular Life Sciences CMLS*, 63(11):1294–1313, 2006.

Yi Peng, Hong Shi, Xue-bin Qi, Chun-jie Xiao, Hua Zhong, Z Ma Run-lin, and Bing Su. The ADH1B Arg47His polymorphism in East Asian populations and expansion of rice domestication in history. *BMC Evolutionary Biology*, 10(1):15, 2010a.

Yi Peng, Zhaohui Yang, Hui Zhang, Chaoying Cui, Xuebin Qi, Xiongjian Luo, Xiang Tao, Tianyi Wu, Hua Chen, Hong Shi, et al. Genetic variations in tibetan populations and high-altitude adaptation at the himalayas. *Molecular Biology and Evolution*, 28(2):1075–1081, 2010b.

Benjamin M Peter, Emilia Huerta-Sanchez, and Rasmus Nielsen. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genetics*, 8(10):e1003011, 2012.

Joseph K Pickrell, Graham Coop, John Novembre, Sridhar Kudaravalli, Jun Z Li, Devin Absher, Balaji S Srinivasan, Gregory S Barsh, Richard M Myers, Marcus W Feldman, et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, 19(5):826–837, 2009.

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.

Molly Prezeworski, Graham Coop, and Jeffrey D Wall. The signature of positive selection on standing genetic variation. *Evolution*, 59(11):2312–2323, 2005.

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.

Jonathan K Pritchard and Molly Przeworski. Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, 69(1):1–14, 2001.

Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

Jonathan K Pritchard, Joseph K Pickrell, and Graham Coop. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, 20(4):R208–R215, 2010.

Molly Przeworski. Estimating the time since the fixation of a beneficial allele. *Genetics*, 164(4):1667–1676, 2003.

Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Jacek Radwan and Wiesław Babik. The genomics of adaptation. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1749):5024–5028, 2012.

Peter Ralph and Graham Coop. The geography of recent genetic ancestry across europe. *PLoS Biology*, 11(5):e1001555, 2013.

Bruce Rannala and Jeff P Reeve. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *The American Journal of Human Genetics*, 69(1):159–178, 2001.

Bruce Rannala and Jeff P Reeve. Joint Bayesian estimation of mutation location and age using linkage disequilibrium. In *Pacific Symposium on Biocomputing*, pages 526–534. World Scientific, 2003.

David Reich, Richard E Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y Durand, Bence Viola, Adrian W Briggs, Udo Stenzel, Philip LF Johnson, et al. Genetic history of an archaic hominin group from denisova cave in siberia. *Nature*, 468(7327):1053, 2010.

Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

Pardis C Sabeti, David E Reich, John M Higgins, Haninah ZP Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J McDonald, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002.

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.

Rena M Schweizer, Arun Durvasula, Joel Smith, Samuel H Vohr, Daniel R Stahler, Marco Galaverni, Olaf Thalmann, Douglas W Smith, Ettore Randi, Elaine A Ostrander, Richard E Green, Kirk E Lohmueller, John Novembre, and Robert K Wayne. Natural selection and the origin of a melanistic allele in north american gray wolves. *Molecular Biology and Evolution*, 2018.

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. Hybridization and adaptive radiation. *Trends in Ecology & Evolution*, 19(4):198–207, 2004.

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.

Laure Ségurel, Emma E Thompson, Timothée Flutre, Jessica Lovstad, Aarti Venkat, Susan W Margulis, Jill Moyse, Steve Ross, Kathryn Gamble, Guy Sella, et al. The abo blood group is a trans-species polymorphism in primates. *Proceedings of the National Academy of Sciences*, 109(45):18493–18498, 2012.

Laure Ségurel, Minyoung J Wyman, and Molly Przeworski. Determinants of mutation rate variation in the human germline. *Annual Review of Genomics and Human Genetics*, 15:47–70, 2014.

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

N. W. Simmonds. Introgression and incorporation. strategies for the use of crop genetic resources. *Biological Reviews*, 68(04):539–562, 1993.

Tatum S Simonson, Yingzhong Yang, Chad D Huff, Haixia Yun, Ga Qin, David J Witherspoon, Zhenzhong Bai, Felipe R Lorenzo, Jinchuan Xing, Lynn B Jorde, et al. Genetic evidence for high-altitude adaptation in tibet. *Science*, 329(5987): 72–75, 2010.

Frederick J Simoons. Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. *The American Journal of Digestive Diseases*, 15(8):695–710, 1970.

Sonal Singhal, Ellen M Leffler, Keerthi Sannareddy, Isaac Turner, Oliver Venn, Daniel M Hooper, Alva I Strand, Qiye Li, Brian Raney, Christopher N Balakrishnan, et al. Stable recombination hotspots in birds. *Science*, 350(6263):928–932, 2015.

Pontus Skoglund, Helena Malmström, Ayça Omrak, Maanasa Raghavan, Cristina Valdiosera, Torsten Günther, Per Hall, Kristiina Tambets, Jüri Parik, Karl-Göran Sjögren, et al. Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science*, 344(6185):747–750, 2014.

Pontus Skoglund, Erik Ersmark, Eleftheria Palkopoulou, and Love Dalén. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current Biology*, 25(11):1515–1519, 2015.

Montgomery Slatkin. Simulating genealogies of selected alleles in a population of variable size. *Genetical Research*, 78(01):49–57, 2001.

Montgomery Slatkin. A Bayesian method for jointly estimating allele age and selection intensity. *Genetics Research*, 90(01):129–137, 2008.

Montgomery Slatkin and Richard R Hudson. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129 (2):555–562, 1991.

Montgomery Slatkin and Bruce Rannala. Estimating allele age. *Annual Review of Genomics and Human Genetics*, 1(1):225–249, 2000.

Joel Smith, Graham Coop, Matthew Stephens, and John Novembre. Estimating time to the common ancestor for a beneficial allele. *Molecular Biology and Evolution*, 2018.

John Maynard Smith and John Haigh. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(01):23–35, 1974.

Ying Song, Stefan Endepols, Nicole Klemann, Dania Richter, Franz-Rainer Matuschka, Ching-Hua Shih, Michael W Nachman, and Michael H Kohn. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Current Biology*, 21(15):1296–1301, 2011.

Daniel R Stahler, Daniel R MacNulty, Robert K Wayne, Bridgett VonHoldt, and Douglas W Smith. The adaptive value of morphological, behavioural and life-history traits in reproductive female wolves. *Journal of Animal Ecology*, 82(1): 222–234, 2013.

Laurie S Stevison, August E Woerner, Jeffrey M Kidd, Joanna L Kelley, Krishna R Veeramah, Kimberly F McManus, Great Ape Genome Project, Carlos D Bustamante, Michael F Hammer, and Jeffrey D Wall. The time scale of recombination rate evolution in great apes. *Molecular Biology and Evolution*, 33(4):928–945, 2015.

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

John R Stinchcombe and Hopi E Hoekstra. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, 100(2):158–170, 2008.

Michael PH Stumpf and Gilean AT McVean. Estimating recombination rates from population-genetic data. *Nature Reviews Genetics*, 4(12):959, 2003.

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

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

Hua Tang, David O Siegmund, Peidong Shen, Peter J Oefner, and Marcus W Feldman. Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics*, 161(1):447–459, 2002.

Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

Jacob A Tennessen, Abigail W Bigham, Timothy D OConnor, Wenqing Fu, Eimear E Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69, 2012.

Russell Thomson, Jonathan K Pritchard, Peidong Shen, Peter J Oefner, and Marcus W Feldman. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proceedings of the National Academy of Sciences*, 97 (13):7360–7365, 2000.

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.

Sarah A Tishkoff, Floyd A Reed, Alessia Ranciaro, Benjamin F Voight, Courtney C Babbitt, Jesse S Silverman, Kweli Powell, Holly M Mortensen, Jibril B Hirbo, Maha Osman, et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, 39(1):31–40, 2007.

Jesper T Troelsen, Jørgen Olsen, Jette Møller, and Hans SjÖstrÖm. An upstream polymorphism associated with lactase persistence has increased enhancer activity. *Gastroenterology*, 125(6):1686–1694, 2003.

Michael Turelli, Jeremy R Lipkowitz, and Yaniv Brandvain. On the Coyne and Orrigin of species: effects of intrinsic postzygotic isolation, ecological differentiation,

X chromosome size, and sympatry on Drosophila speciation. *Evolution*, 68(4): 1176–1187, 2014.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42, 2011.

Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

Joseph J Vitti, Sharon R Grossman, and Pardis C Sabeti. Detecting natural selection in genomic data. *Annual Review of Genetics*, 47:97–120, 2013.

Benjamin F Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. A map of recent positive selection in the human genome. *PLoS Biology*, 4(3):446, 2006.

Jeffrey D Wall and Jonathan K Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4(8):587, 2003.

Binbin Wang, Yong-Biao Zhang, Feng Zhang, Hongbin Lin, Xumin Wang, Ning Wan, Zhenqing Ye, Haiyu Weng, Lili Zhang, Xin Li, et al. On the origin of tibetans and their genetic basis in adapting high-altitude environments. *PLoS One*, 6(2): e17002, 2011.

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

Daniel Wegmann, Darren E Kessner, Krishna R Veeramah, Rasika A Mathias, Dan L Nicolae, Lisa R Yanek, Yan V Sun, Dara G Torgerson, Nicholas Rafaels, Thomas Mosley, et al. Recombination rates in admixed individuals identified by ancestry-based inference. *Nature Genetics*, 43(9):847–853, 2011.

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.

Sandra Wilde, Adrian Timpson, Karola Kirsanow, Elke Kaiser, Manfred Kayser, Martina Unterländer, Nina Hollfelder, Inna D Potekhina, Wolfram Schier, Mark G Thomas, et al. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 years. *Proceedings of the National Academy of Sciences*, 111(13):4832–4837, 2014.

Scott H Williamson, Ryan Hernandez, Adi Fledel-Alon, Lan Zhu, Rasmus Nielsen, and Carlos D Bustamante. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences*, 102(22):7882–7887, 2005.

Carsten Wiuf. On the genealogy of a sample of neutral rare alleles. *Theoretical Population Biology*, 58(1):61–75, 2000.

Carsten Wiuf and Peter Donnelly. Conditional genealogies and the age of a neutral mutant. *Theoretical Population Piology*, 56(2):183–201, 1999.

Shuhua Xu, Shilin Li, Yajun Yang, Jingze Tan, Haiyi Lou, Wenfei Jin, Ling Yang, Xuedong Pan, Jiucun Wang, Yiping Shen, et al. A genome-wide search for signals of high-altitude adaptation in tibetans. *Molecular Biology and Evolution*, 28(2): 1003–1011, 2010.

D Yang, OBSN Chertov, SN Bykovskaia, Q Chen, MJ Buffo, J Shogan, M Anderson, JM Schröder, JM Wang, OMZ Howard, et al. β-defensins: linking innate and adaptive immunity through dendritic and t cell ccr6. *Science*, 286(5439):525–528, 1999.

Xin Yi, Yu Liang, Emilia Huerta-Sanchez, Xin Jin, Zha Xi Ping Cuo, John E Pool, Xun Xu, Hui Jiang, Nicolas Vinckenbosch, Thorfinn Sand Korneliussen, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329 (5987):75–78, 2010.